Advances in Human-Computer Interaction

Emotion-Aware Natural Interaction

Guest Editors: Kostas Karpouzis, Elisabeth Andre, and Anton Batliner

Emotion-Aware Natural Interaction

Emotion-Aware Natural Interaction

Guest Editors: Kostas Karpouzis, Elisabeth Andre, and Anton Batliner

Editorial Board

Contents

Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 309512, 2 pages
doi:10.1155/2010/309512

Editorial
Emotion-Aware Natural Interaction

Kostas Karpouzis,[1] Elisabeth Andre,[2] and Anton Batliner[3]

[1] *National Technical University of Athens, Greece*
[2] *University of Augsburg, Germany*
[3] *Friedrich Alexander University Erlangen-Nuremberg, Germany*

Correspondence should be addressed to Kostas Karpouzis, kkarpou@cs.ntua.gr

Received 15 June 2010; Accepted 15 June 2010

Emotion-aware human-computer interaction has been in the forefront of research and development interest for some time now and many applications integrating concepts of the related theory gradually make it to the users. This interdisciplinary field encompasses concepts from a wide variety of research themes, ranging from psychology and cognition theory to signal processing and understanding, as well as evaluation and software engineering and design. It is the blending and cross-fertilization of those concepts which in the end makes emotion-aware computing and robotic systems perform better and offer a richer experience when humans interact with them.

This special issue was initiated by our joint work in the framework of the Humaine Network of Excellence, which was transformed to the Humaine Association [1] after the end of the funding period. More than thirty partners from Europe and the U.S. participated in Humaine, providing the seeds for a number of research and development projects on affective computing, which maintain the high quality research in the field. The issue consists of five papers, dealing with emotion recognition, emotion-oriented architectures, and assistive computing.

The paper titled "*The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems*" by Marc Schroeder presents the SEMAINE API, an open source framework for building emotion-oriented systems. By encouraging and simplifying the use of standard representation formats, the presented work aims to contribute to interoperability and reuse of system components in the research community. An interactive Sensitive Artificial Listener built within the framework of the Semaine EU project is presented as an example of a full-scale system built on top of this API. Three small example systems are described in detail to illustrate how integration between existing and new video and speech analysis components is realized with minimal effort. Schroeder concludes that if several research teams were to bring their work into a common technological framework, such as the one presented in this paper, this would be likely to speed up the consolidation process, because challenges to integration would become apparent more quickly.

Data preprocessing for speech-based emotion recognition is the subject of the paper titled "*Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach*" by Anton Batliner et al. Authors work on a database with children's emotional speech to illustrate their approach on segmenting emotion-related (emotional or affective) episodes into adequate units for analysis and automatic processing and classification. Using word-based annotations and the subsequent mapping onto different types of higher units, Batliner et al. report classification performances for an exhaustive modeling of this data onto three classes representing valence (positive, neutral, negative), and onto a fourth rest (garbage) class.

Driver assistance is the subject of the next paper titled "*Emotion on the road-necessity, acceptance, and feasibility of affective computing in the car*" by Florian Eyben et al. Authors mention that the ability of a car to understand natural speech and provide a human-like driver assistance system can be expected to be a decisive factor for market success on par with automatic driving systems. Starting with an extensive literature overview of work related to emotions and driving, as well as automatic recognition and control of emotions, Eyben et al. describe various use-case scenarios

as possible applications for emotion-oriented technology in a car. Acceptance from the part of the drivers of such technology is evaluated with a Wizard-Of-Oz study, while feasibility of monitoring driver attentiveness is demonstrated by a real-time experiment.

The fourth paper of this special issue deals with interactions taking place in the Second Life virtual world. In "*EmoHeart: Conveying emotions in second life based on affect sensing from text*", authors Alena Neviarouskaya et al. look at effect sensing from text, which enables automatic expression of emotions in the virtual environment, as a method to avoid manual control by the user and to enrich remote communications effortlessly. A lexical rule-based approach to recognition of emotions from text is described, the results of which trigger animations of avatar facial expressions and visualize emotion by heart-shaped textures. Authors report promising results in fine-grained emotion recognition in real examples of online conversation in both their own, as well as an existing corpus.

The final paper of this special issue is titled "*Emotional communication in finger braille*". Here, authors Yasuhiro Matsuda et al. describe analyses of the features of three emotion classes (joy, sadness, and anger) expressed by Finger Braille interpreters and examine the effectiveness of emotional expression and emotional communication between people unskilled in Finger Braille, targeting the development of a Finger Braille system to teach emotional expression and a system to recognize emotion. Their results indicate that code duration and finger load are correlated with some of the emotion classes and, based on the analysis of effectiveness of emotional expression and emotional communication between unskilled users, expression and communication of emotions using Finger Braille is both feasible and comprehensible.

We believe that this special issue details concepts and implementations from a wide variety of effect- and emotion-related applications, effectively illustrating research challenges and efforts taken to tackle them. Finally, we would like to thank all authors and reviewers and also the Advances in Human-Computer Interaction Journal for hosting this special issue.

Kostas Karpouzis
Elisabeth Andre
Anton Batliner

References

[1] Humaine Association, http://emotion-research.net/.

Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 319406, 21 pages
doi:10.1155/2010/319406

Research Article

The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems

Marc Schröder

German Research Center for Artificial Intelligence DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

Correspondence should be addressed to Marc Schröder, schroed@dfki.de

Received 25 March 2009; Accepted 11 October 2009

Academic Editor: Anton Batliner

This paper presents the SEMAINE API, an open source framework for building emotion-oriented systems. By encouraging and simplifying the use of standard representation formats, the framework aims to contribute to interoperability and reuse of system components in the research community. By providing a Java and C++ wrapper around a message-oriented middleware, the API makes it easy to integrate components running on different operating systems and written in different programming languages. The SEMAINE system 1.0 is presented as an example of a full-scale system built on top of the SEMAINE API. Three small example systems are described in detail to illustrate how integration between existing and new components is realised with minimal effort.

1. Introduction

Systems with some emotional competence, so-called "affective computing" systems, are a promising and growing trend in human-machine interaction (HMI) technology. They promise to register a user's emotions and moods, for example, to identify angry customers in interactive voice response (IVR) systems and to generate situationally appropriate emotional expression, such as the apologetic sound of a synthetic voice when a customer request cannot be fulfilled; in certain conditions they even aim to identify reasons for emotional reactions, using so-called "affective reasoning" technology. Ultimately, such technology may indeed lead to more natural and intuitive interactions between humans and machines of many different kinds, and thus contribute to bridging the "digital divide" that leaves nontechy users helpless in front of increasingly complex technology. This aim is certainly long term; slightly more in reach is the use of emotion-oriented technologies in the entertainment sector, such as in computer games, where emotional competence, even in a rudimentary state, can lead to new effects and user experiences.

In fact, an increasing number of interactive systems deal with emotions and related states in one way or another. Common tasks include the analysis of the user's affective state [1] from the face [2, 3], the voice [4, 5], or physiological measures [6], the evaluation of events in the world according to affective criteria [7], and the generation of emotion-related system behaviour, such as facial expression [8, 9] and voice [10, 11], but also other media such as music and colour [12].

A number of elements are common to the different systems. All of them need to represent emotional states in order to process them; and many of the systems are built from components, such as recognition, reasoning, or generation components, which need to communicate with one another to provide the system's capabilities. In the past, systems used custom solutions to these challenges, usually in clearly delimited ways that were tailor made for their respective application areas (see Section 2 for related work). However, existing emotion-oriented systems neither seem to be explicitly geared towards the use of standard representations nor are they available as open source.

Standards enable interoperability and reuse. Nowadays, standards are taken for granted in such things as the voltage of electricity in a given country, fuel grade, or the dimensions of screw threads [13]. More recently, standards for document formats [14] have entered the public debate, under the perspective of long-term accessibility of documents. Web standards such as the Hyper-Text Markup Language HTML [15] enable access to information in the world wide web through a broad variety of software products supporting the standard format.

Proprietary formats, on the other hand, can be used to safeguard a company's competitive advantage. By patenting, or even by simply not documenting a representation format,

a company can make sure not to open up the market to its competitors.

The same considerations seem to apply in the emerging area of emotion-oriented systems. Agreeing on standard formats and interfaces would enable interoperability and reuse. An initial investment of effort in defining suitably broad but sufficiently delimited standard formats can be expected to pay off in the long run by removing the need to start from scratch with every new system built. Where formats, software frameworks, and components are made generally available, for example, as open source, these can be used as starting points and building blocks for new systems, speeding up development and research.

This paper describes the SEMAINE API, a toolkit and middleware framework for building emotion-oriented systems in a modular way from components that communicate using standard formats where possible. It describes one full-scale system built using the framework and illustrates the issue of reuse by showing how three simple applications can be built with very limited effort.

The paper is structured as follows. Section 2 reviews related work. Section 3 presents the SEMAINE API from the technological point of view, including the support for integrating components into a system and the supported representation formats. Section 4 describes a first larger system based on the API, the SEMAINE system 1.0, and provides some detail on the system architecture and the components used in the system. Section 5 exemplifies the use of the SEMAINE API for building new emotion-oriented systems by showing how to implement three demo systems. Section 6 presents an evaluation of the framework in terms of response times and developer friendliness.

2. Related Work

Several integrated research systems have been built in the recent past which incorporate various aspects of emotional competence. For example, the NECA project [16] generated scripted dialogues between embodied conversational agents in a web environment. Its system was based on a pipeline architecture in which a Rich Representation Language (RRL) [17] was successively enriched with component information. The VirtualHuman project [18] supported dialogues involving multiple humans and multiple agents. Both humans and agents were represented in the system by Conversational Dialogue Engines [19] communicating with each other using the concepts of an application-specific ontology. The FearNot! system [20], an educational application helping children to deal with bullying, uses an architecture involving reactive and deliberative layers and memory components, as well as sensors and effectors. Central to the processing of emotions in FearNot are appraisal processes realised in terms of the OCC model [21]. The project IDEAS4Games realised an emotion-aware poker game, in which two agents and a user played against each others with physical cards carrying RFID tags [22]. The emotions of characters were computed from game events using an affective reasoner [7] and realised through the synthetic voice and through body movements. Whereas all of these systems are conceptually modular, none of them is explicitly geared towards the use of standard representations, and none of the systems is available as open source.

Existing programming environments provide relevant component technologies but do not allow the user to integrate components across programming languages and operating system platforms. For example, the EMotion FX SDK [23] is a character animation engine that supports animation designers to streamline the process of designing the graphical properties of games characters and include them into a game environment. It includes facial animation such as emotional facial expressions and lip synchronisation. Luxand FaceSDK [24] is a facial feature point detection software, which can be used for face detection, the generation of 3D face models, and the automatic creation of animated avatars. Both are relevant component technologies for an emotion-oriented system, but do not solve the issue of how to integrate heterogeneous components across platforms.

When looking beyond the immediate area of emotion-oriented systems, however, we find several toolkits for component integration.

In the area of ubiquitous computing, the project Computers in the Human Loop (CHIL) investigated a broad range of smart space technologies for smart meeting room applications. Its system integration middleware, named CHILix [25], uses XML messages for integrating the components in a smart space application. CHILix uses the freely available NIST DataFlow System II [26] as the low-level message routing middleware. The XML message format used seems to be a domain-specific, custom format; documentation does not seem to be freely available.

In the domain of interactive robots research, the project CognitiveSystems (CoSy) has developed a system integration and communication layer called CoSy Architecture Schema Toolkit (CAST) [27]. The components of a robot's architecture are structured into subarchitectures in which components work on a jointly accessible working memory. Access to data structures is through predefined *types*, similar to objects. Communication passes through the object-oriented middleware Ice [28].

The main features of the CHILix, CAST, and SEMAINE API integration frameworks are summarised in Table 1. It can be seen that out of these frameworks the SEMAINE API is the only one that is based on standard formats and can be flexibly used with closed and open source components due to its less restrictive LGPL license.

3. The SEMAINE API

The SEMAINE API has been created in the EU-funded project "SEMAINE: Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression" [29], as a basis for the project's system integration. The project aims to build a multimodal dialogue system with an emphasis on nonverbal skills—detecting and emitting vocal and facial signs related to the interaction, such as backchannel signals, in order to register and express information such as continued presence, attention or interest, an evaluation of the content, or an emotional connotation. The system

Table 1: Key properties of several component integration frameworks for real-time systems.

	CHILix	CAST	SEMAINE API
Application domain	Smart Spaces	Interactive Robots	Emotion-oriented systems
Integration approach	XML messages	Objects in shared working memory	XML messages
Using standard formats	No	No	Yes
Operating systems	Windows, Linux, Mac	Linux, Mac	Windows, Linux, Mac
Programming languages	C++, Java	C++, Java	C++, Java
Low-level communication platform	NDFS II	Ice	ActiveMQ
Open source	No	Yes, GPL	Yes, LGPL

has strong real-time constraints, because it must react to the user's behaviour while the user is still speaking [30].

The project's approach is strongly oriented towards making basic technology for emotion-oriented interactive systems available to the research community, where possible as open source. While the primary goal is to build a system that can engage a human user in a conversation in a plausible way, it is also an important aim to provide high-quality audiovisual recordings of human-machine interactions, as well as software components that can be reused by the research community.

In front of this background, the SEMAINE API has the following main aims:

(i) to integrate the software components needed by the SEMAINE project in a robust, real-time system capable of multimodal analysis and synthesis,

(ii) to enable others to reuse the SEMAINE components, individually or in combination, as well as to add their own components, in order to build new emotion-oriented systems.

The present section describes how the SEMAINE API supports these goals on a technical level. First, we present the SEMAINE API's approach to system integration, including the message-oriented middleware used for communication between components, as well as the software support provided for building components that integrate neatly into the system and for producing and analysing the representation formats used in the system. After that, we discuss the representation formats used, their status with respect to standardisation, and the extent to which domain-specific representations appear to be needed.

3.1. System Integration. Commercial systems often come as single, monolithic applications. In these systems, the integration of system components is as tight as possible: any system internal components communicate via shared memory access, and any modularity is hidden from the end user.

In the research world, the situation is different. Different research teams, cooperating in research projects in different constellations, are deeply rooted in different traditions; the components they contribute to a system are often extensions of preexisting code. In such situations, the only way to fully integrate all system components into a single binary executable would be to reimplement substantial portions of the code. In most cases, research funding will not provide the resources for that. Therefore, it is often necessary to build an overall system from components that may be running on different operating systems and that may be written in different programming languages.

Key properties of system integration are as follows. The SEMAINE API uses a message-oriented middleware (MOM; see Section 3.1.1) for all communication in the system. As a result, all communication is asynchronous, which decouples the various parts of the system. The actual processing is done in "components", which communicate with one another over "Topics" (see Section 3.1.2) below the named Topic hierarchy `semaine.data.*`. Each component has its own "metamessenger", which interfaces between the component and a central system manager. When a component is started, its metamessenger registers with the system manager over a special metacommunication channel, the Topic `semaine.meta`. At registration time, the metamessenger describes the component in terms of the data Topics that it sends data to and that it receives data from; if the component is an input or output component (in the sense of the user interface), that status is communicated as well. The system manager is keeping track of the components that have been registered and checks at regular intervals whether all components are still alive by sending a "ping". In reply to such a ping, each metamessenger confirms the respective component's status and sends debug information such as the average time spent processing incoming requests. The system manager keeps track of the information about registered components and sends global meta messages informing all components that the overall system is ready or, if a component has an error or is stalled, that the system is not ready. Also, the system manager resets a global timer to zero when the system becomes ready. All components use this global time via their metamessenger, and thus can meaningfully communicate about timing of user and system events even across different computers with potentially unsynchronised hardware clocks.

A centralised logging functionality uses the Topics below `semaine.log.*`. By convention, messages are sent to `semaine.log.<component>.<severity>`, for example, the Topic `semaine.log.UtteranceInterpreter.debug` would be used for debug messages of component UtteranceInterpreter. The severities used are "`debug`", "`info`", "`warn`", and "`error`". Through this design, it is possible for a log reader to subscribe, for example, to all

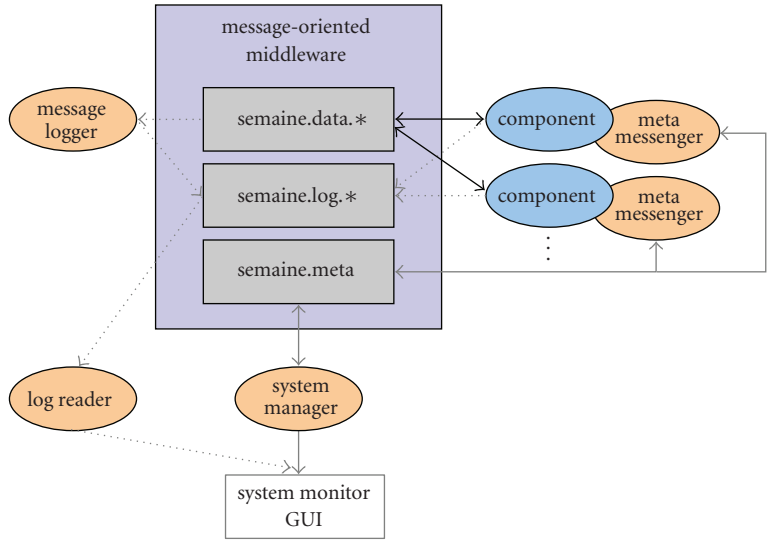


Figure 1: SEMAINE API system architecture.

types of messages from one component or to all messages from all components that have at least severity "`info`", and so forth. Furthermore, a configurable message logger can optionally be used to log certain messages in order to follow and trace them. Notably, it is possible to read log messages in one central place, independently of the computer, operating system, or programming language used by any given component.

Figure 1 illustrates this system architecture. Components communicate with each other via Topics in the `semaine.data` hierarchy (indicated by black arrows). Metainformation is passed between each component's metamessenger and the system manager via the `semaine.meta` Topic (grey arrows). Optionally, components can write log messages, and a message logger can log the content messages being sent; a configurable log reader can receive and display a configurable subset of the log messages (dashed grey arrows).

Optionally, a system monitor GUI visualises the information collected by the system manager as a message flow graph. Input components are placed at the bottom left, output components at the bottom right, and the other components sorted to the extent possible based on the data input/output relationships, along a half-circle from left to right. Component B comes later in the graph than component A if A's output is an input to B or if there is a sequence of components that can process A's output into B's input. This criterion is overly simplistic for complex architectures, especially with circular message flows, but is sufficient for simple quasilinear message flow graphs. If a new component is added, the organisation of the flow graph is recomputed. This way, it is possible to visualise message flows without having to prespecify the layout.

Figure 2 shows the system monitor GUI for the SEMAINE system 1.0 described in Section 4. Components are represented as ovals, whereas Topics are represented as rectangles. Topics are shown in yellow when they have just transported a new message and in grey when they have not seen any recent messages. When the user clicks on the Topic rectangle, the GUI shows a history of the messages transported by a given Topic; debug information about a component is shown when the user clicks on the component oval. A log message reader is shown on the right-hand side. It can be configured with respect to the components and the severity of messages to show.

The remainder of this section describes the various aspects involved in the system in some more detail.

3.1.1. Message-Oriented Middleware. A message-oriented middleware (MOM) [31] is specifically designed to integrate different applications or processes through messages being routed from publishers to subscribers. The aim is to "glue together applications both within and across organisations, without having to reengineer individual components" [31]. One method for describing how messages should be sent from sources to destinations is a message flow graph, in which the nodes represent components and the arcs represent message flows [31]. A major advantage of a generic message-oriented middleware lies in its flexibility. Through a publish-subscribe model, n-to-m connections are trivial to realise; furthermore, the system architecture can be rearranged very easily—adding or removing a component consuming or producing a certain message type does not require any changes elsewhere in the architecture. Communication is asynchronous, so that a publisher does not need to wait for confirmation by subscribers that a message has been received. Where a response is needed, this must be sent as a separate asynchronous message, and the response's receiver needs to match it to the original request.

The SEMAINE API provides an abstraction layer over a MOM that allows the components to deal with messages in a type-specific way. The low-level serialisation and deserialisation processes are encapsulated and hidden from the user code. As a result, it is potentially possible to exchange one MOM against another one without any changes in the user code.

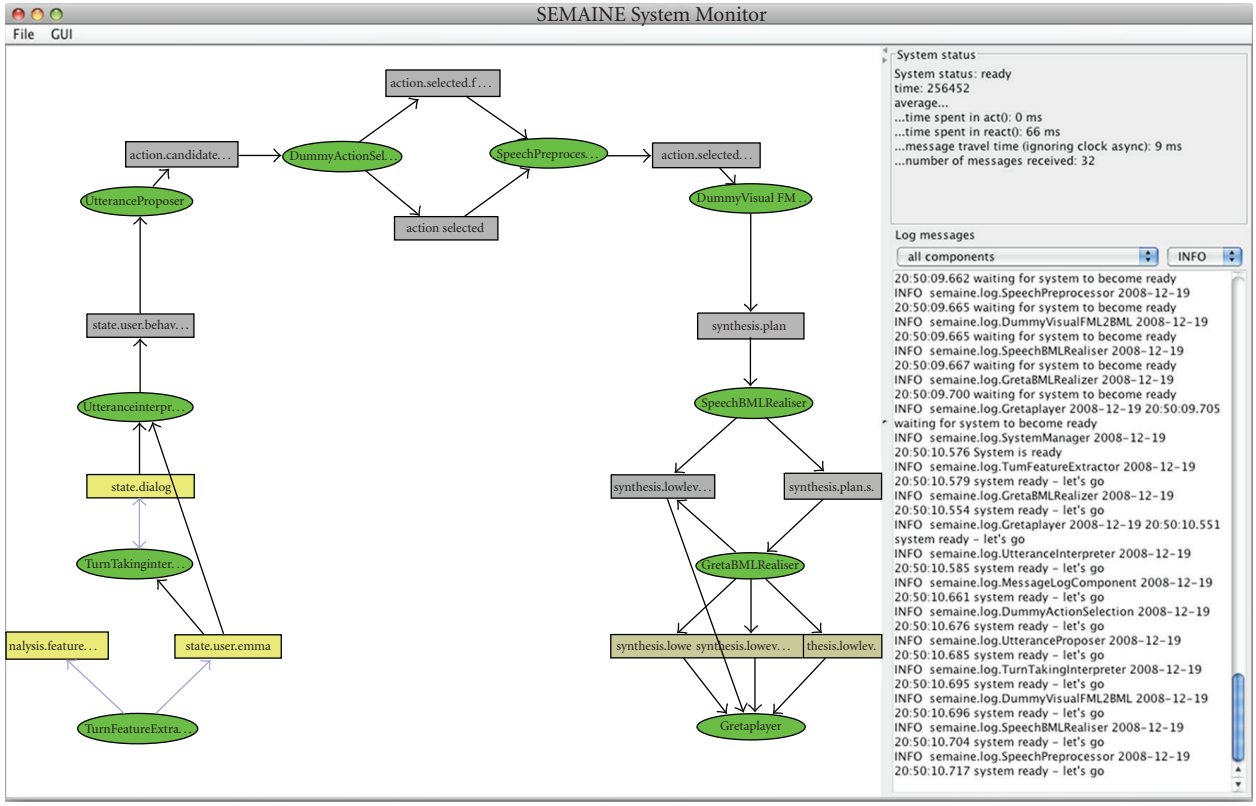


Figure 2: Screenshot of the system monitor GUI showing the implemented SEMAINE system 1.0.

The MOM currently used in the SEMAINE API is ActiveMQ from the Apache project [32]. ActiveMQ is an open-source implementation of the Java Message Service (JMS) server specification [33]. It provides client code in Java, C++, and various other programming languages, is reasonably fast, and is actively used, supported, and developed at the time of this writing.

3.1.2. Topics. In its "publish-subscribe" model, JMS routes messages via so-called *Topics* which can be identified by name. The SEMAINE API adopts this terminology. Names of Topics can be arbitrarily chosen. In order to establish a communication link from component A to component B, it is sufficient for component A to register as a "publisher" to a named Topic and for component B to register as a "subscriber" to the same Topic. Whenever A sends a message to the Topic, B will receive the message. Topics allow for an arbitrary number of publishers and subscribers, so that it is trivial to set up n-to-m communication structures.

For a given system, it is reasonable to choose Topics such that they represent data of a specific type, that is, with a well-defined meaning and representation format. This type of data may be produced by several system components, such as a range of modality-specific emotion classifiers. If there are no compelling reasons why their outputs need to be treated differently, it is possible to use a single Topic for their joint output by registering all the components producing this data type as publishers to the Topic. Similarly, several components may reasonably take a given type of data as input, in which case all of them should register as subscribers to the respective Topic. Using Topics as "information hubs" in this way immensely simplifies the clarity of information flow, and consequently the message flow graph see Figure 2 for example.

3.1.3. Components. The creation of custom components is made simple by the base class `Component` in the SEMAINE API, which provides the basic functionality required for the component to interact with the rest of the system. The `Component` will register as a subscriber and/or as a publisher to a configurable list of Topics using suitable, possibly type-specific message receivers and message senders. Whenever a message is received, the subclass's `react()` method is called, allowing the component to process the data and perform some action, including the emission of an output message. In addition, the method `act()` is called at configurable intervals (every 100 ms by default), allowing for actions to be triggered by component internal mechanisms such as timeouts or custom process logic.

The `Component` base implementation also instantiates the metamessenger (see Figure 1) which handles all meta communication with the system manager, without requiring customisation by user code in subclasses. Examples of simple component classes are provided in Section 5.

3.1.4. API Support for Relevant Representation Types. The SEMAINE API aims to be as easy to use as possible, while allowing for state-of-the-art processing of data. This

principle is reflected in an extensive set of support classes and methods for parsing, interpreting, and creating XML documents in general and the representations specially supported (see Section 3.2) in particular. XML processing is performed by the standards-compliant parser Xerces [34] which converts between a textual representation contained in messages and a user-friendly Document Object Model (DOM) representation [35]. Parsing is done in a namespace-aware manner in order to maintain a clear distinction between the elements used in mixed representations. Examples of mixed representations are the use of the Extensible Multimodal Annotation language (EMMA) to transport a recognition result expressed in EmotionML and the use of the Speech Synthesis Markup Language (SSML) to encode the speech aspect of ECA behaviour in the Behaviour Markup Language (BML). These combinations make perfect sense; namespaces are a suitable method for clearly identifying the markup type of any given element when interpreting a mixed representation.

Support classes exist for the representation formats listed in Section 3.2, notably as dedicated receiver, sender, and message objects. For example, when a component registers an `EmmaReceiver` to listen to a given Topic, it will receive messages directly as `SEMAINEEmmaMessage` objects with methods appropriate for interpreting the content of the EMMA data; a `FeatureSender` will take an array of `float` values and send it as a textual or binary feature message; `BinarySender` and `BinaryReceiver` classes can be used to transport, for example, audio data between components.

In sum, these support classes and methods simplify the task of passing messages via the middleware and help avoid errors in the process of message passing by implementing standard encoding and decoding procedures. Where representations beyond those previewed by the API are required, the user always has the option to use lower-level methods such as plain XML or even text messages and implement a custom encoding and decoding mechanism.

3.1.5. Supported Platforms. The SEMAINE API is currently available in Java and as a shared library in C++, for Linux, Mac OS X, and Windows. State-of-the-art build tools (Eclipse and ant for Java, Visual Studio for C++ on Windows, GNU automake/autoconf for C++ on Linux and Mac) are provided to make the use of the API as simple and portable as possible.

3.1.6. Current Status. As of version 1.0.1, the SEMAINE API is fully functional, but the support for the individual representation formats is preliminary. Not all elements and attributes defined in the specifications mentioned in Section 3.2 are predefined as constants in the API support classes. This limitation is an issue of coverage rather than principle: on the one hand, it is straightforward to add the missing element and attribute names to the lists of string constants; on the other hand, user code can always add custom string constants or use ad hoc strings to create or read XML elements and attributes for which no constants have been defined yet.

Other aspects are more interesting because the practical implementation has hit limits with the current version of draft specifications. For example, for the implementation of symbolic timing markers between words, the draft BML specification [36] proposes to use a `<mark>` element in the default namespace; however, we noticed that treating the speech markup as valid SSML requires the use of an `<ssml:mark>` element in the SSML namespace. Experience of this kind may be helpful in refining specifications based on implementation feedback.

3.2. Representation Formats Supported in the SEMAINE API. In view of future interoperability and reuse of components, the SEMAINE API aims to use standard representation formats where that seems possible and reasonable. For example, results of analysis components can be represented using EMMA (Extensible Multimodal Annotation), a World Wide Web Consortium (W3C) Recommendation [37]. Input to a speech synthesiser can be represented using SSML (Speech Synthesis Markup Language), also a W3C Recommendation [38].

Several other relevant representation formats are not yet standardised, but are in the process of being specified. This includes the Emotion Markup Language EmotionML [39], used for representing emotions and related states in a broad range of contexts, and the Behaviour Markup Language (BML) [40], which describes the behaviour to be shown by an Embodied Conversational Agent (ECA). Furthermore, a Functional Markup Language (FML) [41] is under discussion, in order to represent the planned actions of an ECA on the level of functions and meanings. By implementing draft versions of these specifications, the SEMAINE API can provide hands-on input to the standardisation process, which may contribute to better standard formats.

On the other hand, it seems difficult to define a standard format for representing the concepts inherent in a given application's logic. To be generic, such an endeavour would ultimately require an ontology of the world. In the current SEMAINE system, which does not aim at any sophisticated reasoning over domain knowledge, a simple custom format named SemaineML is used to represent those pieces of information that are required in the system but which cannot be adequately represented in an existing or emerging standard format. It is conceivable that other applications built on top of the SEMAINE API may want to use a more sophisticated representation such as the Rich Description Format (RDF) [42] to represent domain knowledge, in which case the API could be extended accordingly.

Whereas all of the aforementioned representation formats are based on the Extensible Markup Language XML [43], there are a number of data types that are naturally represented in different formats. This is particularly the case for the representations of data close to input and output components. At the input end, low-level analyses of human behaviour are often represented as feature vectors. At the output end, the input to a player component is likely to include binary audio data or player-specific rendering directives.

Table 2 gives an overview of the representation formats currently supported in the SEMAINE API. The following

Table 2: Representation formats currently supported by the SEMAINE API.

Type of data	Representation format	Standardisation status
Low-level input features	String or binary feature vectors	ad hoc
Analysis results	EMMA	W3C Recommendation
Emotions and related states	EmotionML	W3C Incubator Report
Domain knowledge	SemaineML	ad hoc
Speech synthesis input	SSML	W3C recommendation
Functional action plan	FML	Very preliminary
Behavioural action plan	BML	Draft specification
Low-level output data	Binary audio, player commands	Player-dependent

```
0.000860535 rmsEnergy
12.6699 logEnergy
-2.59005e-05 rmsEnergy-De
-0.0809427 logEnergy-De
...
```

Figure 3: Textual representation of a feature vector.

subsections briefly describe the individual representation formats.

3.2.1. Feature Vectors. Feature vectors can be represented in an ad hoc format. In text form (see Figure 3), the feature vectors consist of straightforward key-value pairs—one feature per line—values preceding features.

As feature vectors may be sent very frequently (e.g., every 10 ms in the SEMAINE system 1.0), compact representation is a relevant issue. For this reason, a binary representation of feature vectors is also available. In binary form, the feature names are omitted, and only feature values are being communicated. The first four bytes represent an integer containing the number of features in the vector; the remaining bytes contain the float values one after the other.

3.2.2. EMMA. The Extensible Multimodal Annotation Language (EMMA), a W3C Recommendation, is "an XML markup language for containing and annotating the interpretation of user input" [37]. As such, it is a wrapper language that can carry various kinds of payload representing the interpretation of user input. The EMMA language itself provides, as its core, the `<emma:interpretation>` element, containing all information about a single interpretation of user behaviour. Several such elements can be enclosed within an `<emma:one-of>` element in cases where more than one interpretation is present. An interpretation can have an `emma:confidence` attribute, indicating how confident the source of the annotation is that the interpretation is correct, time-related information such as `emma:start`, `emma:end`, and `emma:duration`, indicating the time span for which the interpretation is provided, information about the modality upon which the interpretation is based, through the `emma:medium` and `emma:mode` attributes, and many more.

Figure 4 shows an example EMMA document carrying an interpretation of user behaviour represented using EmotionML (see below). The interpretation refers to a start time. It can be seen that the EMMA wrapper elements and the EmotionML content are in different XML namespaces, so that it is unambiguously determined which element belongs to which part of the annotation.

EMMA can also be used to represent Automatic Speech Recognition (ASR) output, either as the single most probable word chain or as a word lattice, using the `<emma:lattice>` element.

3.2.3. EmotionML. The Emotion Markup Language EmotionML is partially specified, at the time of this writing, by the Final Report of the W3C Emotion Markup Language Incubator Group [39]. The report provides elements of a specification, but leaves a number of issues open. The language is now being developed towards a formal W3C Recommendation.

The SEMAINE API is one of the first pieces of software to implement EmotionML. It is our intention to provide an implementation report as input to the W3C standardisation process in due course, highlighting any problems encountered with the current draft specification in the implementation.

EmotionML aims to make concepts from major emotion theories available in a broad range of technological contexts. Being informed by the affective sciences, EmotionML recognises the fact that there is no single agreed representation of affective states, nor of vocabularies to use. Therefore, an emotional state `<emotion>` can be characterised using four types of descriptions: `<category>`, `<dimensions>`, `<appraisals>`, and `<action-tendencies>`. Furthermore, the vocabulary used can be identified. The EmotionML markup in Figure 4 uses a dimensional representation of emotions, using the dimension set "valence, arousal, potency", out of which two dimensions are annotated: arousal and valence.

EmotionML is aimed at three use cases: (1) Human annotation of emotion-related data; (2) automatic emotion recognition; and (3) generation of emotional system behaviour. In order to be suitable for all three domains, EmotionML is conceived as a "plug-in" language that can be used in different contexts. In the SEMAINE API, this plug-in nature is applied with respect to recognition, centrally held information, and generation, where EmotionML is used in conjunction with different markups. EmotionML can be used for representing the user emotion currently estimated

```
<emma:emma xmlns:emma="http://www.w3.org/2003/04/emma"
version="1.0">
  <emma:interpretation emma:start="123456789">
    <emotion xmlns="http://www.w3.org/2005/Incubator/emotion">
      <dimensions set="valenceArousalPotency">
        <arousal value="-0.29"/>
        <valence value="-0.22"/>
      </dimensions>
    </emotion>
  </emma:interpretation>
</emma:emma>
```

Figure 4: An example EMMA document carrying EmotionML markup as interpretation payload.

```
<dialog-state xmlns="http://www.semaine-project.eu/semaineml"
version="0.0.1">
   <speaker who="agent"/>
   <listener who="user"/>
</dialog-state>
```

Figure 5: An example SemaineML document representing dialogue state.

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
    xml:lang="en-US">
  <voice gender="female">
    And then <break/> I <emphasis>wanted</emphasis> to go.
  </voice>
</speak>
```

Figure 6: An example standalone SSML document.

from user behaviour, as payload to an EMMA message. It is also suitable for representing the centrally held information about the user state, the system's "current best guess" of the user state independently of the analysis of current behaviour. Furthermore, the emotion to be expressed by the system can also be represented by EmotionML. In this case, it is necessary to combine EmotionML with the output languages FML, BML, and SSML.

3.2.4. SemaineML. A number of custom representations are needed to represent the kinds of information that play a role in the SEMAINE demonstrator systems. Currently, this includes the centrally held beliefs about the user state, the agent state, and the dialogue state. Most of the information represented here is domain specific and does not lend itself to easy generalisation or reuse. Figure 5 shows an example of a dialogue state representation, focused on the specific situation of an agent-user dialogue targeted in the SEMAINE system 1.0 (see Section 4).

The exact list of phenomena that must be encoded in the custom SemaineML representation is evolving as the system becomes more mature. For example, it remains to be seen whether analysis results in terms of user behaviour (such as a smile) can be represented in BML or whether they need to be represented using custom markup.

3.2.5. SSML. The Speech Synthesis Markup Language (SSML) [38] is a well-established W3C Recommendation supported by a range of commercial text-to-speech (TTS) systems. It is the most established of the representation formats described in this section.

The main purpose of SSML is to provide information to a TTS system on how to speak a given text. This includes the possibility to add <emphasis> on certain words, to provide pronunciation hints via a <say-as> tag, to select a <voice> which is to be used for speaking the text, or to request a <break> at a certain point in the text. Furthermore, SSML provides the possibility to set markers via the SSML <mark> tag. Figure 6 shows an example SSML document that could be used as input to a TTS engine. It requests a female US English voice; the word "wanted" should be emphasised, and there should be a pause after "then".

3.2.6. FML. The functional markup language (FML) is still under discussion [41]. Its functionality being needed nevertheless, a working language FML-APML was created [44] as a combination of the ideas of FML with the former Affective Presentation Markup Language (APML) [45].

Figure 7 shows an example FML-APML document which contains the key elements. An <fml-apml> document contains a <bml> section in which the <speech> content contains <ssml:mark> markers identifying points in time in a symbolic way. An <fml> section then refers to those points in time to represent the fact, in this case, that an announcement is made and that the speaker herself is being referred to between marks s1:tm2 and s1:tm4. This information can be used, for example, to generate relevant gestures when producing behaviour from the functional descriptions.

The representations in the <fml> section are provisional and are likely to change as consensus is formed in the community.

For the conversion from FML to BML, information about pitch accents and boundaries is useful for the prediction of plausible behaviour time-aligned with the macrostructure of speech. In our current implementation, a speech preprocessor computes this information using TTS technology (see Section 4.2). The information is added to the end of

```
<fml-apml version="0.1">
  <bml xmlns="http://www.mindmakers.org/projects/BML" id="bml1">
    <speech id="s1" language="en-US" text="Hi, I'm Poppy."
        ssml:xmlns="http://www.w3.org/2001/10/synthesis">
         <ssml:mark name="s1:tm1"/>
         Hi,
         <ssml:mark name="s1:tm2"/>
         I'm
         <ssml:mark name="s1:tm3"/>
         Poppy.
         <ssml:mark name="s1:tm4"/>
    </speech>
  </bml>
  <fml xmlns="http://www.mindmakers.org/fml" id="fml1">
    <performative id="p2" type="announce" start="s1:tm1" end="s1:tm4"/>
    <world id="w1" ref_type="person" ref_id="self" start="s1:tm2"
end="s1:tm4"/>
  </fml>
</fml-apml>
```

Figure 7: An example FML-APML document.

```
<fml-apml version="0.1">
  <bml xmlns="http://www.mindmakers.org/projects/BML" id="bml1">
    <speech id="s1" language="en_US" text="Hi, I'm Poppy."
        ssml:xmlns="http://www.w3.org/2001/10/synthesis">
         <ssml:mark name="s1:tm1"/>
         Hi,
         <ssml:mark name="s1:tm2"/>
         I'm
         <ssml:mark name="s1:tm3"/>
         Poppy.
         <ssml:mark name="s1:tm4"/>
         <pitchaccent id="xpa1" start="s1:tm1" end="s1:tm2"/>
         <pitchaccent id="xpa2" start="s1:tm3" end="s1:tm4"/>
         <boundary id="b1" time="s1:tm4"/>
    </speech>
  </bml>
  <fml xmlns="http://www.mindmakers.org/fml" id="fml1">
    <performative id="p2" type="announce" start="s1:tm1" end="s1:tm4"/>
    <world id="w1" ref_type="person" ref_id="self" start="s1:tm2" end="s1:tm4"/>
  </fml>
</fml-apml>
```

Figure 8: Pitch accent and boundary information added to the FML-APML document of Figure 7.

the <speech> section as shown in Figure 8. This is an ad hoc solution which should be reconsidered in the process of specifying FML.

3.2.7. BML. The aim of the Behaviour Markup Language (BML) [40] is to represent the behaviour to be realised by an Embodied Conversational Agent. BML is at a relatively concrete level of specification, but is still in draft status [36].

A standalone BML document is partly similar to the <bml> section of an FML-APML document (see Figure 7); however, whereas the <bml> section of FML-APML contains only a <speech> tag, a BML document can contain elements representing expressive behaviour in the ECA at a broad range of levels, including <head>, <face>, <gaze>, <body>, <speech>, and others. Figure 9 shows an example of gaze and head nod behaviour added to the example of Figure 7.

While creating an audio-visual rendition of the BML document, we use TTS to produce the audio and the timing information needed for lip synchronisation. Whereas BML in principle previews a <lip> element for representing this information, we are uncertain how to represent exact timing information with it in a way that preserves the information about syllable structure and stressed syllables. For this reason, we currently use a custom representation based on the MaryXML format from the MARY TTS system [46] to represent the exact timing of speech sounds. Figure 10 shows the timing information for the word "Poppy", which is a two-syllable word of which the first one is the stressed syllable.

The custom format we use for representing timing information for lip synchronisation clearly deserves to be revised towards a general BML syntax, as BML evolves.

3.2.8. Player Data. Player data is currently treated as unparsed data. Audio data is binary, whereas player directives are considered to be plain text. This works well with the current MPEG-4 player we use (see Section 4) but may need to be generalised as other players are integrated into the system.

4. The SEMAINE System 1.0

The first system built with the SEMAINE API is the SEMAINE system 1.0, created by the SEMAINE project. It is an early-integration system which does not yet represent the intended application domain of SEMAINE, the Sensitive Artificial Listeners [30], but achieves a first integrated system based on existing components from the project partners.

```
<bml xmlns="http://www.mindmakers.org/projects/BML" id="bml1">
  <speech id="s1" language="en_US" text="Hi, I'm Poppy."
      ssml:xmlns="http://www.w3.org/2001/10/synthesis">
    <ssml:mark name="s1:tm1"/>
    Hi,
    <ssml:mark name="s1:tm2"/>
    I'm
    <ssml:mark name="s1:tm3"/>
    Poppy.
   <ssml:mark name="s1:tm4"/>
    <pitchaccent id="xpa1" start="s1:tm1" end="s1:tm2"/>
    <pitchaccent id="xpa2" start="s1:tm3" end="s1:tm4"/>
    <boundary id="b1" time="s1:tm4"/>
  </speech>
  <gaze id="g1" start="s1:tm1" end="s1:tm4">
    ...
  </gaze>
  <head id="h1" start="s1:tm3" end="s1:tm4" type="NOD">
    ...
  </head>
</bml>
```

Figure 9: An example BML document containing SSML and gestural markup.

```
<bml xmlns="http://www.mindmakers.org/projects/BML" id="bml1">
  <speech id="s1" language="en_US" text="Hi, I'm Poppy."
      ssml:xmlns="http://www.w3.org/2001/10/synthesis"
      mary:xmlns="http://mary.dfki.de/2002/MaryXML">
    ...
    <ssml:mark name="s1:tm3"/>
    Poppy.
    <mary:syllable stress="1">
      <mary:ph d="0.092" end="1.011" p="p"/>
      <mary:ph d="0.112" end="1.123" p="A"/>
      <mary:ph d="0.093" end="1.216" p="p"/>
    </mary:syllable>
    <mary:syllable>
      <mary:ph d="0.141" end="1.357" p="i"/>
    </mary:syllable>
    <ssml:mark name="s1:tm4"/>
   ...
</bml>
```

Figure 10: An excerpt of a BML document enriched with TTS timing information for lip synchronisation.

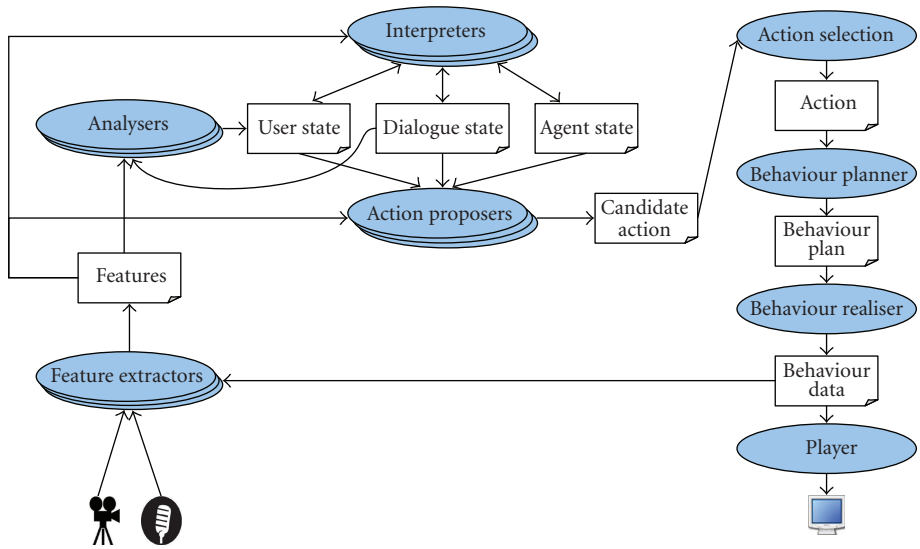


Figure 11: Conceptual message flow graph of the SEMAINE system.

The present section describes the system 1.0, first from the perspective of system architecture and integration, then with respect to the components which at the same time are available as building blocks for creating new systems with limited effort see Section 5, for examples.

4.1. Conceptual System Architecture. Figure 2 shows a message flow graph representing the conceptual system architecture of the intended SEMAINE system [30], which is partially instantiated by the SEMAINE system 1.0 [47]. Processing components are represented as ovals and data as rectangles. Arrows are always between components and data and indicate which data is produced by or is accessible to which component.

It can be seen that the rough organisation follows the simple tripartition of input (left), central processing (middle), and output (right) and that arrows indicate a rough pipeline for the data flow, from input analysis via central processing to output generation.

The main aspects of the architecture are outlined as follows. Feature extractors analyse the low-level audio and video signals and provide feature vectors periodically to the following components. A collection of analysers, such as monomodal or multimodal classifiers, produce a context-free, short-term interpretation of the current user state, in terms of behaviour (e.g., a smile) or of epistemic-affective states (emotion, interest, etc.). These analysers usually have no access to centrally held information about the state of the user, the agent, and the dialogue; only the speech recognition needs to know about the dialogue state, whether the user or the agent is currently speaking.

A set of interpreter components evaluate the short-term analyses of user state in the context of the current state of information regarding the user, the dialogue, and the agent itself and update these information states.

A range of action proposers produce candidate actions, independently of one another. An utterance producer will propose the agent's next verbal utterance, given the dialogue history, the user's emotion, the topic under discussion, and the agent's own emotion. An automatic backchannel generator identifies suitable points in time to emit a backchannel. A mimicry component will propose to imitate, to some extent, the user's low-level behaviour. Finally, a nonverbal behaviour component needs to generate some "background" behaviour continuously, especially when the agent is listening, but also when it is speaking.

The actions proposed may be contradictory, and thus must be filtered by an action selection component. A selected action is converted from a description in terms of its functions into a behaviour plan, which is then realised in terms of low-level data that can be used directly by a player.

Similar to an efferent copy in human motor prediction [48], behaviour data is also available to feature extractors as a prediction of expected perception. For example, this can be used to filter out the agent's speech from the microphone signal.

4.2. Components in the SEMAINE System 1.0. The actual implementation of this conceptual architecture is visualised in the system monitor screenshot in Figure 2. The following paragraphs briefly describe the individual system components. A more detailed description can be found in [47].

Low-level audio features are extracted using the openSMILE (Speech and Music Interpretation by Large-Space Extraction) feature extractor [49]. The SMILE automatic emotion recognition component [50] performs continuous emotion detection in terms of the emotion dimensions arousal and valence. The current models have been trained on preliminary Sensitive Artificial Listener data from the HUMAINE project [51]. The Automatic Speech Recognition component is based on the ATK (http://htk.eng.cam.ac.uk/develop/atk.shtml) real-time speech recognition engine which is a C++ layer sitting on top of HTK (http://htk.eng.cam.ac.uk/). The speaker independent triphone models were trained on the Wall Street Journal corpus [52] and on the AMIDA Meeting corpus [53]. The trigram language model was trained on the preliminary SAL corpus [51] and therefore includes about 1900 words which typically occur in spontaneous emotionally coloured speech. The module for recognition of human interest was trained on the AVIC database [54], a database which contains data of a real-world scenario where an experimenter leads a subject through a commercial presentation. The subject interacted with the experimenter and thereby naturally and spontaneously expressed different levels of interest. The module discriminates three different levels of interest: "bored", "neutral", and "high interest".

Turn taking is a complex conversational system in which the participants negotiate who will be the main speaker in the next period. The SEMAINE system 1.0 implements a simplistic mechanism: when the user is silent for more than 2 seconds, the system decides that the agent has the turn.

When the agent receives the turn, the system will analyse what the user did and said. The User utterance interpreter will look at the utterances of the user that were detected in the previous turn. The utterances are tagged with general semantic features such as the semantic polarity, the time, and the subject of the utterances, and the user state is updated accordingly.

The function of the agent utterance proposer is to select an appropriate response when the agent has to say something. It starts working when it receives an extended user utterance from the user utterance interpreter, because in the current system this also means that the agent has the turn. Using the added features, it searches its response model for responses that fit the current context. This response model is based on the Sensitive Artificial Listener script [55] and contains fitting contexts (i.e., a list of semantic features) for every possible response. When no information about the previous user utterance is available, it will pick a response from a list of generic responses which fit in almost all circumstances.

A backchannel proposer (not shown in Figure 2) can be used to generate behaviour such as the nods and "uh-huh" sounds produced by listeners during the other's turn. The current simplistic implementation triggers a nonspecific backchannel after 300 ms of silence by the user.

In the case of conflicting action proposals, an action selection component is needed to decide which actions to

```
1 public class HelloInput extends Component {
2
3   private Sender textSender = new Sender("semaine.data.hello.text", "TEXT", getName());;
4   private BufferedReader inputReader = new BufferedReader(new InputStreamReader(System.in));
5
6   public HelloInput() throws JMSException {
7     super("HelloInput", true/*is input*/, false);
8     senders.add(textSender);
9   }
10
11  @Override protected void act() throws IOException, JMSException {
12    if (inputReader.ready()) {
13      String line = inputReader.readLine();
14      textSender.sendTextMessage(line, meta.getTime());
15    }
16  }
17 }
```

Figure 12: The HelloInput component sending text messages via the SEMAINE API.

realise. The implementation in the SEMAINE system 1.0 is a dummy implementation which simply accepts all proposed actions.

The speech preprocessing component is part of the conceptual behaviour planner component. It uses the MARY TTS system [46] to add pitch accent and phrase boundary information to FML documents in preparation of the realisation of functions in terms of visual behaviour.

Conceptually, the visual behaviour planner component identifies behaviour patterns that are appropriate for realising the functions contained in an FML document. At this stage, the component is a dummy implementation only which does nothing.

The speech synthesis component is part of the conceptual behaviour realiser component. It uses the MARY TTS system [46] to produce synthetic speech audio data as well as timing information in an extended BML document suitable as input to the visual behaviour realiser. As a proof of concept, the current version of the speech synthesiser also generates vocal backchannels upon request.

The visual behaviour realiser component generates the animation for the Greta agent [9] as MPEG-4 Facial Animation Parameters (FAP) [56] and Body Animation Parameters (BAP). The input of the module is specified by the BML language. It contains the text to be spoken and/or a set of nonverbal signals to be displayed. The list of phonemes and their respective duration, provided by the speech synthesis component, are used to compute the lips movements.

When the Behaviour Realiser receives no input, the agent does not remain still. It generates some idle movements whenever it does not receive any input. Periodically a piece of animation is computed and is sent to the player. It avoids unnatural "freezing" of the agent.

The Greta player [9] receives the animation generated by the behaviour realiser and plays it in a graphic window. The animation is defined by the Facial Animation Parameters (FAPs) and the Body Animation Parameters (BAPs). Each FAP or BAP frame received by the player carries also the time intended for its visualisation as computed by the behaviour realiser.

4.3. System Properties. The combination of the system components described above enables a simple kind of dialogue interaction. While the user is speaking, audio features are extracted. When silence is detected, estimates of the user's emotion and interest during the turn are computed, and the ASR produces an estimate of the words spoken. When the silence duration exceeds a threshold, backchannels are triggered (if the system is configured to include the backchannel proposer component); after a longer silence, the agent takes the turn and proposes a verbal utterance from the SAL script [55]. Even where no meaningful analysis of the user input can be made, the script will propose a generic utterance such as "Why?" or "Go on." which is suitable in many contexts. The utterances are realised with a generic TTS voice and rendered, using either the audiovisual Greta player or an audio-only player.

This description shows that the system is not yet capable of much meaningful interaction, since its perceptual components are limited and the dialogue model is not fully fleshed out yet. Nevertheless, the main types of components needed for an emotion-aware interactive system are present, including emotion analysis from user input, central processing, and multimodal output. This makes the system suitable for experimenting with emotion-aware systems in various configurations, as the following section will illustrate.

5. Building Emotion-Oriented Systems with the SEMAINE API

This section presents three emotion-oriented example systems, in order to corroborate the claim that the SEMAINE API is easy to use for building new emotion-oriented systems out of new and/or existing components. Source code is provided in order to allow the reader to follow in detail the steps needed for using the SEMAINE API. The code is written in Java and can be obtained from the SEMAINE sourceforge page [57]. The SEMAINE API parts of the code would look very similar in C++.

5.1. Hello World. The "Hello" example realises a simple text-based interactive system. The user types arbitrary text; an analyser component spots keywords and deduces an affective state from them; and a rendering component outputs an emoticon corresponding to this text. Despite its simplicity,

Table 3: Ad hoc emoticons used to represent positions in the arousal-valence plane.

		Valence		
		—	**0**	+
Arousal	+	8-(	8-|	8-)
	0	:-(	:-|	:-)
	—	*-(	*-|	*-)

the example is instructive because it displays the main elements of an emotion-oriented system.

The input component (Figure 12) simply reads one line of text at a time and sends it on. It has an input device (Figure 12, line 4) and a Sender writing TEXT data to the Topic `semaine.data.hello.text` (line 3). In its constructor, the component registers itself as an input component (l. 7) and registers its sender (l. 8). Its `act()` method, which is automatically called every 100 ms while the system is running, checks for new input (l. 12), reads it (l. 13), and sends it to the Topic (l. 14).

As a simplistic central processing component, the HelloAnalyser (Figure 13) makes emotional judgements about the input. It registers a Receiver (l. 7) for the Topic that HelloInput writes to and sets up (l. 3) and registers (l. 8) an XML Sender producing data of type EmotionML. Whenever a message is received, the method react() is called (l. 11). It receives (l. 13) and analyses (l. 14–17) the input text and computes values for the emotion dimensions arousal and valence from the text. Finally, it creates an EmotionML document (l. 18) and sends it (l. 19).

As the SEMAINE API does not yet provide built-in support for standalone EmotionML documents, the component uses a generic `XMLSender` (l. 3) and uses the `XMLTool` to build up the EmotionML document (l. 23–30).

The output of the Hello system should be an emoticon representing an area in the arousal-valence plane as shown in Table 3. The EmoticonOutput component (Figure 14) registers an XML Receiver (l. 5) to the Topic that the HelloAnalyser sends to. Whenever a message is received, the `react()` method is called (l. 8), which analyses the XML document in terms of EmotionML markup (l. 10–12) and extracts the arousal and valence values (l. 14-15). The emotion display is rendered as a function of these values (l. 17–19).

In order to build a system from the components, a configuration file is created (Figure 15). It includes the SystemManager component as well as the three newly created components. Furthermore, it requests a visible system manager GUI providing a message flow graph.

The system is started in the same way as all Java-based SEMAINE API systems: `activemq; java eu.semaine.system.ComponentRunner example-hello.config`. Figure 16 shows a screenshot of the resulting message flow graph. As the communication passes via the middleware ActiveMQ, the system would behave in the exact same way if the four components were started as separate processes, on different machines, or if some of them were written in C++ rather than Java.

5.2. Emotion Mirror. The Emotion mirror is a variant of the Hello system. Instead of analysing text and deducing emotions from keywords, it uses the openSMILE speech feature extraction and emotion detection (see Section 4.2) for interpreting the user's emotion. The output is rendered using the same EmoticonOutput component from the Hello system in Section 5.1.

Only one new component is needed to build this system. EmotionExtractor (Figure 17) has an emotion Sender (l. 2 and l. 7) just like the one HelloAnalyser had, but uses an EMMA Receiver (l. 6) to read from the topic that the Emotion detection component from the SEMAINE system (see Section 4.2) publishes to, as documented in [47]. Upon reception of an EMMA message, the method `react()` is called (l. 10). As the only receiver registered by the component is an EMMA receiver, the message can be directly cast into an EMMA message (l. 11) which allows for comfortable access to the document structure to extract emotion markup (l. 12-13). Where emotion markup is present, it is inserted into a standalone EmotionML document (l. 16–18) and sent to the output Topic (l. 19).

The configuration file contains only the components SystemManager, EmotionExtractor, and EmoticonOutput. As the SMILE component is written in C++, it needs to be started as a separate process as documented in the SEMAINE wiki documentation [58]. The resulting message flow graph is shown in Figure 18.

5.3. A Game Driven by Emotional Speech: The Swimmer's Game. The third example system is a simple game application in which the user must use emotional speech to win the game. The game scenario is as follows. A swimmer is being pulled backwards by the stream towards a waterfall (Figure 19). The user can help the swimmer to move forward towards the river bank by cheering him up through high-arousal speech. Low arousal, on the other hand, discourages the swimmer and drives him more quickly to the waterfall.

The system requires the openSMILE components as in the Emotion mirror system, a component computing the swimmer's position as time passes and considering the user's input, and a rendering component for the user interface. Furthermore, we will illustrate the use of TTS output in the SEMAINE API by implementing a commentator providing input to the speech synthesis component of the SEMAINE system 1.0 (Section 4.2).

The PositionComputer (Figure 20) combines a `react()` and an `act()` method. Messages are received via an EMMA receiver and lead to a change in the internal parameter `position` (l. 22). The `act()` method implements the backward drift (l. 29) and sends regular position updates (l. 30) as a plain-text message.

The SwimmerDisplay (Figure 21) implements the user interface shown in Figure 19. Its messaging part consist of a simple text-based Receiver (l. 5) and an interpretation of the text messages as single float values (l. 10).

Due to the separation of position computer and swimmer display, it is now very simple to add a Commentator

```
1 public class HelloAnalyser extends Component {
2
3   private XMLSender emotionSender =
        new XMLSender("semaine.data.hello.emotion", "EmotionML", getName());
4
5   public HelloAnalyser() throws JMSException {
6     super("HelloAnalyser");
7     receivers.add(new Receiver("semaine.data.hello.text"));
8     senders.add(emotionSender);
9   }
10
11  @Override protected void react(SEMAINEMessage m) throws JMSException {
12    int arousalValue = 0, valenceValue = 0;
13    String input = m.getText();
14    if (input.contains("very")) arousalValue = 1;
15    else if (input.contains("a bit")) arousalValue = -1;
16    if (input.contains("happy")) valenceValue = 1;
17    else if (input.contains("sad")) valenceValue = -1;
18    Document emotionML = createEmotionML(arousalValue, valenceValue);
19    emotionSender.sendXML(emotionML, meta.getTime());
20  }
21
22  private Document createEmotionML(int arousalValue, int valenceValue) {
23    Document emotionML = XMLTool.newDocument(EmotionML.ROOT_ELEMENT, EmotionML.namespaceURI);
24    Element emotion = XMLTool.appendChildElement(emotionML.getDocumentElement(),
                                                   EmotionML.E_EMOTION);
25    Element dimensions = XMLTool.appendChildElement(emotion, EmotionML.E_DIMENSIONS);
26    dimensions.setAttribute(EmotionML.A_SET, "arousalValence");
27    Element arousal = XMLTool.appendChildElement(dimensions, EmotionML.E_AROUSAL);
28    arousal.setAttribute(EmotionML.A_VALUE, String.valueOf(arousalValue));
29    Element valence = XMLTool.appendChildElement(dimensions, EmotionML.E_VALENCE);
30    valence.setAttribute(EmotionML.A_VALUE, String.valueOf(valenceValue));
31    return emotionML;
32  }
33 }
```

Figure 13: The HelloAnalyser component. It receives and analyses the text messages from HelloInput and generates and sends an EmotionML document containing the analysis results.

```
1 public class EmoticonOutput extends Component {
2
3   public EmoticonOutput() throws JMSException {
4     super("EmoticonOutput", false, true /*is output*/);
5     receivers.add(new XMLReceiver("semaine.data.hello.emotion"));
6   }
7
8   @Override protected void react(SEMAINEMessage m) throws MessageFormatException {
9     SEMAINEXMLMessage xm = (SEMAINEXMLMessage) m;
10    Element dimensions = (Element) xm.getDocument().getElementsByTagNameNS(
                             EmotionML.namespaceURI, EmotionML.E_DIMENSIONS).item(0);
11    Element arousal = XMLTool.needChildElementByTagNameNS(dimensions, EmotionML.E_AROUSAL,
                             EmotionML.namespaceURI);
12    Element valence = XMLTool.needChildElementByTagNameNS(dimensions, EmotionML.E_VALENCE,
                             EmotionML.namespaceURI);
13
14    float a = Float.parseFloat(arousal.getAttribute(EmotionML.A_VALUE));
15    float v = Float.parseFloat(valence.getAttribute(EmotionML.A_VALUE));
16
17    String eyes  = a > 0.3 ? "8"/*active*/   : a < -0.3 ? "*"/*passive*/  : ":"/*neutral*/;
18    String mouth = v > 0.3 ? ")"/*positive*/ : v < -0.3 ? "("/*negative*/ : "|"/*neutral*/;
19    System.out.println(eyes+"-"+mouth);
20  }
21 }
```

Figure 14: The EmoticonOutput component. It receives EmotionML markup and displays an emoticon according to Table 3.

```
semaine.components = \
    |eu.semaine.components.meta.SystemManager| \
    |eu.semaine.examples.hello.HelloInput| \
    |eu.semaine.examples.hello.HelloAnalyser| \
    |eu.semaine.examples.hello.EmoticonOutput|

semaine.systemmanager.gui = true
```

Figure 15: The configuration file `example-hello.config` defining the Hello application.

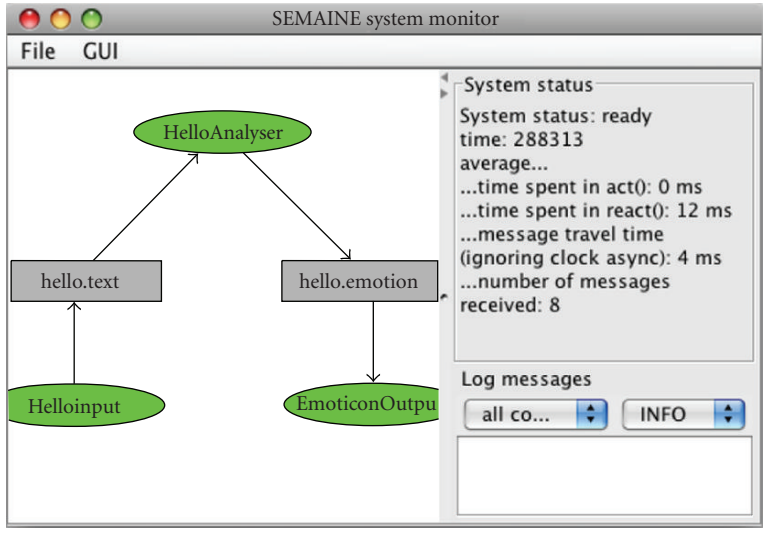


Figure 16: Message flow graph of the Hello system.

```
1 public class EmotionExtractor extends Component {
2   private XMLSender emotionSender =
        new XMLSender("semaine.data.hello.emotion", "EmotionML", getName());
3
4   public EmotionExtractor() throws JMSException {
5     super("EmotionExtractor");
6     receivers.add(new EmmaReceiver("semaine.data.state.user.emma"));
7     senders.add(emotionSender);
8   }
9
10  @Override protected void react(SEMAINEMessage m) throws JMSException {
11    SEMAINEEmmaMessage emmaMessage = (SEMAINEEmmaMessage) m;
12    Element interpretation = emmaMessage.getTopLevelInterpretation();
13    List<Element> emotionElements = emmaMessage.getEmotionElements(interpretation);
14    if (emotionElements.size() > 0) {
15      Element emotion = emotionElements.get(0);
16      Document emotionML = XMLTool.newDocument(EmotionML.ROOT_ELEMENT, EmotionML.namespaceURI);
17      emotionML.adoptNode(emotion);
18      emotionML.getDocumentElement().appendChild(emotion);
19      emotionSender.sendXML(emotionML, meta.getTime());
20    }
21  }
22 }
```

Figure 17: The EmotionExtractor component takes EmotionML markup from an EMMA message and forwards it.

component (Figure 22) that generates comments using synthetic speech, as a function of the current position of the swimmer. It subscribes to the same Topic as the SwimmerDisplay (l. 7) and sends BML output (l. 2) to the Topic serving as input to the speech synthesis component of the SEMAINE system 1.0 [47]. Speech output is produced when the game starts (l. 18–20) and when the position meets certain criteria (l. 13-14). Generation of speech output consists in the creation of a simple BML document with a <speech> tag enclosing the text to be spoken (l. 25–28) and sending that document (l. 29).

The complete system consists of the Java components SystemManager, PositionComputer, SwimmerDisplay, Commentator, SpeechBMLRealiser, and SemaineAudioPlayer, as well as the external C++ component openSMILE. The resulting message flow graph is shown in Figure 23.

6. Evaluation

One important aspect in a middleware framework is message routing time. We compared the MOM ActiveMQ, used in the SEMAINE API, with an alternative system, Psyclone [59], which is used in systems similar to ours (e.g., [60]). In order to compute the mere message routing time ignoring network latencies, we ran both ActiveMQ 5.1.0 and Psyclone 1.1.7 on a Windows Vista machine with a Core2Duo 2.5 GHz processor and 4 GB RAM with no other load and connected to each using a Java client sending and receiving messages in sequence from the localhost machine. We sent text messages of different lengths to each middleware in a loop, averaging measures over 100 repetitions for each message length. We used plain string messages with lengths between 10 and 1 000 000 characters. The message routing times are shown

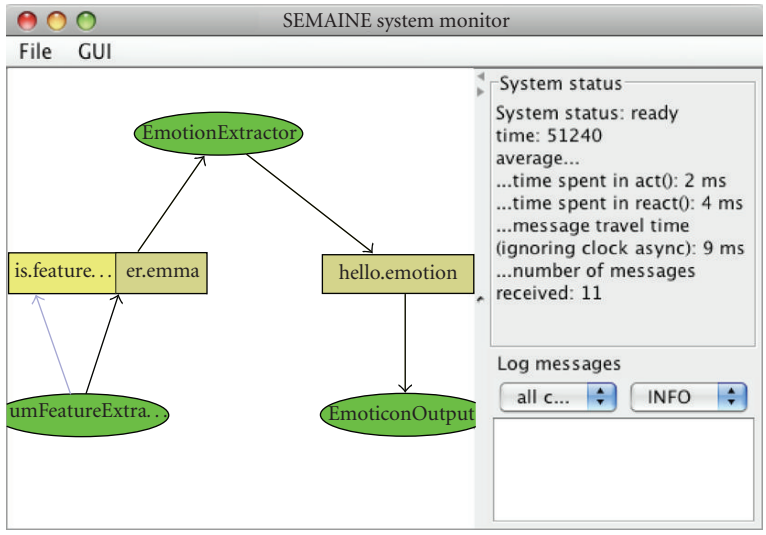


Figure 18: Message flow graph of the Emotion mirror system.

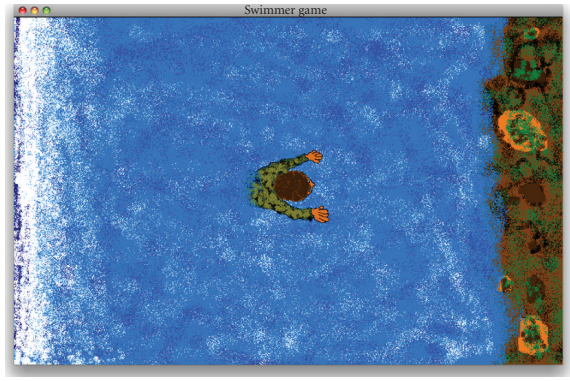


Figure 19: Swimmer's game user interface.

in Figure 24. Between 10 and 1000 characters, round trip message routing times for ActiveMQ are approximately constant at around 0.3 ms; the times rise to 0.5 ms for 10,000 characters, 2.9 ms for 100 000, and 55 ms for messages of 1 000 000 characters length. Psyclone is substantially slower, with routing times approximately constant around 16 ms for messages from 10 to 10 000 characters length, then rising to 41 ms at 100 000 characters length and 408 ms at 1 000 000 characters message length.

These results show that in this task ActiveMQ is approximately 50 times faster than Psyclone for short messages and around 10 times faster for long messages. While it may be possible to find even faster systems, it seems that ActiveMQ is reasonably fast for our purposes.

Other evaluation criteria are more difficult to measure. While it is an aim of the SEMAINE API to be easy to use for developers, time will have to tell whether the system is being embraced by the community. A first piece of positive evidence is the adoption of the SEMAINE API for a real-time animation engine [61].

One aspect that should be added to the current SEMAINE API when representation formats settle is the validation of representation formats per Topic. Using XML schema, it is possible to validate that any message sent via a given Topic respects a formally defined syntax definition for that Topic. At the time of developing and debugging a system, this feature would help identify problems. At run time, the validation could be switched off to avoid the additional processing time required for XML validation.

7. Availability

The SEMAINE API and the SEMAINE system 1.0 are available as open source [57, 58]. The API is covered by the GNU Lesser General Public License (LGPL) [62], which can be combined with both closed-source and open source components. The components of the system 1.0 are partly released under the LGPL, partly under the more restrictive GNU General Public License (GPL) [63], which prohibits the proliferation together with closed-source components.

```
1 public class PositionComputer extends Component {
2   private Sender positionSender =
        new Sender("semaine.data.swimmer.position", "TEXT", getName());
3   private float position = 50;
4
5   public PositionComputer() throws JMSException {
6     super("PositionComputer");
7     receivers.add(new EmmaReceiver("semaine.data.state.user.emma"));
8     senders.add(positionSender);
9   }
10
11  @Override protected void react(SEMAINEMessage m) throws MessageFormatException {
12    SEMAINEEmmaMessage emmaMessage = (SEMAINEEmmaMessage) m;
13    Element interpretation = emmaMessage.getTopLevelInterpretation();
14    List<Element> emotionElements = emmaMessage.getEmotionElements(interpretation);
15
16    for (Element emotion : emotionElements) {
17      Element dimensions = XMLTool.getChildElementByTagNameNS(emotion, EmotionML.E_DIMENSIONS,
                                EmotionML.namespaceURI);
18      if (dimensions != null) {
19        Element arousal = XMLTool.needChildElementByTagNameNS(dimensions, EmotionML.E_AROUSAL,
                                EmotionML.namespaceURI);
20        float arousalValue = Float.parseFloat(arousal.getAttribute(EmotionML.A_VALUE));
21        // Arousal influences the swimmer's position:
22        position += 10*arousalValue;
23      }
24    }
25  }
26
27  @Override protected void act() throws JMSException {
28    // The river slowly pulls back the swimmer:
29    position -= 0.1;
30    positionSender.sendTextMessage(String.valueOf(position), meta.getTime());
31  }
32 }
```

Figure 20: The PositionComputer component.

```
1 public class SwimmerDisplay extends Component {
2
3   public SwimmerDisplay() throws JMSException {
4     super("SwimmerDisplay", false, true/*is output*/);
5     receivers.add(new Receiver("semaine.data.swimmer.position"));
6     setupGUI();
7   }
8
9   @Override protected void react(SEMAINEMessage m) throws JMSException {
10    float percent = Float.parseFloat(m.getText());
11    updateSwimmerPosition(percent);
12    String message = percent <= 0 ? "You lost!" : percent >= 100 ? "You won!!!" : null;
13    if (message != null) {
        ...
      }
    }
    ...
  }
```

Figure 21: The SwimmerDisplay component (GUI code not shown).

The examples in Section 5 are available from the SEMAINE sourceforge page [57] as an add-on to the SEMAINE system 1.0.

8. Conclusion

This paper has presented the SEMAINE API as a framework for enabling the creation of simple or complex emotion-oriented systems with limited effort. The framework is rooted in the understanding that the use of standard formats is beneficial for interoperability and reuse of components. The paper has illustrated how system integration and reuse of components can work in practice.

More work is needed in order to make the SEMAINE API fully suitable for a broad range of applications in the area of emotion-aware systems. Notably, the support of representation formats needs to be completed. Moreover, several crucial representation formats are not yet fully specified, including EmotionML, BML and FML. Agreement on these specifications can result from an ongoing consolidation process in the community. If several research teams were to bring their work into a common technological framework, this would be likely to speed up the consolidation process, because challenges to integration would become apparent more quickly. An open source framework such as the SEMAINE API may be suited for such an endeavour.

```
1 public class Commentator extends Component {
2   private BMLSender bmlSender = new BMLSender("semaine.data.synthesis.plan", getName());
3   private boolean started = false;
4
5   public Commentator() throws JMSException {
6     super("Commentator");
7     receivers.add(new Receiver("semaine.data.swimmer.position"));
8     senders.add(bmlSender);
9   }
10
11  @Override protected void react(SEMAINEMessage m) throws JMSException {
12    float percent = Float.valueOf(m.getText());
13    if (percent < 30 /*danger*/) say("Your swimmer needs help!");
14    else if (percent > 70 /*nearly there*/) say("Just a little more.");
15  }
16
17  @Override protected void act() throws JMSException {
18    if (!started) {
19      started = true;
20      say("The swimmer needs your support to reach the river bank. Cheer him up!");
21    }
22  }
23
24  private void say(String text) throws JMSException {
25    Document bml = XMLTool.newDocument(BML.ROOT_TAGNAME, BML.namespaceURI);
26    Element speech = XMLTool.appendChildElement(bml.getDocumentElement(), BML.E_SPEECH);
27    speech.setAttribute("language", "en-US");
28    speech.setTextContent(text);
29    bmlSender.sendXML(bml, meta.getTime());
30  }
31 }
```

Figure 22: The Commentator component, producing TTS requests.

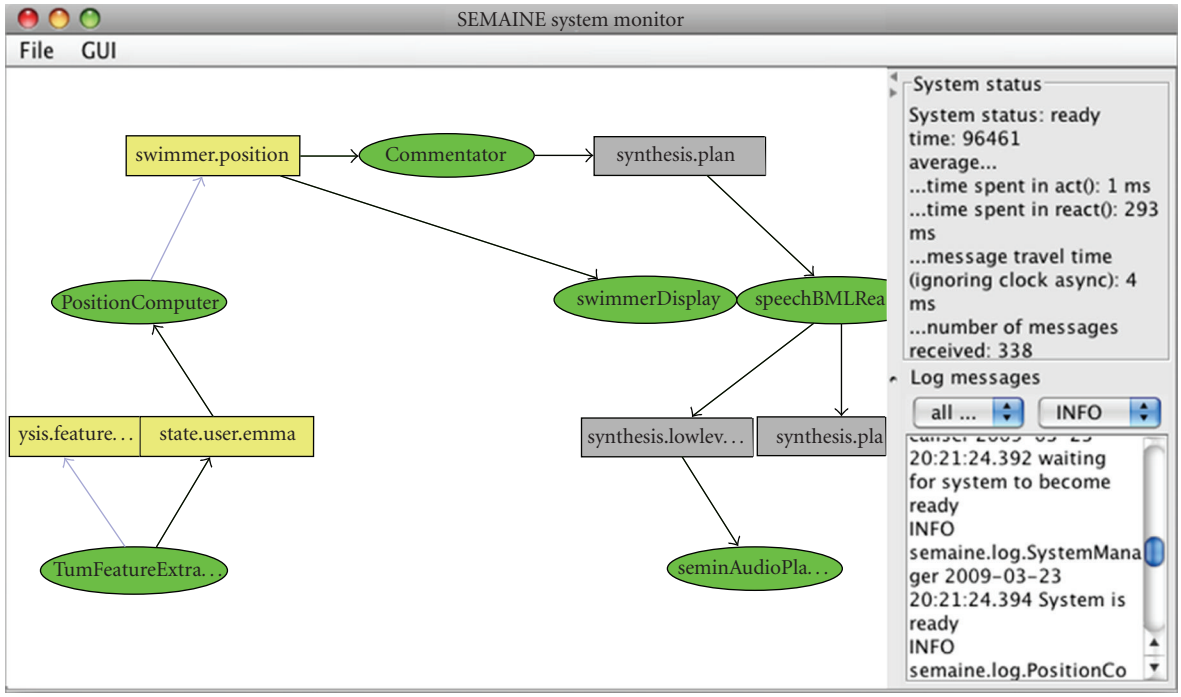


Figure 23: Message flow graph of the swimmer's game system.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under Grant agreement no. 211486 (SEMAINE). The work presented here has been shaped by discussions about concepts and implementation issues with many people, including Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Sathish Pammi, Maja Pantic, Catherine

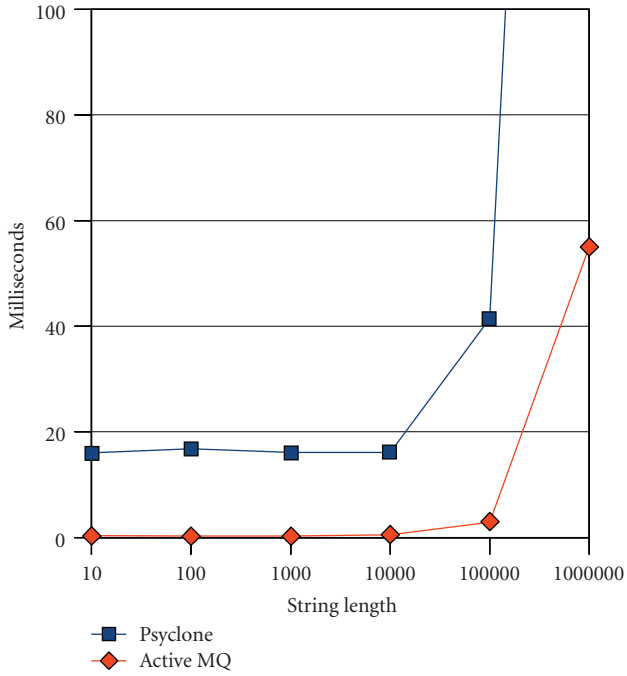


Figure 24: Round-trip message routing times as a function of message length.

Pelachaud, Björn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wöllmer. Thanks to Jonathan Gratch who pointed us to ActiveMQ in the first place. Thanks also to Oliver Wenz for designing the graphics of the swimmer's game.

References

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," in *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI '07)*, pp. 126–133, ACM, Nagoya, Japan, 2007.

[2] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[3] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," *Neural Networks*, vol. 18, no. 4, pp. 423–435, 2005.

[4] A. Batliner, S. Steidl, B. Schuller, et al., "Combining efforts for improving automatic classification of emotional user states," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, Ljubljana, Slovenia, 2006.

[5] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 4, pp. 941–944, 2007.

[6] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," *Interacting with Computers*, vol. 18, no. 2, pp. 139–170, 2006.

[7] P. Gebhard, "ALMA—a layered model of affect," in *Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '05)*, pp. 177–184, Utrecht, The Netherlands, 2005.

[8] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie, "Emotion recognition and synthesis based on MPEG-4 FAPs," in *MPEG-4 Facial Animation—The Standard, Implementations, Applications*, I. S. Pandzic and R. Forchheimer, Eds., John Wiley & Sons, Hillsdale, NJ, USA, 2002.

[9] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud, "An expressive ECA showing complex emotions," in *Proceedings of the AISB Annual Convention*, pp. 208–216, Newcastle, UK, 2007.

[10] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant synthesis," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 151–156, Newcastle, UK, 2000.

[11] M. Schröder, "Approaches to emotional expressivity in synthetic speech," in *The Emotion in the Human Voice*, K. Izdebski, Ed., Plural, San Diego, Calif, USA, 2008.

[12] G. Castellano, R. Bresin, A. Camurri, and G. Volpe, "Expressive control of music and visual media by full-body movement," in *Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, pp. 390–391, ACM, New York, NY, USA, 2007.

[13] ISO—International Organization for Standardization, "ISO 261: ISO general purpose metric screw threads—general plan," 1998, http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=4165.

[14] ISO—International Organization for Standardization, "ISO/IEC 26300:2006: Information technology—Open Document Format for Office Applications (OpenDocument) v1.0," 2006, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485.

[15] D. Raggett, A. Le Hors, and I. Jacobs, *HTML 4.01 Specification*, 1999, http://www.w3.org/TR/html401/.

[16] K. van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann, "Fully generated scripted dialogue for embodied agents," *Artificial Intelligence*, vol. 172, no. 10, pp. 1219–1244, 2008.

[17] P. Piwek, B. Krenn, M. Schröder, M. Grice, S. Baumann, and H. Pirker, "RRL: a rich representation language for the description of agent behaviour in NECA," in *Proceedings of the AAMAS Workshop Conversational Agents*, Bologna, Italy, 2002.

[18] B. Kempe, N. Pfleger, and M. Löckelt, "Generating verbal and nonverbal utterances for virtual characters," in *Proceedings of the 3rd International Conference on Virtual Storytelling (ICVS '05)*, vol. 3805 of *Lecture Notes in Computer Science*, pp. 73–76, 2005.

[19] M. Löckelt and N. Pfleger, "Multi-party interaction with self-contained virtual characters," in *Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue (DIALOR '05)*, pp. 139–142, Nancy, France, 2005.

[20] R. Aylett, A. Paiva, J. Dias, L. Hall, and S. Woods, "Affective agents for education against bullying," in *Affective Information Processing*, J. Tao and T. Tan, Eds., pp. 75–90, Springer, London, UK, 2009.

[21] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotion*, Cambridge University Press, Cambridge, UK, 1988.

[22] P. Gebhard, M. Schröder, M. Charfuelan, et al., "IDEAS4Games: building expressive virtual characters for computer games," in *Proceedings of the 8th International Conference on Intelligent Virtual Agents (IVA '08)*, vol. 5208 of *Lecture Notes in Computer Science*, pp. 426–440, Springer, Tokyo, Japan, 2008.

[23] Mystic Game Development, "EMotion FX," http://www.mysticgd.com/site2007/.

[24] Luxand, Inc., "Luxand—Detect Human Faces and Recognize Facial Features with Luxand FaceSDK," http://www.luxand.com/facesdk/.

[25] N. Dimakis, J. K. Soldatos, L. Polymenakos, P. Fleury, J. Curín, and J. Kleindienst, "Integrated development of context-aware applications in smart spaces," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 71–79, 2008.

[26] US National Institute of Standards and Technology (NIST), "NIST Data Flow System II," 2008, http://www.nist.gov/smartspace/sf_presentation.html.

[27] N. Hawes, J. L. Wyatt, A. Sloman, et al., "Architecture and representations," in *Cognitive Systems*, H. I. Christensen, A. Sloman, G. Kruijff, and J. Wyatt, Eds., pp. 53–95, 2009.

[28] M. Henning, *Choosing Middleware: Why Performance and Scalability do (and do not) Matter*, ZeroC, 2009, http://www.zeroc.com/articles/IcePerformanceWhitePaper.pdf.

[29] "Semaine Project," http://www.semaine-project.eu/.

[30] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[31] G. Banavar, T. Chandra, R. Strom, and D. Sturman, "A case for message oriented middleware," in *Proceedings of the 13th International Symposium on Distributed Computing (DISC '99)*, p. 846, 1999.

[32] "Apache ActiveMQ," http://activemq.apache.org/.

[33] M. Hapner, R. Burridge, R. Sharma, J. Fialli, and K. Stout, *Java Message Service (JMS) Specification Version 1.1*, Sun Microsystems, 2002, http://java.sun.com/products/jms/docs.html.

[34] "The Apache Xerces Project—xerces.apache.org," http://xerces.apache.org/.

[35] A. Le Hors, P. Le Hégaret, L. Wood, et al., *Document Object Model (DOM) Level 3 Core Specification*, 2004, http://www.w3.org/TR/DOM-Level-3-Core/.

[36] "Behavior Markup Language (BML) Wiki," 2008, http://wiki.mindmakers.org/projects:BML:main.

[37] M. Johnston, P. Baggia, D. C. Burnett, et al., "EMMA: Extensible MultiModal Annotation markup language," February 2009, http://www.w3.org/TR/emma/.

[38] D. C. Burnett, M. R. Walker, and A. Hunt, "Speech Synthesis Markup Language (SSML) Version 1.0," 2004, http://www.w3.org/TR/speech-synthesis/.

[39] M. Schröder, P. Baggia, F. Burkhardt, et al., *Elements of an EmotionML 1.0*, World Wide Web Consortium, 2008, http://www.w3.org/2005/Incubator/emotion/XGR-emotionml-20081120/.

[40] S. Kopp, B. Krenn, S. Marsella, et al., "Towards a common framework for multimodal generation: the behavior markup language," in *Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA '06)*, vol. 4133 of *Lecture Notes in Computer Science*, pp. 205–217, 2006.

[41] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjálmsson, "Why conversational agents do what they do functional representations for generating conversational agent behavior," in *Proceedings of the Workshop on Functional Markup Language at the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '08)*, Estoril, Portugal, 2008.

[42] D. Becket and B. McBride, *RDF/XML Syntax Specification (Revised)*, 2004, http://www.w3.org/TR/rdf-syntax-grammar/.

[43] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler, and F. Yergeau, *Extensible Markup Language (XML) 1.0 (Fifth Edition)*, 2008, http://www.w3.org/TR/xml/.

[44] M. Mancini and C. Pelachaud, "The FML-APML language," in *Proceedings of the Workshop on Functional Markup Language at the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '08)*, Estoril, Portugal, 2008.

[45] B. De Carolis, C. Pelachaud, I. Poggi, and M. Steedman, "APML, a markup language for believable behavior generation," in *Life-Like Characters*, H. Prendinger and M. Ishizuka, Eds., pp. 65–86, Springer, New York, NY, USA, 2004.

[46] M. Schröder, M. Charfuelan, S. Pammi, and O. Türk, "The MARY TTS entry in the Blizzard challenge 2008," in *Proceedings of the Blizzard Challenge*, Brisbane, Australia, 2008.

[47] M. Schröder, M. ter Maat, C. Pelachaud, et al., *SEMAINE deliverable D1b: 1st integrated system*, 2008, http://semaine.sourceforge.net/SEMAINE-1.0/D1b%20First%20integrated%20system.pdf.

[48] D. M. Wolpert and J. R. Flanagan, "Motor prediction," *Current Biology*, vol. 11, no. 18, pp. R729–R732, 2001.

[49] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *Proceedings of the Affective Computing and Intelligent Interaction*, IEEE, Amsterdam, The Netherlands, 2009.

[50] M. Wöllmer, F. Eyben, S. Reiter, et al., "Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, 2008.

[51] E. Douglas-Cowie, R. Cowie, I. Sneddon, et al., "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII '07)*, vol. 4738 of *Lecture Notes in Computer Science*, pp. 488–500, Lisbon, Portugal, September 2007.

[52] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*, pp. 357–362, Association for Computational Linguistics, Harriman, NY, USA, 1992.

[53] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[54] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, pp. 30–37, ACM, Nagoya, Japan, 2007.

[55] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC '08)*, pp. 1–4, Marrakech, Morocco, May 2008.

[56] J. Ostermann, "Face animation in MPEG-4," in *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, I. S. Pandzic and R. Forchheimer, Eds., pp. 17–55, John Wiley & Sons, London, UK, 2002.

[57] "SEMAINE sourceforge page," http://sourceforge.net/projects/semaine/.

[58] "SEMAINE-1.0 wiki documentation," http://semaine.opendfki.de/wiki/SEMAINE-1.0.

[59] CMLabs, "Psyclone," 2007, http://www.mindmakers.org/projects/Psyclone.

[60] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive ECA system," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, vol. 2, pp. 1399–1400, 2009.

[61] A. Heloir and M. Kipp, "EMBR—a realtime animation engine for interactive embodied agents," in *Proceedings of the 9th International Conference on Intelligent Virtual Agents (IVA '09)*, pp. 393–404, Springer, Amsterdam, The Netherlands, 2009.

[62] "GNU Lesser General Public License, version 3," http://www.gnu.org/licenses/lgpl.html.

[63] "GNU General Public License, version 3," http://www.gnu.org/licenses/gpl-3.0.html.

Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 782802, 15 pages
doi:10.1155/2010/782802

Research Article

Segmenting into Adequate Units for Automatic Recognition of Emotion-Related Episodes: A Speech-Based Approach

Anton Batliner,[1] Dino Seppi,[2] Stefan Steidl,[1] and Björn Schuller[3]

[1] *Pattern Recognition Laboratory, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), D-91058 Erlangen, Germany*
[2] *ESAT, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium*
[3] *Institute for Human-Machine Communication, Technische Universität München (TUM), D-80333 Munich, Germany*

Correspondence should be addressed to Anton Batliner, batliner@informatik.uni-erlangen.de

Received 1 April 2009; Accepted 12 December 2009

Academic Editor: Elisabeth Andre

We deal with the topic of segmenting emotion-related (emotional/affective) episodes into adequate units for analysis and automatic processing/classification—a topic that has not been addressed adequately so far. We concentrate on speech and illustrate promising approaches by using a database with children's emotional speech. We argue in favour of the word as basic unit and map sequences of words on both syntactic and "emotionally consistent" chunks and report classification performances for an exhaustive modelling of our data by mapping word-based paralinguistic emotion labels onto three classes representing valence (positive, neutral, negative), and onto a fourth rest (garbage) class.

1. Introduction

It is not only difficult to define "emotion," it is difficult as well to find out where an emotional episode—whatever it is—begins and where it ends. It is difficult both for theoretical reasons—in order to know where it begins and ends, we have to know what it is—and for methodological/practical reasons as well, which we will detail below. By and large, studies on emotion have bypassed this topic by dealing with episodes delimited by external criteria.

1.1. The Phenomena: Emotions or …. There is definitely no agreement on an extensional or intensional definition of "emotion"—or of any other term that could be used instead such as affect, attitude, and mood, to replace it or to denote similar phenomena that have to be told apart from the core semantics of this term. The core phenomena consist of the big *n* emotions such as despair, anger, joy—*n* being some figures between 4 and 8 or more; this concept is mainly rooted in psychology, and has been challenged, elaborated, and extended amongst others in cognitive psychology, for instance in the OCC model [1]. Perhaps "the" alternative concept is a wider definition, encompassing all the fringe phenomena that are "present in most of life but absent when people are emotionless" [2]; this is the concept of *pervasive emotion*, which often implicitly forms the basis of engineering approaches in a sometimes vague use of the term, addressing states such as interest, stress, and boredom.

These fiercely disputed terminological debates are, however, not relevant for our topic. Segmentation for dealing with such para-linguistic phenomena is pivotal—no matter which definition we use to describe them. They are related to but are not by definition coextensive with linguistic units such as sentences, utterances, dialogue acts, and salience. We will elaborate on this topic in the next subsection. To prevent fruitless debates, we use the rather vague term "emotion-related episodes" in the title to denote both emotions in a strict sense and related phenomena in a broader sense, which are found in the database our experiments are based on. In the text, we will often use "emotion" as the generic term, for better readability. This resembles the use of generic "he" instead of "he/she;" note, however, that in our context, it is not a matter of political correctness that might make a more cumbersome phrasing mandatory, it is only a matter of competing theoretical approaches, which are not the topic of the present paper.

Implicitly, it is normally taken for granted that these states to be modelled are produced and not only perceived.

This difference can be illustrated by the following example: a father can get really angry with his son, and this can be heard in his tone of voice and seen in the outcome of physiological measurements. He can, however, only pretend being angry—because as father, he has to, even if he perhaps likes his son's "misbehaviour." In such a case, we can hear the change in his tone of voice but most likely we will not be able to measure marked physiological changes. The son might notice such a "fake" emotion—if he is clever and has experienced it often enough—or not. We cannot assume that machines are clever enough to notice. Thus, the emotion-related states we are dealing with have to be taken as "perceived" *surface* phenomena, at face value—at least as long as we do not employ physiological measurements, trying to find out a real ground truth. (Strictly speaking, physiological measurements are most likely closer to—but not necessarily constituting—any ground truth.)

The components of speech are vocal expression and linguistic content. Both components can be employed for signalling denotations and semantics, and for constituting illocutions (such as dialogue acts), and for expressing connotations as paralinguistic messages (such as emotions). The same scenario as above can illustrate this usage: the father can get really angry with his son, but instead of expressing his anger in his tone of voice, he simply can say, in a low and calm voice: "*Now I'm really getting angry.*" It could be argued that this is a describing "meta" statement and not an indication of "real" anger. However, the son will be well advised to react as if the father had expressed "real" anger in his tone of voice as well. Moreover, it cannot be argued that this is not an indication of negative valence—note that in this paper, we map our raw labels onto main classes representing positive, neutral, or negative valence. Again, the son can take this at face value and stop his misbehaviour, or he can misconceive his father's anger as pretense because it is not expressed in the father's tone of voice. Again, machines should not try to be too clever; the only possibility they have is to take the linguistic content of the user's utterances at face value.

Thus, both vocal and linguistic expression of emotions should be taken by machines along the lines of Grice's cooperative principle, at face value, and not assuming any indirect use [3]; this excludes, for example, irony, metaphor, and meiosis. (It is sometimes claimed that irony can be recognised by a system; this will never work under real-life conditions, at least not in the foreseeable future.) In this vein, we will employ both acoustic and linguistic features for the automatic classification of emotion-related user states.

1.2. The Need for Segmentation. In this paper, we want to address different possibilities to segment emotional episodes. We will concentrate on speech but, at the same time, we want to argue that in many applications to be imagined, speech will possibly be the best modality to base segmentation upon; of course, this only holds for speech to be found together with other modalities. As has been noted in [4], for all modalities the segmentation into emotion units is one of the most important issues if we aim at real applications but has been "largely unexplored so far." The same holds for another, equally important aspect: even if most authors agree that it is high time to go over from acted data to realistic, spontaneous data, normally, only a subset of the full database is used consisting of somehow clear, that is, more or less prototypical cases. This is not only a clever move to push classification performance, it simply has grown out from the problem of class assignment in emotion processing: there is no simple and unequivocal ground truth. We have elaborated on the use of prototypes and their impact on classification performance in [5, 6].

In the transition from read speech to spontaneous speech in Automatic Speech Recognition (ASR), normally all data have been employed apart from, for instance, nonlinguistic vocalisations, which are treated as garbage; but they are still treated and not removed from the signal before processing. Note that a rough estimate for the difference between read and spontaneous data was that, at least at the beginning, one could expect an error rate for spontaneous data twice the size than the one for read data [7]. Of course, we cannot simply transfer this empirically obtained estimate onto Automatic Emotion Recognition (AER). Yet we definitely will have to deal with a plainly lower classification performance. However, this constitutes the last step before AER, including full ASR and automatic segmentation, really can be used "in the wild," that is, in real applications.

As mentioned in the last subsection, we use "emotion" in a broad sense, following the definition of "pervasive emotions" in the Network of Excellence HUMAINE [2], where emotion is defined as the absence of non-emotion. Thus, it is a foreground-background phenomenon: emotional has to be different from "not emotional", that is, from *neutral* (emotionally *idle*). However, in the four different modalities that have been mostly investigated in emotion studies—speech, vision (i.e., facial gestures), body posture and movements, and physiology—telling apart emotionally idle from emotionally active, that is, not-idle, poses different questions because the modalities themselves behave in a different way: speech versus non-speech is easy to tell apart—at least for a human being. Note that of course, such statements have to be taken with a grain of salt: depending on the signal-to-noise ratio, it can be difficult; telling apart a non-linguistic vocalisation from a linguistic one can be difficult, sometimes; and for a machine, it is more difficult than for a human listener. Yet to start speaking and to finish speaking is a voluntary act, which normally can be detected and delimited by the listener. However, you cannot start or finish your face—you always have it, even while asleep. And you might be able to control your physiological signals, but only up to a certain extent—you cannot stop your heart beats. (Interestingly, perception in these different modalities is different as well: you cannot stop hearing even while sleeping—you only can use some ear protection decreasing the noise—but you can stop looking by simply closing your eyes, while awake or while sleeping.)

If there was an unequivocal ground truth, at least for the reference data used in automatic processing, we could define begin and end of such episodes easily. However, there is none, irrespective of the modalities. Thus, we have to use and rely

on human annotations or on some external criteria; a well-known example for the latter is taking hanging-up the phone abruptly in a telephone human-machine communication as an indication of anger so we know that there has been some anger before—but we do not know yet whether and where it could be noticed. Arousal might be traced back in physiological signals by defining a threshold criterion based on, for example, Feeltrace annotations [8], but this is way more difficult for valence. Moreover, in many applications, physiological signals cannot be recorded.

We mentioned in the beginning that almost all studies have bypassed somehow the decision where to start or to end an emotion. For decades, the bulk of evidence came from acted data in the lab; for such data, beginning and end are given trivially: either some short episodes had to be produced, for example, using a semantically void utterance as carrier, or longer periods have been integrated. Even in a—more or less realistic—verbal human-human or human-machine communication, a dialogue act/move or an "utterance" can be delimited easily by the act of turn taking, when the one partner finishes speaking and the other partner takes over. As long as such a strategy works sufficiently well, there is no pressure to go over to other criteria. The longer such a unit is, however, the higher is the probability that it does not only entail one emotional episode but two or more, and that it is "smeared," that is, not unequivocal. We can compare this situation with dialogue acts: often, one dialogue move (turn) constitutes just one dialogue act. However, it can consist of sequences of dialogue acts as well; for instance, in appointment scheduling dialogues [9] often sequences of rejection (of the dialogue partner's suggestion), statement (of problems/facts), and suggestion (of alternative dates) can be found.

Several different subunits have been investigated as for their impact on improving classification performance such as frame-based processing, or taking some other fixed interval (percentage of whole utterance, n ms, or voiced/unvoiced decisions, just to mention the most important ones, cf. [10]). But all this has rather been independent from higher processing; yet in a full system such as SmartKom [11] or Semaine [12], time constraints make it mandatory not to wait with ASR and other processing modules until the speaker has finished his/her full turn. For a (close to) real-time processing, it might not matter much whether frames, or syllables, or words, or short chunks are processed; when we assume 1.5 real time, for a short chunk lasting 2 seconds, a user has to wait 3 seconds before a system answer has been generated. This can be tolerated. However, for a turn lasting 10 seconds, the user had to wait 15 seconds—which simply is far too long. Taking any unit below chunk level of course results in even shorter processing time.

We can suppose that emotional changes within a word are difficult to produce and therefore very rare. As a further advantage of word-based processing of emotional speech, we see the better dovetailing of emotion and speech processing. We do not have to align the "emotional time axis" in an additional step with Word Hypotheses Graphs (WHG); for instance, each word in a WHG can be annotated with either its individual emotion label or with the label that has been attributed to the higher unit this specific word belongs to. These are practical considerations; yet it might be plausible conceiving the *word* as the "smallest meaningful emotional unit" as well. We thus can speak of an "*ememe*" in analogy to the phoneme, the morpheme, and especially to the sememe and claim that normally, the word and the ememe are coextensive. A sememe consists of either a morpheme or a word indicating both semantic denotation and/or connotation either encoding a holistic meaning or being constituted by a feature bundle. We introduce the ememe as constituting "pure" connotation, indicated both by acoustic and linguistic means. Such a concept definitely makes sense from a practical point of view; thus, we do not have to care too much whether sometimes it might make sense to go over to subword units. Of course, this only holds for speech; there is no equivalent—at least no one that can be defined and segmented easily—in the other modalities. Note that our "ememe" is the *smallest* emotional unit. The same way as it makes more sense to process meaningful sequences of n words (sememes) constituting something like a syntactically meaningful chunk or a dialogue act, the same way it pays off to combine ememes into higher units. The charm of such an approach is that it is relatively easy to find a word, and where it begins and where it ends. In other modalities, it is way more difficult to delimit units.

In this paper, we want to pursue different emotion units based on speech. We will start with the word, and later combine words into syntactically/semantically meaningful chunks or into consistent "ememe sequences," that is, sequences of words belonging to the same emotion class. By that, we model two different approaches: in the one approach, emotion is sort of modelled as being part of linguistics, in the other one, emotion is an independent layer, in parallel to linguistics. The latter one might be more adequate for theoretical reasons—emotion is not (fully or only) part of linguistics. On the other hand, in a communication conveyed partly or mostly via speech, emotion might really be structured along the speech layer; for instance, the emotional load of content words is normally higher than the one of function words, and the "emotional message" might really be coextensive with dialogue acts. To give an example: laughter is a non-linguistic indication of emotions/user states and often co-occurs with joy. It can be stand alone or modulated onto speech (speech laughter). Laughter and speech laughter are mainly found at the end of syntactic units. This does not necessarily mean that laughter and speech/language are processed and generated in the same module, but it demonstrates a close relationship. Moreover, in an end-to-end system, we always need to align emotion and linguistic processing somehow. However, in this paper, we can only deal with performance measures such as classification rates as criteria. Note that we will not deal with data, use cases, or applications without speech. If we take into account more than one modality we always have to align the unit of one modality with the unit(s) found in the other modality/modalities. Of course, this can be done with some criteria for overlapping on the time axis. At least for the time being, it seems to us that speech, if available, is advantageous over the other modalities to start with.

This is, of course, an assumption that has to be validated or falsified. We can imagine that researchers working in other modalities but speech prefer having their own units of analysis and late fusion of channels [13]. From a theoretical point of view, this might be easier to accomplish; from a practical point of view, it will be a matter of performance, of ease of handling, and—perhaps most important—of the weight a modality has in specific scenarios. Thus, a fall-back solution for our approach is, of course, to use it within a uni-modal speech scenario. Time synchronous or adjacent emotional messages conveyed by different modalities can be congruent or incongruent; speech can even distort the emotional message conveyed via facial gesture or bio-signals because of lip and jaw movements. Moreover, we have to tell apart different types of systems: on the one hand, there are end-to-end systems that take into account emotions as a way of "colouring" the intended message, triggering decisions on part of the dialogue manager in, for instance, call-centre interactions. Here we find a high functional load on speech. On the other hand, there are pure "emotion systems" with a low functional load on speech, for instance in video games—here, "non-verbal" grunts and affect bursts might be more relevant, together with facial gestures.

1.3. Overview. In Section 2, we present the database and the annotations performed, as well as the mapping onto main classes used in this paper. Section 3 presents the units of analysis we want to deal with: we start with the word as basic unit, and then discuss two different types of units, one based on syntactic criteria—to be dovetailed with higher processing modules such as dialogue act processing, the other one simply based on "emotional consistency;" adjacent words belonging to the same class are aggregated within the same unit. In Section 4, we describe the acoustic and linguistic features used in this study, as well as the classifier chosen for this task. Classification results are presented in Section 5 and discussed in Section 6. The paper closes with concluding remarks in Section 7.

2. Database and Annotation

The general frame for our FAU Aibo Emotion Corpus is human-robot communication, children's speech, and the elicitation and subsequent recognition of emotional user states. The robot is Sony's (dog-like) robot Aibo. The basic idea is to combine a so far rather neglected type of data (children's speech) with "natural" emotional speech within a Wizard-of-Oz task. The children were not told to use specific instructions but to talk to the Aibo like they would talk to a friend. They were led to believe that the Aibo is responding to their commands, but the robot is actually being controlled by a human operator, using the "Aibo Navigator" software over a wireless LAN (the existing Aibo speech recognition module is not used). The wizard causes the Aibo to perform a fixed, predetermined sequence of actions, which takes no account of what the child says. For the sequence of Aibo's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the Aibo was reacting to their orders—albeit often not immediately. In fact, it was the other way round: the Aibo always strictly followed the same screen-plot, and the children had to align their orders to its actions.

The data was collected from 51 children (age 10–13, 21 male, 30 female). The children were from two different schools, Mont and Ohm. The recordings took place in a classroom at each school. The child, the wizard, and two supervisors were present. The disjoint school recordings will be used to obtain a natural partitioning into train (Ohm) and test (Mont) in the ongoing. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz; quantisation is 16 bit. The data is downsampled to 16 kHz. Each recording session took some 30 minutes. The speech data were segmented automatically into speech files (turns), triggering a turn boundary at pauses ≥ 1 second. Note that here, the term "turn" does not imply any linguistic meaning; however, it turned out that only in very few cases, this criterion wrongly decided in favour of a turn boundary instead of (implicitly) modelling a hesitation pause. Because of the experimental setup, these recordings contain a huge amount of silence (reaction time of the Aibo), which caused a noticeable reduction of recorded speech after raw segmentation; finally we obtained about 8.9 hours of speech.

Five labellers (advanced students of linguistics with German as native language, four females, one male) listened to the speech files in sequential order and annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes, which were obtained by inspection of the data. This procedure was iterative and supervised by an expert. The sequential order of the labelling process does not distort the linguistic and paralinguistic message. Needless to say, we do not claim that these classes represent children's emotions (emotion-related user states) in general, only that they are adequate for the modelling of these children's behaviour in this specific scenario. We resort to majority voting (henceforth MV): if three or more labellers agree, the label is attributed to the word; if four or five labellers agree, we assume some sort of prototypes. The following raw labels were used; in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, that is, irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, that is, non-neutral, but not belonging to the other categories (3), *neutral* (39169); 4707 words had no MV; all in all, there are 48401 words. *Joyful* and *angry* belong to the "big" emotions, the other ones rather to "emotion-related/emotion-prone" user states but have been listed in more extensive catalogues of emotion/emotion-related terms, for example, "reproach" (i.e., *reprimanding*), *bored*, or *surprised* in [1]. The state *emphatic* has been introduced because it can be seen as a possible indication of some (starting) trouble in communication and by that, as a sort of "pre-emotional," negative state [5, 14, 15]; note that all the states, especially *emphatic*, have only been annotated when they differed from the (initial) baseline of the speaker.

Table 1: Emotion classes and their word-based frequencies.

Class	No. train	No. test	No. total	% total
P(ositive)	1110	299	1409	2.9
I(dle)	20471	18698	39169	80.9
R(est)	685	550	1235	2.5
N(egative)	3891	2697	6588	13.6
Total	26157	22244	48401	100.0

In this paper, we do not preselect a subcorpus out of the whole database but model valence, that is, positive, idle (neutral), and negative, for all the data; the remaining cases are attributed to a rest class. Thus, we map *motherese* and *joyful* onto *P(ositive)*, *neutral* onto *I(dle)*, *emphatic*, *touchy*, *reprimanding*, and *angry* onto *N(egative)*, and *surprised*, *helpless*, *bored*, and *rest* onto *R(est)*. Cases without an MV were mapped onto *R(est)* as well. The confusion matrices and the subsequent one- and two-dimensional plots based on Nonmetric Multidimensional Scaling (NMDS) in [14], both for labelling correspondences and for confusion matrices based on automatic classification, corroborate these mappings.

Table 1 displays the frequencies of these four classes; interestingly, *I(dle)* versus all other classes is Pareto distributed, that is, 80/20, as was the case for the emotion-related user states in the SmartKom [16] and in the AVIC [17] corpus as well.

Our database might seem to be atypical as it deals with children's speech; however, children represent just one of the usual partitions of the world's population into subgroups such as women/men, upper/lower class, or different dialects. Of course, automatic procedures have to adapt to this specific group—children's speech is a challenge for an Automatic Speech Recognition (ASR) system [18, 19], as both acoustic and linguistic characteristics differ from those of adults [20]. However, this necessity to adapt to a specific sub-group is a frequent issue in speech processing. Pitch, formant positions, and not yet fully developed co-articulation vary strongly, especially for younger children due to anatomical and physiological development [21]. Moreover, until the age of five/six, expression and emotion are strongly linked: children express their emotions even if no one else is present; the expression of emotion can be rather intense. Later on, expressions and emotions are decoupled [22] when children start to control their feelings. Thus so far, we found no indication that our children (age 10–13) behave differently from adults in a *principled* way, as far as speech/linguistics in general or emotional states conveyed via speech are concerned.

3. Units: Words, Syntactic Chunks, and Ememe Chunks

In this section, we present the three units we will deal with in the following: the word (ememe) as basic unit, and as higher units syntactic chunks (SCs) consisting of 1 to n words, and consistent sequences of ememes belonging to the same class, as well consisting of 1 to n words. The word (WO) is a straightforward unit, as is the ememe chunk (EC). There are many different syntactic theories yielding different representations of deep structure. However, we resort to a shallow, surface structure, thus, neutralising many of these differences. We are dealing not with syntactically well-formed speech but with spontaneous, natural speech—this will be the rule and not the exception if we aim at applications in real-life scenarios. For such data, no agreed-upon syntactic representation exists; thus, we have to establish one ourselves, based on the phenomena we can observe in our data.

3.1. Words: WO. Based on the orthographic transcription (transliteration) of the speech data, a lexicon has been compiled consisting of 1146 words, of which 333 are word fragments. Beginning and end of each word were presegmented automatically using a forced alignment of the spoken word chain, and eventually manually corrected. Throughout this paper, we will use this manual segmentation for extracting acoustic features. (Different smaller units of analysis for the FAU Aibo Emotion Corpus were pursued in [23].)

3.2. Syntactic Chunks: SC. Finding the appropriate unit of analysis for emotion recognition has not posed a problem in studies involving acted speech with different emotions, using segmentally identical utterances (cf. [24, 25]). In realistic data, a large variety of utterances can be found, from short commands in a well-defined dialogue setting, where the unit of analysis is obvious and identical to a dialogue move, to much longer utterances, and from syntactically well-defined units to all kinds of spontaneous phenomena such as elliptic speech and disfluencies [26]. In [27] it has been shown that in a Wizard-of-Oz-scenario (appointment scheduling dialogues), it is beneficial not to model whole turns but to divide them into smaller, syntactically and semantically meaningful chunks along the lines of [9]. Our scenario differs in one pivotal aspect from most of the other scenarios investigated so far; there is no real dialogue between the two partners; only the child is speaking, and the Aibo is only acting. Thus, it is not a "tidy" stimulus-response sequence that can be followed by tracking the very same channel; we are using only the recordings of the children's speech. Therefore, we do not know what the Aibo is doing at the corresponding time or has been doing shortly before or after the child's utterance. Moreover, the speaking style is rather special; there are not many "well-formed" utterances but a mixture of some long and many short sentences and one- or two-word utterances, which are often commands. The statistics of the observable turn lengths (in terms of the number of words) for the whole database is as follows: 1 word (2538 times), 2 words (2800 times), 3 words (2959 times), 4 words (2134 times), 5 words (1190 times), 6–9 words (1560 times), ≥10 words (461 times). We see that on the one hand, the threshold for segmentation of 1 s is meaningful; on the other hand, there are still many turns having more than 5 words per turn. This means that they tend to be longer than

one intonation unit, one clause, or one elementary dialogue act unit, which are common in this restricted setting "giving commands."

We observe neither "integrating" prosody as in the case of reading, nor "isolating" prosody as in the case of TV reporters. Many pauses of varying length are found, which can be hesitation pauses—the child produces slowly while observing the Aibo's actions—or pauses segmenting into different dialogue acts—the child waits until he/she reacts to the Aibo's actions. Thus, there is much overlap between two different channels: speech produced by the child, and vision, based on the Aibo's actions, which is not used for our annotation. We therefore decided in favour of hybrid syntactic-prosodic criteria: higher syntactic boundaries always trigger chunking; whereas lower syntactic boundaries do so only if the adjacent pause is ≥500 ms. By that, we try, for example, to tell apart vocatives that simply function as "relators" from vocatives with specific illocutive functions meaning, for example, "*Hi, I am talking to you*" or "*Now I am getting angry*" (illocution "command": "*Listen to me!*").

Note that in earlier studies, we found out that there is a rather strong correlation of more than 90% between prosodic boundaries, syntactic boundaries, and dialogue act boundaries [9]. Using only prosodic boundaries as chunk triggers results in (somehow) worse classification performance (in [9], some five percentage points lower). Moreover, from a practical point of view, it would be more cumbersome to time-align the different units—prosodic, that is, acoustic units, and linguistic, that is, syntactic or dialogue units, based on ASR and higher level segmentation—at a later stage in an end-to-end processing system, and to interpret the combination of these two different types of units accordingly. Preliminary experiments with chunks of different granularity, that is, length, showed that using our longer turns actually results in suboptimal classification performance, while the chunking procedure presented below, which was used for the experiments dealt with in this paper, results in better performance. This might partly result from the fact that more training instances are available, but partly as well from the fact that shorter units are more "consistent," [6].

The syntactic and pause labels are explained in Table 2. For this type of data, we could use a simplified version of the full set of syntactic-prosodic boundaries which is described in detail in [9], for both German and English. Chunk boundaries are triggered by higher syntactic boundaries after main clauses (s3) and after free phrases (p3), and boundaries between vocatives *Aibo Aibo* (v2v1) because here, the second *Aibo* is most likely not simply a relator but is conveying specific illocutions, as discussed above. Single instances of vocatives (v1, v2) are treated the same way as dislocations (d2). If the pauses at those lower syntactic boundaries given in Table 2 (s2, d2, v1, and v2) are at least 500 ms long, we insert a chunk boundary as well. s3 and s2 delimit syntactically "well-formed" clauses containing a verb; p3 characterises not-well-formed units, functioning like clauses but without a verb. d2 is annotated between clauses and some dislocated units to the left or to the right, which could have been integrated into the clause as well. Any longer pause at words within all these units was defined as a nontriggering hesitation pause. Each end-of-turn was redefined as triggering a clause/phrase boundary as well. Note that our turn-triggering threshold of 1 s works well because in the whole database, only 17 end-of-turn (<eot>) triggers were found that obviously denote within clause word boundaries as in the case of "*dieses Mal musst du nach* <eot> *links gehen* <eot>" (*this time you have to go to* <eot> *the left* <eot>). The boundary s1 had to be introduced because *komm* can function both as a sentence initial particle (corresponding to English *well, ...*) as well as an imperative (corresponding to English *come! ...*); only the imperative constitutes a clause.

Table 2: Syntactic and pause labels; frequencies given in Table 3.

label	Description
eot	End-of-turn, recoded as s3 (p3)
s3	Main clause/main clause
s2	Main/subord. clause or subord./subord. clause
s1	Sentence-initial particle or imperative "*komm*"
p3	Free phrases/particles
d2	Dislocations to the left/right
v2	Postvocative
v1	Praevocative
v2v1	Between "*Aibo*" instances
0	Pause 0–249 ms
1	Pause 250–499 ms
2	Pause 500–749 ms
3	Pause 750–1000 ms

With our simple and automatic threshold of 1 s between turns, we obtained turns as long as the one below, which we present for illustration. We denote the chunk triggering boundaries described in Tables 2 and 3 with the symbol "|". In Table 3, rows and columns with chunk triggering labels are shaded in grey. Between angle brackets, first the syntactic, and after the colon the pause labels are given; note that pause length has been corrected manually. For the last part, five chunks are defined. The degree of emotional homogeneity (confidence) given at the end of each chunk is simply the number of the labels the whole chunk is attributed to, divided by the number of all labels. These five chunks in this example belong all to *N(egative)*.

One long turn, German original word sequence with syntactic and pause labels, and chunk boundaries:

> *und stopp* <v1:1> *Aibo* <v2:0> *stehenbleiben* <s3:0> | *darein musst du laufen* <s3:0> | *da links in die Straße* <p3:2> | *so is gut* <v1:0> *Aibo* <v2:0> *und jetzt laufen* <s3:3> | *fein gemacht* <v1:0> *Aibo* <v2:2> | *und* <A> *weiter* <p3:3> <A> | *und jetzt da nach links in die Straße abbiegen* <d2:0> | *zu dem blauen Napf* <s3:1> | *nein* <v1:0> *Aibo* <v2:0> *nein* <p3:0> | *stopp* <v1:1> *Aibo* <v2:0> *nicht* <p3:2> <G> | *nein* <v1:0> *Aibo* <v2:0> *stopp* <s3:0> | *stehenbleiben* <v1:3> | *Aibo* <v2:1> *stehenbleiben* <s3:eot> |

Table 3: Frequencies of syntactic and pause labels in the full database (48401 words), grey rows and columns indicate chunk triggering boundaries; "% tr." displays the percentage of triggering boundaries within the specific row/column.

Label	Pause length				Sum	% tr.
	0	1	2	3		
eot					13642	100
s3	801	407	340	273	1821	100
p3	885	276	183	135	1479	100
s2	165	10	12	10	197	11
s1	328	48	32	28	436	25
d2	56	5	7	2	70	13
v2	3278	498	376	228	4380	14
v1	3226	217	204	178	3825	10
v2v1	20	49	59	69	197	100
Sum	8759	1510	1213	923		
% tr.	19%	48%	100%	100%		

English translation with chunk boundaries:

> *and stop Aibo stand still | go this way | to the left towards the street | well done Aibo and now go on | well done Aibo | and further on | and now turn into the street to the left | to the blue cup | no Aibo no | stop Aibo no | no Aibo stop | stand still | Aibo stand still |*

Last part, German original, emotion labels per word, syntactic and pause labels, chunk boundaries, and "confidence":

> *nein* NNNNI <v1:0> *Aibo* NNNNI <v2:0> *nein* NNNNN <p3:0> 0.87 | *stopp* NNNNN <v1:1> *Aibo* NNNNI <v2:0> *nicht* NNNNI <p3:2> 0.87 | *nein* NNNIN <v1:0> *Aibo* NNNII <v2:0> *stopp* NNNNN <s3:0> 0.80 | *stehenbleiben* NNNNN <v1:3> 1.0 | *Aibo* NNNNN <v2:1> *stehenbleiben* NNNNN <s3:eot> 1.0 |

If all 13642 turns are split into chunks, the chunk triggering procedure results in a total of 18216 chunks. Note that the chunking rules have been determined in a heuristic, iterative procedure; we corroborated our initial hypotheses, for instance, that pauses between adjacent vocatives are longer on average than pauses after or before single vocatives, with the descriptive statistics given in Table 3. The basic criteria have been formulated in [9]; of course, other thresholds could be imagined if backed by empirical results. The rules for these procedures can be automated fully; in [9] multilayer perceptrons and language models have successfully been employed for an automatic recognition of similar syntactic-prosodic boundaries, yielding a classwise average recognition rate of 90% for two classes (boundary versus no boundary). Our criteria are "external" and objective and are not based on intuitive notions of an "emotional" unit of analysis as in the studies by [28–30]. Moreover, using syntactically motivated units makes processing in an end-to-end system more straightforward and adequate.

In order to obtain emotion labels for the chunks, we first mapped the word level decisions of the five labellers (the raw labels) onto the four main classes *P(ositive)*, *I(dle)*, *N(egative)*, and *R(est)*. A whole chunk is considered to be *P(ositive)* if either the absolute majority (≥50%) of all raw labels is positive or if the proportion of positive raw labels is at least one third and the remaining raw labels are mostly neutral, that is, the positive and the neutral raw labels make up at least 90% of all raw labels. By that, chunks that are mostly neutral but where some words clearly signal the subject's positive state are considered to be *P(ositive)* as well. The heuristic thresholds are adjusted by inspecting the resulting chunk labels. *N(egative)* and *R(est)* chunks are defined along the same lines. If according to these definitions a chunk does not belong to one of these three main classes and the proportion of neutral raw labels is at least 90%, the chunk is considered to be neutral, that is, *I(dle)*. If the proportion of neutral raw labels is lower but at least 50% and raw labels of only one other main class appear, the chunk is assigned *I(dle)* as well. These are the cases where single words signal one nonneutral main class but where the proportion of these words is too low. In all other cases, the raw labels belong to too many main classes and the whole chunk is assigned to the *R(est)* class. The frequencies of the four main classes on the chunk level are given in Table 4.

Our word-based labelling makes it possible to try out different types and sizes of chunks. The other way round would be to attribute the same label to a word that the chunk it belongs to has been annotated with. This has two disadvantages: first, there is only one possibility to map chunk labels onto word labels—each word has to be annotated with the chunk label. Thus, we could, for instance, not contrast SC with EC. Second, the result would be "smeared" because of the contra-factual assumption that all words belonging to a chunk necessarily belong to the same emotion class. This can be, but need not be the case. (Of course, sometimes chunking together words belonging to different classes to the rest class, as we do, results in some "smearing" as well—but at least we do know where and up to what extent. Thus, thresholds can be altered and more or less prototypical cases can be established [5, 6].)

3.3. Eme Chunks: EC. The last unit of analysis investigated in this work consists of eme chunks (ECs). ECs are obtained from the eme sequence by clustering together adjacent emes belonging to the *same* main class. An EC is

therefore an i-tuple of ememes characterised by an identical emotional content. In general, from an utterance of n ememes, we can obtain from 1 to n EC. The practical motivation behind EC is that homogeneous groups of ememes might be easier to classify than single ememes; for the recognition of each emotional class, we are exploiting the largest amount of contextual information, that is, the entire EC. From a theoretical point of view, this approach might be most adequate when we model emotional episodes fully independently from linguistic processing.

In the example below, we draw the ECs that belong to the same utterance described in the previous section. Chunks of ememes (denoted as pairs of spoken word and emotion label, that is, "*word* `emotion_label`") are delimited by markers (symbol "|").

Sequence of ememe chunks:

> *und* `I` | *stopp* `R` | *Aibo* `N` *stehenbleiben* `N` | *darein* `I` *musst* `I` *du* `I` *laufen* `I` *da* `I` | *links* `N` | *in* `I` *die* `I` *Straße* `I` *so* `I` **is* `I` *gut* `I` *Aibo* `I` *und* `I` *jetzt* `I` *laufen* `I` *fein* `I` *gemacht* `I` *Aibo* `I` *und* `I` *weiter* `I` *und* `I` *jetzt* `I` *da* `I` *nach* `I` *links* `I` *in* `I` *die* `I` *Straße* `I` *abbiegen* `I` *zu* `I` *dem* `I` | *blauen* `N` | *Napf* `I` | *nein* `N` *Aibo* `N` *nein* `N` *stopp* `N` *Aibo* `N` *nicht* `N` *nein* `N` *Aibo* `N` *stopp* `N` *stehenbleiben* `N` *Aibo* `N` *stehenbleiben* `N` |

Combining ememes into EC is trivial that way; we use a simple finite state automaton. However, this is only the case because our processing is sort of trivial; we are able to map mixed cases onto one class because we have previously performed MV and adopted threshold criteria (Section 2). Taking into account other dimensions or mixtures of annotations [29] would have required a more sophisticated clustering strategy and would not have been feasible for our data, due to the severe sparse data problem. A mixed case in our data—albeit a rather seldom one—is this sequence of words: "*so* `PNNPI` *weit* `PNNPI` **simma* `PNNPI` *noch* `PNNPI` *nicht* `PNNPI`" (*we ain't that far yet*) which is attributed to *R(est)* both as SC and EC. In fact, this is a good example of a mixture of *motherese* and *reprimanding*, the latter being indicated by the wording, the former by the tone of voice. However, as these are very rare cases, we cannot model them reliably for automatic processing and have to map them onto *R(est)*.

Furthermore, in this paper we will not deal with the problem of automatically obtaining EC given a set of features. Instead, we will assume the emotional labels as given. Thereby, we are avoiding segmentation errors for EC, as we do for SC, in both cases assuming a 100% correct segmentation; this can be considered as an upper bound for classification performance. Note that in preliminary experiments, we found out that Hidden Markov Models trained on EC obtained on the training set lead to a segmentation of the test data that achieves a classification performance comparable to the SC approach. To keep the two approaches fully apart, we do not combine sequences of SC with the same label into one higher "SC/EC-unit."

Table 4: Emotion classes and their SC-based frequencies.

Class	No. train	No. test	No. total	% total
P(ositive)	674	215	889	4.9
I(dle)	5260	5083	10343	56.8
R(est)	667	494	1161	6.4
N(egative)	3358	2465	5823	32.0
Total	9959	8257	18216	100.0

Table 5: Emotion classes and their EC-based frequencies.

Class	No. train	No. test	No. total	% total
P(ositive)	518	199	717	3.9
I(dle)	6410	6185	12595	69.0
R(est)	479	391	870	4.8
N(egative)	2276	1789	4065	22.3
Total	9683	8564	18247	100

3.4. Ememe Chunks versus Syntactic Chunks. Tables 4 and 5 reveal that the overall frequencies of the two chunk types are almost the same; however, there are 2.2 k more *I(dle)* EC-chunks than SC-chunks, counterbalanced by more *P(ositive)*, *N(egative)*, and *R(est)* SC-chunks. Figure 1 displays for each of the four classes and for all classes taken together, frequencies in percent for SC and EC with the length 1 to n words. The same information is given in the stacked histograms of Figure 2; in Figure 1, relationships within classes and differences between type of chunk can be seen, whereas Figure 2 concentrates on frequencies across classes within one plot. One-word ECs are more frequent than one-word SCs; especially for the three "marked" classes *P(ositive)*, *N(egative)*, and *R(est)*, there is a decline in frequencies, especially for chunks with 2, 3, or 4 words, which display higher frequencies for SC than for EC. These differences can be traced back to the different MV and thresholds. EC in our case are "pure," that is, after the initial, word-based MV, the labels are fixed, and only adjacent words with identical labels are combined into EC; as we mentioned above, this is not a necessary condition for EC but had to be chosen for our database, to avoid the sparse data problem. In contrast, if it is a chunk with more than one word, individual words belonging to the very same SC can be attributed to different classes but the combined threshold for the whole SC overrides such differences.

4. Features and Classifiers

4.1. Acoustic Features. The main focus has been on prosodic features in the past, in particular pitch, durations, and intensity [31]. Comparably small feature sets (10–100) were first utilised. In only a few studies, low-level feature modelling on a frame level was pursued, usually by Hidden Markov Models (HMMs) or Gaussian Mixture Models (GMMs). The higher success of static feature vectors derived by projection of the Low-Level Descriptors (LLDs) such as pitch or energy by descriptive statistical functional application such as lower order moments (mean, standard deviation) or

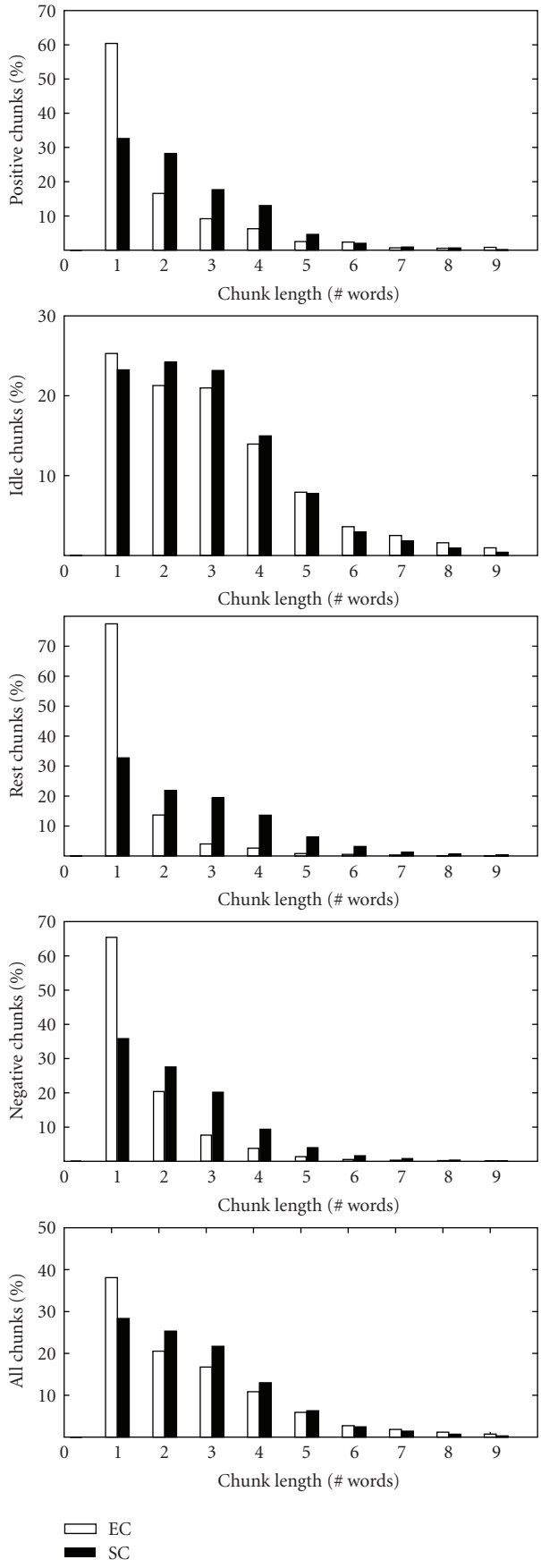


Figure 1: Chunk histogram with frequencies.

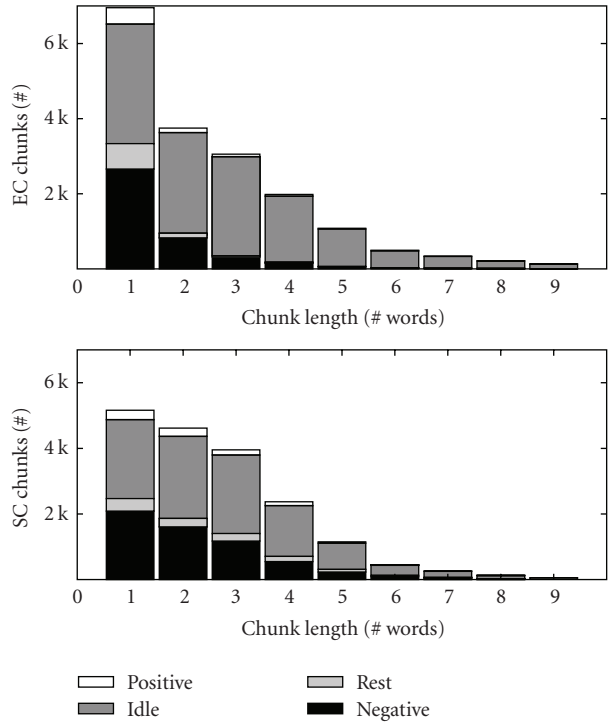


Figure 2: Chunk histogram with frequencies, stacked.

extrema is probably justified by the supra-segmental nature of the phenomena occurring with respect to emotional content in speech. In more recent research, also voice quality features such as Harmonics-to-Noise Ratio (HNR), jitter, or shimmer, and spectral and cepstral features such as formants and Mel-Frequency Cepstral Coefficients (MFCCs) have been successfully added to prosodic features. At the same time, brute-forcing of features (1000 up to 50000), for example, by analytical feature generation, partly also in combination with evolutionary generation, has become popular. It seems as if this (slightly) outperforms hand-crafted features while the individual worth of automatically generated features seems to be lower. Within expert-based hand-crafted features, perceptually more adequate features have been investigated, reaching from simple log-pitch to Teager energy or more complex features such as articulatory features (e.g., (de-)centralisation of vowels).

In this study, a feature set is employed that shall best cover the described gained knowledge. We therefore stick to the findings in [32] by choosing the most common and at the same time promising feature types and functionals covering prosodic, spectral, and voice quality features. Furthermore, we limit to a systematic generation of features. For the highest transparency, we utilise the open source openSMILE feature extraction and choose the basic set used in the only official challenge on emotion recognition from speech to the present day (cf. "Classifier Sub-Challenge" [33]). In detail, the 16 Low-Level Descriptors chosen are: zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), HNR by autocorrelation function, and MFCCs 1–12. To each of

Table 6: Acoustic Low-Level Descriptors (LLD) and functionals.

LLD (16 · 3)	Functionals (12)
(Δ, ΔΔ) ZCR	Mean
(Δ, ΔΔ) RMS Energy	Standard deviation
(Δ, ΔΔ) F0	Kurtosis, skewness
(Δ, ΔΔ) HNR	Extremes (max, min): value (2), rel. position (2), range (1)
(Δ, ΔΔ) MFCC 1-12	Linear regression: offset, slope, MSE

these, the delta coefficients are additionally computed. We further add double-delta coefficients for a better modelling of context. Then the 12 functionals mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their Mean Square Error (MSE) are applied on a chunk basis as depicted in Table 6. Thus, the total feature vector per chunk contains $16 \cdot 3 \cdot 12 = 576$ attributes.

4.2. Linguistic Features. Spoken or written text also carries information on the underlying affective state [34–36]. This is usually reflected in the usage of certain words or grammatical alterations—which means in turn, in the usage of specific higher semantic and pragmatic entities.

From the many approaches existing we chose vector space modelling, that is, bag of words [37]. This is a well-known numerical representation form of text in automatic document categorisation introduced in [38]. It has been successfully ported to recognise sentiments in [39] or emotion and interest in [40]. The possibility of early fusion with acoustic features helped make this technique very popular as shown in [37].

For the FAU Aibo Emotion Corpus, the vocabulary size is 1146 entries. But only a fraction of these words conveys relevant information about the underlying emotional states of a person. In order to reduce the information in a meaningful way, two methods can be applied: stopping and stemming. Stopping uses simple rules or a data-driven evaluation to exclude single words from the vocabulary. A simple yet popular method for reducing the vocabulary is exploited: a minimum training database word frequency, here two, determines the necessary minimum number of occurrences for a word in the database for being part of the vocabulary. Very rare words are therefore discarded. Stemming instead is a method for reducing different morphological forms of a word to its base form. The Iterated Lovins Stemmer [41] is used for the experiments in this paper.

The main idea of the bag of words approach is the representation of words (or lexemes if stemming is applied) as numeric features. For each word (i.e., term) in the vocabulary, a corresponding feature that represents its frequency of occurrence in the unit exists, resulting in a high-dimensional feature vector space. Each unit can therefore be mapped to a vector in this feature space.

This frequency can be transformed in various ways [42, page 311], [38]. The logarithmic term frequency normalised to the inverse document (here database) frequency (TFIDF) in combination with normalisation to the unit length proved to be the best in our experiments.

Within this paper, the linguistic analysis is based on the correct transcription of the spoken content. Therefore it describes the performance under perfect speech recognition conditions. This follows the typical reporting of linguistic analysis results in emotion recognition, as it allows for better comparability of results [37]; the corpus comes with the transcription, while speech recognition results would differ from site to site. Also, some practical relevance exists: consider media retrieval from broadcasts; here the close captions are usually available. However, to close the gap to the real world where spoken content has to be determined by an ASR engine first, we had carried out experiments employing ASR for this corpus in other studies: though recognition of affect related speech is a rather difficult problem which has not been solved yet to complete satisfaction [43], this did not yield marked differences, as reported, for example, in [44, 45] for this corpus. This derives from the fact that not the perfect word chain is needed as, for example, in transcription of speech. Some minor mistakes are caught by stemming and stopping, and not all words are necessarily needed. Insertions and substitutions are only critical if they change the "tone" of the affective content. As additional features in linguistic analysis, we utilise each word's start and end time, as well as the derived duration. This is motivated by the fact that an ASR engine would also provide this information.

4.3. Classifiers. Classifiers typically used for the recognition of emotion from speech comprise a broad variety: depending on the feature type considered for classification either dynamic algorithms as Hidden Markov Models [46] or Multiinstance Learning techniques [10] for processing on a frame-level, and static classifiers for processing on the supra-segmental functional level are found. With respect to static classification the list of classifiers seems endless: Neural Networks (mostly Multilayer Perceptrons), Naïve Bayes, Bayesian Networks, Gaussian Mixture Models, Decision Trees, Random Forests, Linear Discriminant Classifiers, k-Nearest Neighbour distance classifiers, and Support Vector Machines are found most often [4]. Also a selection of ensemble techniques has been applied, as Boosting, Bagging, Multiboosting, and Stacking with and without confidences [47]. Finally, the two general types may also be mixed by fusion of dynamic and static classification [48].

As we consider acoustic and linguistic information, the two information streams need to be integrated. In this respect, all experiments found in the literature use static classification techniques [17, 37, 49]: an early fusion is usually the best choice for preserving all information prior to the final decision. Thus, the acoustic features introduced in Section 4.1 and the linguistic ones introduced in Section 4.2 are combined in one feature vector on the respective unit level (i.e., word or chunk), which demands for static classification.

The classifier of choice to this aim in this paper is a discriminatively learned simple Bayesian Network, namely Discriminative Multinomial Naïve Bayes (DMNB) [50]

instead of Support-Vector Machines (SVMs) and Random Forests as applied in our previous investigations [23, 32, 33]. The reason is twofold: first, DMNB only requires lower memory and only a fraction of the computation time of SVM. (Sequential Minimal Optimisation training of SVM with linear Kernel demanded 200 times higher computation time than DMNB in parameterisation as below using [42] on an 8 GB RAM, 2.4 GHz, 64 Bit industry PC.) At the same time, the mean recall values resulted in a slight absolute improvement over SVM in our experiments on the FAU Aibo Emotion Corpus (−0.9/+1.3 weighted/unweighted average recall on average for acoustic features; +6.9/+2.3 for linguistic features). Second, the parameter learning is carried out by discriminative frequency estimation, whereby the likelihood information and the prediction error are considered. Thus, a combination of generative and discriminative learning is employed. This method is known to work well in highly correlated spaces (as in our case), to converge quickly, and not to suffer from overfitting.

For optimal results we found it best to ignore the frequency information in the data and select a number of ten iterations. Numeric variables are discretised using unsupervised ten-bin discretisation [42]. Multiclass decision is obtained by transformation into binary problems by taking the two largest classes, each.

5. Classification

As mentioned above and carried out within [33], we split the FAU Aibo Emotion Corpus into train and test partitions by schools of recording. Thus, utmost independence of the speaker, room acoustics, general intonation and articulation patterns, and wording of the children is ensured. To better cope with this variety, all features are standardised per partition (speaker group normalisation). Due to the high imbalance among classes (cf. Table 1), balancing of the training instances is further mandatory to achieve reasonable values of unweighted recall and thus avoid overfitting of strong classes (here *I(dle)* and *N(egative)* [17]. The chosen straightforward strategy is random mixed up-sampling of sparse and downsampling of majority classes' instances enforcing unit distribution while preserving the total number of instances. Note that the order of operations has an influence on (un)weighted recall figures [33]; we first standardise and then balance the training. Next we classify with DMNB as described. At this point constant parameterisation is preferred over individual optimisation; thus, no alterations are undertaken with respect to number of iterations, quantisation, and so forth, among the different units and feature types to be classified.

Table 7 displays weighted average recall (WA), that is, the overall recognition rate (RR) or recall (number of correctly classified cases divided by total number of cases), and unweighted average recall (UA) (or "class-wise" computed recognition rate (CL)), that is, the mean along the diagonal of the confusion matrix in percent, for three sets of features: only acoustic features, only linguistic features, and both acoustic and linguistic features (early fusion).

Table 7: Evaluation in percent correct; (un-)weighted average recall (UA/WA). Note that the chunk- and the word-based evaluations of word units coincide; words can be seen a the smallest possible chunk.

Unit	Acoustic		Linguistic		Ac. + ling.	
	WA	UA	WA	UA	WA	UA
Basic unit of evaluation: the chunk						
WO	49.73	46.15	44.27	45.50	53.59	**48.56**
SC	46.43	44.40	43.04	42.30	50.02	**46.33**
EC	57.23	51.63	62.42	51.80	64.89	**55.38**
Basic unit of evaluation: the word						
WO	49.73	46.15	44.27	45.50	53.59	**48.56**
SC	44.56	45.10	40.63	44.37	48.82	**48.35**
EC	65.84	53.10	71.98	53.00	73.66	**56.77**

There are two different types of evaluation: first we evaluate for the whole units WO, SC, and EC; note that the total number of chunks is different for each of these units. Then, we evaluate WO, SC, and EC by checking each word in these units whether it has been classified correctly, that is, attributed to the class the higher unit it belongs to has been annotated with. Obviously, the evaluation of the unit WO is identical under the two methods, as words can be seen as smallest possible chunks. WA tends to be higher because of the bias in class distribution; UA is more relevant for applications which are, most of the time, more interested in the "marked" classes, that is, in our case, not in the frequent *I(dle)* class; thus, we concentrate on the interpretation of UA. All results are above chance level (25% correct). The chunk-based figures might be more relevant if we have applications in mind, the word-based figures are more balanced. We can see that the early fusion of acoustic and linguistic features pays off, always yielding higher WA and UA. As there are more—and especially longer *I(dle)* chunks containing more words, it could be expected that word-based evaluation for SC and EC yields better results; the differences are, however, not marked. As ECs are more consistent—all words belonging to an EC belong to the same class—UA for EC is higher than for SC.

On average, the unit "word" contains less information than the units SC and EC; each unit consists of only one word whereas SC and EC mostly consist of more than one word. The number of SC and EC is in the same range, as can be seen in Tables 4 and 5, although EC should be more consistent than SC. In Table 8, we display classification results for cross-unit evaluation, that is, we use different units for the training and for testing partitions. This could only be done for acoustic features because of the unbalanced distribution of linguistic features in the different units. We see that performance is really worst when we use the unit "word" both as train or test unit, with UA being consistently below 40 % correct. Overall, there is again almost no difference between chunk-based and word-based evaluation of UA. Although the figures are of course lower than for the within-unit evaluation displayed in Table 7, it is reassuring that

Table 8: Cross unit: train ≠ test; evaluation in percent correct; (un-)weighted average recall (UA/WA); acoustics only.

Train	Test	Chunk based		Word based	
		UA	WA	WA	UA
EC	SC	47.77	42.15	51.51	42.10
WO	SC	44.01	38.43	46.83	39.90
SC	EC	45.22	45.40	47.85	46.98
WO	EC	32.53	30.73	31.23	33.08
EC	WO	37.81	38.30	37.81	38.30
SC	WO	27.50	39.35	27.50	39.35

Table 9: Confusion matrix, acoustics + linguistics, word evaluation in percent correct.

Class. as →	*P*	*I*	*R*	*N*	Total
P(ositive)	**47.49**	24.41	16.38	11.70	299
I(dle)	9.60	**51.76**	12.95	25.67	18698
R(est)	21.27	26.54	**21.63**	30.54	550
N(egative)	3.15	13.12	10.30	**73.41**	2697

performance does not break down when we train with EC and test with SC or vice versa.

Table 9 displays the confusion matrix for the fusion of acoustics and linguistics for the word evaluation (cf. the two last columns in Table 7, row WO), giving an impression of the confusion between classes. The confusion between the classes is as expected; no much confusion between *P(ositive)* and *N(egative)*, most confusion between *I(dle)* (and partly *R(est)*) and the other classes. Table 10 displays recall-rates (i.e., only the figures of the diagonal of the confusion matrices) as correctly classified cases per class, for the remaining four constellations from Table 7. This gives an impression of the performance per class across all constellations. Basically, the picture is always the same: the mixed *R(est)* class is recognised worst and almost evenly smeared across all classes. The highest recognition rates can be observed for the rather acoustically and linguistically marked *N(egative)* instances—both for SC and EC; *P(ositive)* is in between.

6. Discussion

6.1. Classification Performance: The Reality Shock. The scientific community has been used to good or almost perfect classification performance in emotion recognition; it is such figures that are remembered and implicitly defined as standard. We have to realise, however, that such figures have been obtained only within specific constellations: acted data [51], prototypical cases preselected out of the whole database, or a focus on one specific class, modelling all other classes as rest/garbage; for this last constellation, high recall can be obtained if we can live with many false alarms in the rest classes. Normally, the data have not been processed fully automatically but the experiments have been based on the spoken word chain. In the present study, we aim at realistic conditions—apart from the last step to use fully automatic ASR. In [23] we could show that—depending on the recording conditions and the feature set used—ASR errors do not always deteriorate emotion recognition.

Table 10: Classwise recall values (i.e., diagonal values of confusion matrices), acoustics + linguistics; chunk evaluation in percent correct. Note that the chunk- and the word-based evaluations of WO coincide.

Constellation	*P(ositive)*	*I(dle)*	*R(est)*	*N(egative)*
WO, chunk based	47.49	51.76	21.63	73.41
SC, chunk based	51.62	44.26	22.06	67.34
EC, chunk based	44.72	65.40	40.66	70.65
WO, word based	47.49	51.76	21.64	73.41
SC, word based	55.18	46.43	21.45	70.26
EC, word based	48.49	75.93	33.81	68.85

We simply do not know yet which type of realistic databases—amongst them our FAU Aibo Emotion Corpus—could be conceived as being representative, as far as distinctiveness of classes and by that, goodness of performance is concerned. Chances are, however, that we will never achieve such high performance as we did, using only acted and/or prototypical data, and that approaches such as the present one—trying to model all phenomena present is a database—will give way to more focused approaches, aiming at specific classes for specific application tasks.

6.2. Deciding between Types of Units. At least three aspects are relevant for deciding between WO, SC, and EC—or any other type of sequencing emotional episodes: first, performance; second, adequacy in real-life applications; third, perceptual, cognitive adequacy.

Performance has been significantly better for EC than for SC. Note that in this paper, we used the spoken word chain simulating 100% correct word recognition, and a manual segmentation into SC and EC. For a fair comparison between SC and EC, this had to be done automatically. We know that SC can be established automatically with high reliability even for spontaneous speech [9]. As for EC, this might look like a "Münchhausen" approach; finding the boundaries of phenomena we afterwards want to recognise; however, preliminary experiments showed that it can be done using an HMM approach, albeit yielding lower classification performance in the range of SC. Semantically "rich" words, that is, content words such as nouns, adjectives, and verbs, tend to be marked emotionally to a higher extent than function words such as particles. For instance, in our data, more EC (22.5%) than SC (21.3%) consist only of 1 to n content words. A modelling of part-of-speech (POS) sequences yields a classification performance, not much lower than one obtained with acoustic modelling [32]. POS modelling is rather robust because ASR confusions between words within one POS class have no effect. Factors like these make it likely that LM modelling of EC is as promising as LM modelling of SC. Thus, it is an empirical question to be addressed

whether EC will be classified better than SC, if the process is fully automated. The best compromise between automation and performance seems to be WO. Here, we obtain better results than for SC—but still worse results than for EC. Word segmentation is obtained for free if ASR has been applied. However, due to the reasons sketched passim and in the following, the single ememe, that is, the word, might not be the optimal unit, if it comes to processing in both higher linguistic and emotion modules.

Applications are different. If we look at the attempt towards a taxonomy of applications in [52], most important for segmentation might be the difference between online and offline applications. We mentioned in the beginning that for online applications such as SmartKom [11] or Semaine [12], incremental processing will be mandatory because of time constraints. Matters are similar in any interaction between users and Embodied Conversational Agents (ECAs) or robots. In such online applications, there is normally an interaction between system and user. The system does not only monitor somehow the user's emotional states but has to recognise and process linguistic content and semantics and illocutions (dialogue acts) in order to react appropriately. This makes a close dovetailing of linguistics and para-linguistics such as monitoring emotional states most adequate, and this in turn might favour the processing of SC instead of EC or WO. It is different in offline applications; the processing of movie databases in search for emotional episodes needs not be incremental and can be done in several passes. Thus, we can imagine one pass for emotion monitoring within the whole movie, and then a second pass for segmentation, and so on.

To our knowledge, there are not many studies on the relationship between human speech/linguistic processing and human emotion processing. We know, however, that phonetic/psycholinguistic studies on the localisation of nonverbal signals within speech showed that listeners tend to structure the perception of these phenomena along the perception and comprehension of linguistic phenomena (sentence processing) [53]. Unpublished studies on the localisation of laughter in our data showed that this is the case for the production of paralinguistic events as well. Thus, it might be that linguistics and emotions are more intertwined—at least within interactions where emotional and non-emotional episodes alternate. If this is the case, the modelling of SC seems to be most adequate also from the point of view of cognition and comprehension.

7. Concluding Remarks

The unique contribution of the present study is the use of word-based annotations and the subsequent mapping onto different types of higher units, to investigate promising possibilities of segmenting emotional episodes. However, word-based annotation is very time-consuming and thus expensive. Perhaps it should not be established as a new standard but only be used for basic research. The higher units "syntactic chunk" and "emotion/ememe chunk" introduced in this study are, in our opinion, representative for two different types of most promising units. However, a great variety of different thresholds or mapping procedures can be imagined. Most of them will not differ considerably, as far as usability or performance is concerned. Although being a truism, we definitely need more realistic databases for deciding between such alternative approaches.

Acknowledgments

This work originated in the CEICES initiative (Combining Efforts for Improving Automatic Classification of Emotional User States) taken in the European Network of Excellence HUMAINE [37]. The research leading to these results has received funding from the European Community under Grant (FP7/2007–2013) no. 211486 (SEMAINE), Grant no. IST-2002–50742 (HUMAINE), and Grant no. IST-2001-37599 (PF-STAR). The responsibility lies with the authors.

References

[1] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, NY, USA, 1988.

[2] R. Cowie, N. Sussman, and A. Ben-Ze'ev, "Emotions: concepts and definitions," in *Humaine Handbook on Emotion*, P. Petta, Ed., Springer, Berlin, Germany, 2010, to appear.

[3] H. Grice, "Logic and conversation," in *Syntax and Semantics*, P. Cole and J. Morgan, Eds., vol. 3 of *Speech Acts*, pp. 41–58, Academic Press, New York, NY, USA, 1975.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[5] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of tuning—prototyping for automatic classification of emotional user states," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 489–492, Lisbon, Portugal, 2005.

[6] D. Seppi, A. Batliner, B. Schuller, et al., "Patterns, prototypes, performance: classifying emotional user states," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '08)*, pp. 601–604, Brisbane, Australia, September 2008.

[7] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[8] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "Feeltrace: an instrument for recording perceived emotion in real time," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 19–24, Newcastle, Northern Ireland, 2000.

[9] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, no. 4, pp. 193–222, 1998.

[10] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: a multi-corpus study," in *Speaker Classification II*, C. Müller, Ed., vol. 4441 of *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence*, pp. 43–56, Springer, Berlin, Germany, 2007.

[11] M. Streit, A. Batliner, and T. Portele, "Emotions analysis and emotion-handling subdialogues," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed., pp. 317–332, Springer, Berlin, Germany, 2006.

[12] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[13] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of ACM Multimedia*, pp. 669–676, Singapore, 2005.

[14] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech," *User Modelling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 175–206, 2008.

[15] S. Steidl, *Automatic classification of emotion-related user states in spontaneous children's speech*, Ph.D. thesis, Logos, Berlin, Germany, 2009.

[16] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth, "We are not amused—but how do you know? User states in a multi-modal dialogue system," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Interspeech '03)*, pp. 733–736, Geneva, Switzerland, September 2003.

[17] B. Schuller, R. Müller, F. Eyben, et al., "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[18] M. Blomberg and D. Elenius, "Collection and recognition of children's speech in the PF-Star project," in *Proceedings of the Swedish Phonetics Conference (Fonetik '00)*, pp. 81–84, Umeå, Sweden, 2003.

[19] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1044–1046, 2007.

[20] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 2, pp. 137–140, Hong Kong, 2003.

[21] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[22] M. Holodynski and W. Friedlmeier, *Development of Emotions and Emotion Regulation*, Springer, New York, NY, USA, 2006.

[23] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 941–944, Honolulu, Hawaii, USA, 2007.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 1517–1520, Lisbon, Portugal, 2005.

[25] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 1695–1698, Rhodes, Greece, 1997.

[26] J. Schwitalla, *Gesprochenes Deutsch: Eine Einführung*, Erich Schmidt, Berlin, Germany, 1997.

[27] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Communication*, vol. 40, no. 1-2, pp. 117–143, 2003.

[28] Z. Inanoglu and R. Caneel, "Emotive alert: HMM-based emotion detection in voicemail messages," in *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*, pp. 251–253, San Diego, Calif, USA, 2005.

[29] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[30] F. de Rosis, A. Batliner, N. Novielli, and S. Steidl, "'You are sooo cool, Valentina!' recognizing social attitude in speech-based dialogues with an ECA," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds., pp. 179–190, Springer, Berlin, Germany, 2007.

[31] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[32] B. Schuller, A. Batliner, D. Seppi, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 2, pp. 2253–2256, Antwerp, Belgium, August 2007.

[33] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech '09)*, pp. 312–315, Brighton, UK, 2009.

[34] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 2675–2678, Aalborg, Denmark, September 2001.

[35] Z.-J. Chuang and C.-H. Wu, "Emotion recognition using acoustic features and textual content," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 1, pp. 53–56, Taipei, Taiwan, 2004.

[36] K. Dupuis and K. Pichora-Fuller, "Use of lexical and affective prosodic cues to emotion by younger and older adults," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 2, pp. 2237–2240, Antwerp, Belgium, August 2007.

[37] A. Batliner, S. Steidl, B. Schuller, et al., "Combining efforts for improving automatic classification of emotional user states," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, pp. 240–245, Ljubljana, Slovenia, 2006.

[38] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning (ECML '98)*, C. Nédellec and C. Rouveirol, Eds., pp. 137–142, Chemnitz, Germany, 1998.

[39] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '02)*, pp. 79–86, Philadelphia, Pa, USA, 2002.

[40] B. Schuller, N. Köhler, R. Müller, and G. Rigoll, "Recognition of interest in human conversational speech," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, vol. 2, pp. 793–796, Pittsburgh, Pa, USA, 2006.

[41] J. B. Lovins, "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22–31, 1968.

[42] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2005.

[43] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: clarifying the issues and enhancing performance," *Neural Networks*, vol. 18, no. 4, pp. 437–444, 2005.

[44] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Does affect affect automatic recognition of children's speech?" in *Proceedings of the 1st Workshop on Child, Computer and Interaction*, Chania, Greece, 2008.

[45] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: putting ASR in the loop," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, pp. 4585–4588, Taipei, Taiwan, 2009.

[46] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[47] B. Schuller, R. Jiménez Villar, G. Rigoll, and M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 1, pp. 325–328, Philadelphia, Pa, USA, March 2005.

[48] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition ofemotions within speech," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech '07)*, vol. 4, pp. 2249–2252, Antwerp, Belgium, August 2007.

[49] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[50] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative parameter learning for Bayesian networks," in *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 1016–1023, Helsinki, Sweden, 2008.

[51] A. Batliner, K. Fischer, R. Huber, J. Spilker, and R. Nöth, "Desperately seeking emotions: actors, wizards, and human beings," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 195–200, Newcastle, Northern Ireland, 2000.

[52] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, "A taxonomy of applications that utilize emotional awareness," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, pp. 246–250, Ljubljana, Slovenia, 2006.

[53] M. Garrett, T. Bever, and J. Fodor, "The active use of grammar in speech perception," *Perception and Psychophysics*, vol. 1, pp. 30–32, 1966.

Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 263593, 17 pages
doi:10.1155/2010/263593

Review Article

Emotion on the Road—Necessity, Acceptance, and Feasibility of Affective Computing in the Car

Florian Eyben,[1] Martin Wöllmer,[1] Tony Poitschke,[1] Björn Schuller,[1] Christoph Blaschke,[2] Berthold Färber,[2] and Nhu Nguyen-Thien[3]

[1] *Institute for Human-Machine Communication, Technische Universität München, 80333 München, Germany*
[2] *Human Factors Institute, Universität der Bundeswehr München, 85577 Neubiberg, Germany*
[3] *Continental Automotive GmbH, Interior BU Infotainment & Connectivity, Advanced Development and Innovation, 93055 Regensburg, Germany*

Correspondence should be addressed to Florian Eyben, eyben@tum.de

Received 25 February 2010; Accepted 6 May 2010

Academic Editor: Kostas Karpouzis

Besides reduction of energy consumption, which implies alternate actuation and light construction, the main research domain in automobile development in the near future is dominated by driver assistance and natural driver-car communication. The ability of a car to understand natural speech and provide a human-like driver assistance system can be expected to be a factor decisive for market success on par with automatic driving systems. Emotional factors and affective states are thereby crucial for enhanced safety and comfort. This paper gives an extensive literature overview on work related to influence of emotions on driving safety and comfort, automatic recognition, control of emotions, and improvement of in-car interfaces by affect sensitive technology. Various use-case scenarios are outlined as possible applications for emotion-oriented technology in the vehicle. The possible acceptance of such future technology by drivers is assessed in a Wizard-Of-Oz user study, and feasibility of automatically recognising various driver states is demonstrated by an example system for monitoring driver attentiveness. Thereby an accuracy of 91.3% is reported for classifying in real-time whether the driver is attentive or distracted.

1. Introduction

More than 100 years of history of the automobile are marked by milestones as the combustion engine and mechanical components followed by electrical and electronic device integration, increasing usage of control technique and software. Apart from reduction of fuel consumption, and thus alternative actuation and light weight construction, the main research interest in automobile development in the near future is dominated by driver assistance and natural, intuitive driver-car communication. This statement is supported by various EU-funded research projects such as PREVENT (http://www.prevent-ip.org/), SASPENCE (subproject of PREVENT), and PROSPER [1], which aim at advancing the state-of-the-art in the area of driver assistance systems, and a body of literature on in-car signal processing, for example, [2]. In this respect the ability of a car to talk naturally and provide a virtual companion can be expected to be a market success decisive factor in future automotive systems as "next big thing" on par with automatic driving systems and intelligent measures to improve driving safety.

Emotional factors are decisive for enhanced safety and comfort while driving a car, as we will show in Section 2.1. It is thus necessary for a car to sense these and by that better understand a driver's intention and/or state. The aim of in-car emotion recognition should be to support the driver in performing primary, secondary, and tertiary driving tasks. Thereby the primary driving task, which includes steering, accelerating, braking, and choosing the correct lane, speed, route, and distance to other vehicles, as well as the secondary driving task, denoting activities like dimming, operating windscreen wipers, coupling, changing gears, and blinking, can be seen as rather safety-related whereas the tertiary driving task (operating air conditioner, seat heater, radio, and phone) mainly refers to comfort [3].

Constantly increasing provision of speech technology as well as gaze detection and eye/head movement monitoring mark the beginning of more natural ways of human-machine interactions which are based on intuitive communication modalities. Recognition of emotion from vocal and facial expression, physiological measurement, and contextual knowledge will be the next key-factor driving improved naturalness in many fields of Human-Computer Interaction [4]. Next to the modalities speech, facial expression, physiological measurements, and contextual knowledge, driving parameters can be used as an important and reliable modality for driver emotion and state recognition in a car.

This paper will give an introduction to in-car affective computing with an extensive literature overview on studies and existing work in Section 2. The section includes a discussion of the influence of emotions on the driving performance, and lists methods for recognising emotion and control of emotion. Moreover, the concept of a "virtual companion" will be presented. We will outline various use-case examples which illustrate how emotion-oriented technology can be used in a vehicle in Section 3. An open issue is that of user acceptance of emotion aware technology. In order to assess this acceptance, we conduct a pilot study where we interrogate users on their experiences using a Wizard-of-Oz (WoZ) prototype system (see Section 4). To show the feasibility of automatic driver state recognition for safety and infotainment-related tasks we finally and briefly present a system for detecting driver distraction in Section 5, based on measuring various driving style parameters and tracking the driver's head motion.

2. Literature Review

This section gives an extensive literature overview on the topic of affective computing and the role of emotions and other driver states in the car. We investigate the influence of affective states on the driving performance in Section 2.1. Road rage, fatigue, stress, confusion, nervousness, and sadness are picked as main factors with respect to driving safety. Further, we deal with techniques for recognition and control of various driver states. The two main strategies of countersteering negative emotions, and alternatively adapting car functionalities to the driver's emotion are the focus of Section 2.2. They are followed by a discussion of the modalities used to actually recognise driver emotion in the car. Next, we discuss the feasibility and benefits of a "socially competent" car in Section 2.3. This type of car is supposed to enhance the driving experience, that is, the driver's pleasure, as "happy drivers" were shown to be the better drivers in any respect (e.g., [5]).

2.1. The Influence of Affective States on Driving Performance. The role that emotions and other mental states (such as fatigue) play while driving a car becomes evident when considering essential driver abilities and attributes that are affected by emotion: perception and organisation of memory [6, 7], categorisation and preference [8], goal generation, evaluation, decision-making [9], strategic planning [10], focus and attention [11], motivation and performance [12], intention [13], communication [14–16], and learning [17]. Taking into account the great responsibility a driver has for her or his passengers, other road users, and her- or himself, as well as the fact that steering a car is an activity where even the smallest disturbance potentially has grave repercussions, keeping the driver in an emotional state that is best suited for the driving task is of enormous importance. Of course, similar to a simple control circuit, for an "intelligent" car the first step towards *influencing* or even *controlling* a drivers emotional state is to *measure* emotion. But what kind of emotion would be ideal to optimally perform primary and secondary driving tasks? Obviously the driver's emotion should support capabilities like attention, accurate judgement of traffic situations, driving performance, compliance, fast and correct decision making, strategic planning, and appropriate communication with other road users.Literature answers the question of the optimum emotional state with the statement "happy drivers are better drivers" [5, 18, 19]. Research and experience demonstrate that being in a good mood is the best precondition for safe driving and that happy drivers produce fewer accidents [20]. The following sections take a closer look at how other affective states can influence driving performance.

2.1.1. Aggressiveness and Anger. Aggressiveness and anger are emotional states that extremely influence driving behaviour and increase the risk of causing an accident [21]. "Road rage" denotes an extreme case of aggressive driving implying specific incidents of anger intentionally directed at another driver, vehicle or object. Approximately 16 million people in the US might suffer from road rage disorder, as was reported by CNN news in 2006 and is cited in [22]. Extreme forms even involve physical attacks, confrontation with other drivers, "tailgating" (i.e., driving too closely at a distance), and cutting another driver off the road. Apart from these grave misbehaviours, slightly milder forms of road rage like provocation of other road users, obscene gestures, and expressing anger by yelling or honking are part of day-to-day traffic interactions and concern a significantly larger number of traffic participants [23]. Even those comparatively mild levels of aggressiveness disrupt the drivers attention and preclude the driver from concentrating on the traffic, increasing the risk of an accident [24].

On the other hand, a too low level of activation (e.g., resulting from emotional states like sadness or fatigue) also leads to reduced attention as well as prolonged reaction time and therefore lowers driving performance. As stated by Yerkes and Dodson [25], a medium level of activation results in the best performance, whereas the precise optimum level of activation (OLA) varies, depending on the difficulty of the task. This relationship is commonly known as the Yerkes-Dodson Law.

2.1.2. Fatigue. Another example for dangerous driver states is sleepiness, which affects all abilities that are important for driving, in a negative way. The fact that even when people recognise they are tired, they often force themselves

not to take a rest but to go on driving, makes sleepiness a severe problem in today's car traffic [20]. According to [26] up to 3% of all motor vehicle crashes happen due to sleepiness whereas the main risk factors are youth, shift work, alcohol, drugs, and sleep disorders. Since fatigue degrades alertness as well as quick and accurate perception, judgement, and action, tired drivers not only risk accidents from falling asleep while driving, but also from slowing down of reactions or loss of attention during time-critical manoeuvres like abruptly breaking when observing the end of a traffic jam on a highway or avoiding to hit a pedestrian. Various surveys demonstrate that many people experienced excessive sleepiness during driving [27–29]: 55% of 1 000 interviewed individuals had driven while being sleepy during the year preceding the survey [30], 23% had fallen asleep while driving once or more than once in their life, and almost 5% already had an accident due to sleepiness. Those figures indicate that fatigue is a serious problem in traffic and belongs to the most alarming states of a car driver.

2.1.3. Stress. As automobile driving itself can often be a source of stress, it seems obvious that stress is an affective state which is very likely to occur in a car. Driving closely behind other vehicles, changing traffic lanes during rush hour, receiving a phone call, getting to ones destination on time, and paying attention to traffic rules are only some of the tasks which partly have to be fulfilled simultaneously by the driver and therefore cause mental overload. A frequently experienced event is rush hour traffic congestion which is interpreted as stressful by almost every automobile driver and causes many people to use public transport and to dismiss the private car in urban areas. Similar to anger and aggressiveness, stress usually implies a high level of arousal which in turn leads to a lack of focus and attention and therefore lowers driving performance [31]. Excessive workload during driving, for example due to distraction caused by using the cell phone in the car, was shown to downgrade driving style [32]: using a cell phone causes drivers to have higher variations in accelerator pedal position and to drive more slowly with high variation in speed. Obviously such a decrease of driving performance and concentration is also caused by other in-car information and entertainment systems. (cf. http://www.nuance.com/distracteddriving/, and Department for Transport Project: Scoping Study of Driver Distraction (2008), Ref. T201T, summary available at: http://www.dft.gov.uk/rmd/project.asp?intProjectID=12560) which suggests that with the growth of car functionality the need for monitoring the drivers' stress level increases.

2.1.4. Confusion. Confusion or irritation is a further state which can lead to a loss of self-control and control over the vehicle, increasing the probability of committing a traffic violation or even being involved in an accident [33]. Sources of confusion can on the one hand be nonintuitive user interfaces or defective systems (like e.g., a badly designed navigation system or an error prone speech recogniser; in the latter case detected confusion could be used to increase the speech recognisers robustness). On the other hand irritating traffic situations like route diversions, mistakable signs or complicated routing of a road can cause confusion. Just like stress, irritation leads to disturbance of driver capabilities such as decision-making, attention, perception, and strategic planning. Particularly older people tend to be confused by the amount of information they have to process simultaneously during driving [34] as today neither car information systems nor all traffic signs or routes are designed for elderly people who often have longer reaction times and slower perception. Supporting irritated drivers through intelligent emotion-sensitive assistance systems will become indispensable for future car generations as confusion potentially increases with the number of car functionalities.

2.1.5. Nervousness. Nervousness is an affective state that implies a level of arousal which is above the degree of activation that is best suited for the driving task. Reasonable decision-making as well as strategic planning and concentration are affected when being nervous. Reasons for nervousness are variable and can be related directly to the driving task (e.g., for novice drivers) or to other personal or physical circumstances. In [35] the nervousness induced by the use of drugs is examined with respect to effects on driving: nervous drivers tend to perform worse as far as driving ability is concerned—mainly due to poorer concentration. Also Li and Ji name nervousness as one of the most dangerous driver states and point out the importance to detect nervousness to provide intelligent assistance and appropriate alerts [36].

2.1.6. Sadness. Also negative emotions with a rather low level of arousal, like sadness or frustration can have perturbing effects on driving ability [37]. An example is shown in [38] where the influence of terror attacks on driving performance was examined: in Israel an increase of traffic accidents by 35% was observed on the third day after terrorist attacks. Sadness seriously affects the level of attention of a driver and therefore endangers the safe operation of a vehicle. As frustration and sadness usually coincide with a certain degree of passiveness or resignation, reaction time in critical situations increases.

Apart from safety aspects, when thinking of the car as a "virtual companion", the automatic recognition of sadness as an emotional state maybe one day enable the system to cheer up the driver and thus deliver also enhanced driving pleasure besides increased safety.

2.2. Recognition and Control of Driver States. So far, we have pointed out the enormous effect that affective states and emotions have on driving abilities and listed the most dangerous affective states which prevent safe driving. However, the need for automatic in-car emotion recognition and driver state detection only becomes evident when examining adaptation or even "countersteering" strategies that can easily be implemented provided that the drivers emotion is determined accurately. The aim of affect recognition is to provide a kind of "state variable" which serves as input

for subsequent processing in emotion-sensitive accessories, aiming to improve not only driving comfort but also safety [5, 36, 39, 40]. Thereby secure driving can be supported by either attempting to improve the affective state of *the driver*, which would mean making the driver "happy" or at least directing her or him into a neutral emotional state or by adapting *the car* with the emotion of the driver [41, 42]. Both strategies rely on proper emotion recognition and were shown to improve driving performance and reduce the risk of having an accident.

2.2.1. Countersteering Negative Emotions. To reduce the stress level of the driver, dialogue strategies can be adapted to the current workload [43, 44]. Since voice messages potentially distract the driver, an adaptive system would deliver messages only when the driver's stress level is low. A method to avoid stress caused by traffic jams could be to warn the driver in time as soon as he or she intends to use a route for which congestion had been reported to the system. This, of course, is a desirable feature regardless of affect aware technology, but has the possible benefit of reducing negative emotions.

A possible approach towards making, for example, an angry driver aware of the dangerous driving style, resulting from her or his increased level of arousal, would be to encourage better driving via voice response [45] or to make appropriate alerts [36] (e.g., making the driver aware that the current traffic situation demands more thoughtful actions than emotional actions). Calming down the driver or increasing the consciousness of critical manoeuvres caused by aggressive driving would be a typical method of verbal countersteering that a "virtual companion" might perform in the far future, thereby replacing a reasonable, observant human codriver.

In [5] it is suggested that the car could become more or less talkative depending on the affective state of the driver. Sleepiness is a driver state which requires increased communicativeness of the virtual companion in order to involve the tired driver into a conversation and thereby aiming to make her or him attentive again or even to prevent her or him from falling asleep. This is equivalent to what a good codriver would do to compensate the sleepiness of the driver. However, according to [26], the only safe countermeasure against driving while being sleepy is to stop driving. This advice also counts to the useful alerts an intelligent car could give, provided that the driver state "fatigue" is recognised reliably.

To countersteer detected confusion, an emotion-sensitive system could provide help or more detailed explanations concerning the functionality which the driver is about to use. Complicated information or entertainment systems benefit from automatic guidance through menus. that could be triggered by the detection of irritation. As far as confusion or nervousness due to traffic situations is concerned, it was shown that particularly elderly people profit by the recognition of irritation and subsequent driving support [46] leading to better driving performance and to higher confidence while driving.

2.2.2. Adapting Car Functionalities to Driver Emotion. Apart from trying to influence the driver's emotion in a positive way, adapting user interfaces to the user's affective state can also reduce the risk of accidents and potentially leads to higher driving pleasure. Experiments indicate that matching the in-car voice with the driver's state not only encourages users to communicate with the system, but also improves driving performance [41]. Choice of words, intonation, and tone of voice are important aspects of communication and should be adapted to the emotion of the conversational partner to make the dialogue more natural. Further, it is known that the words which are used to inform the driver of bad driving performance are much more effective when they ascribe the bad performance to the driving environment rather than to the drivers themselves [47]. A voice that matches the driver's emotion increases the connection between the user and the voice and, like most other adaption strategies, corresponds to what a human codriver would do.

As an important step towards enhanced and reliable speech recognition, adaptation of speech recognition engines to the driver's current emotion is a technique that has prevailed in increasing the robustness of speech recognition systems [48, 49]. Both the acoustic realisation of a spoken utterance and the choice of words are highly dependent on the speaker's emotion [50] which makes it necessary to adapt acoustic models (emotionally coloured speech has a different spectral pattern than normal speech, e.g.) as well as the language model of the speech recogniser in order to maintain automatic speech recognition performance for emotionally coloured speech. This again stresses the need for emotion recognition in the car as a major component to guarantee not only safety and comfort but also accuracy and robustness of other functionalities like automatic speech recognition.

In this context the design of emotion dependent speech dialogues for natural interaction with in-car systems will be an upcoming challenge. Besides speech technology improvements, new concepts regarding the interaction design with other input and output modalities are also relevant. Flat menu hierarchy, "one click" solution, user interfaces with seamless multimodality and usage of handwriting recognition (for almost blind text input on a touch display without having to look at buttons) are some examples.

2.2.3. Modalities for In-Car Emotion Recognition. As in many pattern recognition disciplines, the best emotion recognition results are reported for multimodal recognisers [44, 51] that make use of more than one of the four major modalities. These are audio (e.g., [52–54]), video (e.g., [55, 56]), physiology (e.g., [57–59]), and driving style (Section 5). However, not every affective state can be detected equally well from every modality. It will, for example, be hard to reliably recognise sleepiness from the speech signal since a tired driver probably does not talk. Yet, visual information (frequency and duration of twinkling as a trivial example)—if combined with infrared illumination at night time—will be a well-suited indicator for fatigue, and driving style is a good indicator of distraction as we will see in Section 5.

For the recognition of driver states like anger, irritation, or nervousness however, the audio channel was proven to be valuable [19, 40, 60]. This seems little surprising when considering how strongly for example anger is correlated to simple speech features like volume (and energy, resp.) or pitch. The great advantages of the speech modality are low hardware costs, relatively low apparent observation, and high reliability. Furthermore the user is able to control how much emotion is shown, which of course is a disadvantage for constant and reliable driver monitoring; audio information simply is not continuously present if the driver does not constantly speak. Further, not all speech captured by the microphone may be relevant in such an open microphone scenario, which makes recognition tasks more difficult [61].

Recognisers exploiting visual information have been applied for the detection of emotions like anger, sadness, happiness, disgust, fear, irritation, and surprise [44, 62, 63]—mostly from facial expressions. For interest detection, for example, the visual modality also seems to be superior to the audio modality [64]. In contrast to speech, video is omnipresent, however, the use of visual information implies slightly higher hardware costs, and increased observation feeling.

More extensive approaches (at least from the sensory point of view) to measure emotion also include physiology exploiting data from electromyograms, electrocardiograms, respiration, and electrodermal activity [65] measuring quantities like heart rate or skin conductance. These methods are at present mainly used for research and not for Human-Machine Interaction as they require a great amount of noncommercial hardware. Depending on the type of signal, the hardware costs of physiological measurements can also be marginal, however, user acceptance will still be low as the driver—apart from the process of wearing such devices—has a strong feeling of "being watched" and controllability is not granted.

The use of driving style as a modality for emotion recognition is quite self-evident and less costly although not investigated very intensely so far. The fact that driving style and affective state are highly correlated was outlined in Section 2.1 and can be utilised for in-car emotion recognition. A small study on a system using this modality is presented as an example in Section 5.

Motion of the driver in her or his seat is another method to measure nervousness or activity [66], for example. However, we have to be cautious not to confuse nervousness with backache or other problems which could make the driver move more than usual. It has to be carefully researched if backache and nervousness produce different movement patterns. Contextual knowledge is mostly to be seen as additional knowledge source, yet certainly highly reasonable to improve audio- or video-based recognition performance.

2.3. The Socially Competent Human-Like Car. Besides improvements in driving safety by monitoring the driver's emotional state, the upper class car of tomorrow will also be "socially competent", that is, more human-like with respect to verbal and nonverbal communication and interaction skills and, possibly somewhat limited, understanding of nonverbal meaning and contextual information. The car can be expected to be able to interact with driver and passengers in a way quite natural to us as humans. It could be able to serve as a virtual companion or secretary and assist the driver in difficult situations. The car will likely be more like a real human codriver than like the touchscreen based interfaces found in today's cars [67, 68], however ensuring controllability at all times, that is, the driver will be able to switch off the "talking interface". One might now ask the question why cars have to be more human-like. The answer is simple: because people want it and a more human-like interface simplifies interaction with advanced technology. The trend is clear, already today people driving upper class cars demand for the latest state-of-the-art technologies in their cars. There exist route guidance systems, finest HiFi entertainment and integrated cell-phones including messaging, calendar and e-mail systems, all synchronisable with your home, work and laptop computer. Still in a development stage are topics like real internet access, just-in-time information about the environment, and real time traffic alerts and warnings [69]—all functions controlled via natural language voice interaction between driver and car. In Section 4, we present the results of a small survey regarding the user acceptance of such technology in the car.

A major problem arising along with the growing number and complexity of in-car entertainment and communication systems is the increased distraction of the driver caused by these systems. When changing your route while driving, for example, the display of your route guidance system will capture your visual and cognitive attention for some time. The same is true for changing the radio station or your music selection in the on-board entertainment system. If these distractions remain few, the driving safety is not affected notably. However, if more tasks are added, especially reading e-mails, retrieving background information about points of interest or communication with other people over the phone, driving safety will certainly suffer, if these systems do not change their way of interfacing with the user [68]. State-of-the-art in-car systems can be controlled via Automatic Speech Recognition (ASR), however robustness is still an issue here. Moreover the user is restricted to a well-defined grammar, and cannot say what he would like the system to do in his own words or even nonverbally (e.g., confirmation or rejection only by tone of voice).

In this section we list reasons and show various sources that indicate a demand for in-car Human-Machine Interfaces to become more human-like in the near future. As mentioned in the previous paragraph, there is primarily the issue of improving driving safety. Section 2.1, shows that this is of major importance and Section 2.2 demonstrates that various ways exist in which technological systems can directly improve safety by incorporating human factors and giving useful feedback. However, it has also been shown that driving safety can be increased indirectly by making driving a more pleasureful experience [19, 41, 42, 70]. Therefore, if the pleasure of the driver is increased and the interfaces are designed more human-like as reported by [68], safety will automatically benefit, even though in-car systems are getting

more complex and are able to offer things not primarily related to driving.

This is an important factor in competitive markets. Users' demands for more technology in the car are obvious. The literature indicates that people experience more driving pleasure in newer, safer, and more comfortable cars, as proven, for example, in [70].

In the following section the demand and feasibility of enhancing users' driving pleasure and enabling the user to be more productive, while still focussing on driving, is discussed in more detail. Section 2.3.1 will summarise existing work that deals with the effects of driving pleasure on driving performance and methods shown to enhance driving pleasure. Section 2.3.2 shows methods that enable the driver to be more productive without affecting the driver's focus on the road.

2.3.1. Enhancing Driving Pleasure. Users of in-car Human-Machine Interaction nowadays most often get frustrated, if a system does not understand their intentions and interface design is complex and nonintuitive. Numerous publications exist that try to improve the interfaces by optimising the amount of time users spend on data input and by re-structuring menus in order to access items more quickly, for example, [71]. However, all of these approaches focus only on traditional input and output modalities: haptics/touch and speech for input and displays and speech synthesis for output. The speech-based systems take the first step in making the interaction easier by allowing the driver to operate the system in a hands-free mode. However, the dialogue structure usually follows a fixed set of rules the user has to adapt to [68]. If communication errors occur, due to faulty speech recognition, for example, the user can quickly become annoyed and stressed [68, 72]. Hearing a prerecorded response like "sorry, I did not understand you" over and over again, is likely to provoke anger.

More human-like systems should use multiple modalities (especially adding the visual channel) and thus be able to detect communication problems as quickly as possible, as pointed out by [72]. If the system shows awareness of communication problems and quickly offers alternative solutions and personalised help [73] or uses partial information to re-request specific details that were incorrectly recognised by the speech recognition unit, for example, general perceived reliability and robustness of the system improves. Literature suggests that generally, the more robust a system is, the better the user acceptance is [74]. User studies, such as [73] further suggest that if a driver needs help, the system should be very sensitive to her or his needs and preferences in order to solve the driver's problem as easy and quickly as possible.

A "socially competent" car in the role of a virtual companion can engage the driver into conversation and thus will give the driver the feeling of not being alone [75]. It can further give helpful assistance and driving hints, assuring the driver that there always is somebody to help. Such a form of communication is in strong contrast to the traditional and impersonal way of interfacing via menu structures or well defined dialogue schemes. According to traffic statistics in the USA published by the US department of Transportation (http://www.bts.gov/) for the year 2001, the overall mean occupancy of passenger vehicles varies between 1.1 and 1.3, indicating a large amount of single drivers. Considering this fact, a virtual companion might be well appreciated by a large number of drivers. According to these reports, the mean occupancy is lowest for drives to work (1.14) and work-related trips (1.22). The overall mean occupancy for all observed trips not being much above the number for work-related trips, is an indication that a large amount of traffic is caused by commuters and work-related driving, where car-pooling often is not possible or not appropriate.

Another way to improve the driving experience is personalisation [76]. This trend can be observed in almost all other user interfaces, from full-blown computers to basic mobile phones. Practically all of these devices allow to change settings like the background image, color of the interface or configure favourite menu items. A "socially competent" car should be able to detect who is currently driving, judge emotion and behaviour based on past experiences and automatically adapt a conversation to the preferences of the current driver. This will give the driver the feeling that she or he is in control and the car adapts to her or his wishes and needs and thus also increases driving pleasure.

The car may also offer a good personalised music selection, for example. Music is known to improve mood by directly affecting physical brain processes [77]. Music thus contributes to the overall driving pleasure. However, care has to be exercised when selecting the style of music. Choosing too relaxing music may make the driver tired, choosing music that the driver does not like may annoy her or him, choosing happy music if the driver is sad is also likely to lead to opposite results. The user's personal preferences and the user's current mood have to be considered in such a decision.

Modern upper class vehicles fine tune the engine sound perceived by the driver using a considerable number of hidden speakers throughout the passenger cabin. In some situations the driver might be in a bad mood and bothered by the disturbing sound of her or his engine. An emotionally sensitive car could sense the driver's mood and adjust the engine (the motor of the car) sound (especially as perceived inside the car) based on good guesses or learnt preferences.

2.3.2. Enabling the Driver to be More Productive. Since time is precious, many drivers want to be able to use the time while driving to communicate with other people, access information like news and weather forecast or check reservations and bookings, for example. Today's in-car information systems in combination with mobile phones practically allow drivers to do all these tasks, however, most of these tasks cannot be done safely while driving. Interfaces are still designed in a traditional way like most other Human-Machine Interfaces, using a screen to display information combined with haptic input via buttons, knobs, and touch devices. Some systems use speech input in certain areas such as dialling of phone numbers. Yet, the driver has to spend some cognitive and visual effort on communicating with the system. He or she must learn to interact with the system—it is not the system

that learns how to interact with the user. The latter, however, should be the case in a user-friendly, human-like system [68].

Most users, especially elderly people or people with less practice in interacting with computers, will experience problems properly using an in-car driver interface and thus will require more time to access and use various features [78]. During driving they do not have the time to deal with the system, which leads to poorer acceptance of the system. One could imagine situations where there is a human codriver present in the car. Most of the time, the driver will certainly instruct his or her codriver to enter the new route into the route guidance system, check the weather forecast or call somebody via the cell-phone, for example, instead of doing these tasks on his or her own.

This is exactly where it becomes obvious that a "socially competent" virtual companion would be indeed very helpful for single drivers, especially those unfamiliar with computer interfaces. As many cars are occupied only by one driver, as pointed out in Section 2.3.1 this is an important issue. Communicating with the virtual companion like with a human codriver would increase user acceptance. In the literature the term virtual codriver (VICO) [67] is often used to refer to a human-like virtual companion. Of course a short period of time will be required to get used to the, at first, strange idea of an actually naturally talking car.

Besides being a helpful aid for single drivers, a "socially competent" virtual companion can also be of great help for drivers with passengers requiring special assistance. Children in the back seats can significantly distract the driver if they require too much of her or his attention. Such a virtual companion could detect such a situation and take some load off the driver by engaging the children in conversation or begin telling them stories or showing them cartoons, for example, via a rear seat entertainment system.

At this point it becomes most obvious that for a "socially competent" car it is also necessary to estimate the interest level of the conversation partner. If the entertainment system detects that the children are not interested in, for example, the film currently shown, it probably is time to change to something different in order to keep the children's attention. Also, the driver should not be bored by noninteresting information.

In Section 3 we summarise use-cases like the above that could be handled by a "socially competent" virtual companion, like reading and writing e-mails and making reservations while driving.

3. Exemplary Use-Cases

In order to design human-machine communication in future upper class cars more naturally and intuitive, the incorporation of innovative applications of pattern recognition and machine learning into in-car dialogue interfaces becomes more and more important. As discussed in the previous sections, emotion recognition is an essential precondition to create a social competent car that can talk to the driver and provide a "virtual companion". In this section we discuss specific use-cases for emotion related technology in the car for both fields, namely the safety-related tasks of driver state monitoring and control of driver emotions, and the tasks related to enhancement of driving pleasure and productivity, such as multimodal and affect sensitive interfaces. We start our use-case overview by giving a brief summary of the state-of-the-art in in-car driver assistance and entertainment systems.

3.1. State-of-the-Art. While affect aware technology is missing in today's automobiles due to the lack of user adaptable and autonomous, reliable technology, speech recognition has started to mature in the automobile market. The most obvious example is navigation systems where the destination selection can be performed via speech input. This speech recognition is based on templates which are stored during a training phase when the user adds a new destination and pronounces its name several times. More advanced systems, are based on subword modelling (phonemes) and include a universal acoustic model. They are thus able to recognise speech input without the need of recording several templates. Some minor voice adaptation might need to be performed in the same way as in modern dictation systems. These systems allow for a voice based command-like interface, where the user can change routes by command ("fast route", "short route"), change the view, or have traffic information read out aloud. Entertainment systems can be controlled in a similar fashion by commands such as "change station", "next song", or even by pronouncing a song title or artist. Yet, these systems are restricted to a set of predefined commands and do not allow for flexible interaction. The user has to know the capabilities of the system, he has to know "what" he can say. Future systems, as proposed in the use-cases in the following sections, must be able to accept all input, filter out information they understand, associate it with available car functions, and "tell" the user what his options are.

3.2. Safety-Related Use Cases. For the safety-related tasks we present three different categories of use-cases, which are countersteering strategies, adaptation strategies, and communicating the driver's emotional state (e.g., anger/rage, fatigue, and high workload, stress, or uncertainty) to other vehicles.

3.2.1. Countersteering Strategies. This category contains use-cases which aim to "countersteer" negative affective states in order to guide the driver into a happy or neutral state which is known to be best suited for safe driving [5, 18, 19], since most other emotions (anger, fatigue, stress, confusion, nervousness, sadness, etc.) negatively affect driving capabilites like goal generation, evaluation, decision-making, strategic planning, focus, and attention [9–11]. Depending on the context, different voice responses for angry drivers can be given, intending to encourage better driving, make appropriate alerts or calming down the driver. Further, a virtual codriver can react to detected sleepiness—which constitutes another dangerous driver state—by keeping the driver awake or bringing the vehicle to a safe halt in case of danger, if the traffic situation permits (e.g., stopping on a busy highway

is too dangerous, the car has to be directed towards the side lane before it is stopped). Possible measures against stress, confusion, nervousness, and sadness can also be addressed by the virtual assistant through intelligent dialogue strategies. Thereby the responses or actions of the intelligent car always depend on the amount of available contextual and background information regarding the reason for the specific affective state. Especially in situations like stress, the virtual codriver can actively help to reduce the driver's workload, for example, by offering intelligent solutions for tasks related to the on-board entertainment and communication system or temporarily disabling such functions if the traffic situations require the driver's full attention.

3.2.2. Adaptation Strategies. Adapting the personality of an automated in-car assistant to the mood of the driver can also be important. A badly synthesised voice or an overly friendly, notoriously the same voice is likely to annoy the driver which soon will lead to distraction. Therefore, as an important adaptation strategy, matching in-car voice with the driver's emotion is beneficial, as has been found in, for example, [41, 47]. Different parameter settings for the synthesis of emotional speech for different emotions need to be used, as given in [79–82], for example. Other use-cases related to adaptation are emotion dependent spoken language understanding and model adaptation for speech recognition engines. These techniques serve the purpose of improving the accuracy of the in-car speech recogniser, since an inaccurate system is also likely to annoy and distract the user, instead of assisting the driver.

3.2.3. Communicating the Driver's Emotional State. The third category consists of use-cases that describe how a driver's state can be communicated to others. Locating potentially dangerous drivers can aid the driver assistance systems in other vehicles to warn their drivers more timely. Methods of car-to-car communication for preventing road rage are developed by some automobile manufacturers, for example. Further applications include monitoring passengers—especially children—and other road users while driving, to reduce the driver's cognitive workload, logging the driver's emotion to derive statistics for research purposes, and automatically triggering emergency calls in case of accidents, severe pain or dangerous situations.

3.3. Driving Pleasure Related Use-Cases. Similar to the safety-related applications of in-car emotion recognition, the use-cases related to driving pleasure can also be grouped into three different categories: enabling of a mood adequate human-machine dialogue, adaptation of surroundings, and increasing productivity.

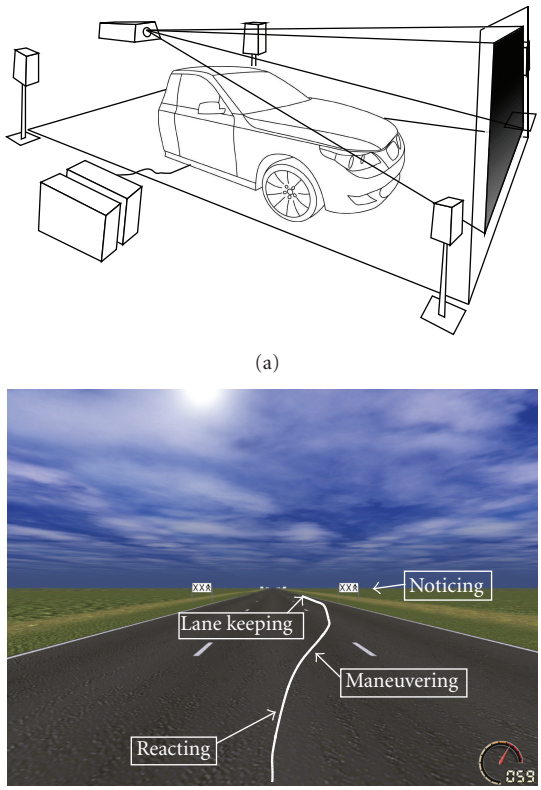
3.3.1. Mood Adequate Human-Machine Dialogue. Personalised and "socially competent" small-talk belongs to the first category and is a key feature of a "virtual companion". Thereby emotion serves as contextual knowledge that indicates how the dialogue system has to interpret the output of the automatic speech recogniser (e.g., the use of irony may depend on the user's emotional state, also there seems to be a reduced vocabulary in highly emotional speech, such as short angry commands or comments). Such dialogues do not only depend on the current words uttered by the user, but depend also on contextual information like time of day or weather. Similar use-cases are adaptive topic suggestion and switching, dialogue grounding, and reactions to nonlinguistic vocalisations like moaning or sneezing. Further, multimedia content analysis methods enable the car to deliver information from the internet which suits the current interest and affective state of the driver (e.g., love poems if the driver is in love, or only allowing happy news if the driver is in a happy state). Observing the driver's workload also enables the car to adapt the level of entertainment to the current traffic situation. Incoming and outgoing calls can be managed by a "phone guide" who takes into account the affective state of both the driver and the conversational partner. The latter can be determined from speech while the system converses with the caller (i.e., asking for the caller's identification and purpose/importance of his call) before putting him through to the driver.

3.3.2. Adaptation of Surroundings. Depending on the driver's mood, the in-car ambience can be adjusted. This can be done by automatic selection of mood adequate music, for example. Moreover, engine sound, ambient light, and air conditioning can be adapted according to the driver's affect.

3.3.3. Increasing Productivity. Finally, potential use-cases for a virtual codriver can be derived from the goal to increase the driver's productivity. Thereby calendar functions, handling of e-mails, internet access, and automatic translation are relevant as aspects that are likely to be welcomed by car buyers. However, the role affective computing takes in such technological advances is not fully researched, yet. Also increasing productivity on the other hand means higher workload for the driver, and thus reduced focus on the road leading to reduced safety. The aspect of increasing productivity thus should only be addressed if it can be ensured that these tasks do not in any major way keep the driver from his primary task of controlling the vehicle. This would be the case if the virtual codriver had a fully natural speech interface and the capability to robustly understand the driver's intentions from minimal input.

4. User Acceptance

It is important to assess acceptance and success of any new technology as soon as possible to determine whether efforts in developing the technology are well spent. Since it is a well known issue that too much technology might irritate or confuse users or make them feel observed, we address these issues in a user study designed for in-car affective computing. The basic idea is to set up a car with a simulated virtual codriver in a Wizard-of-Oz experiment. Users are asked to perform several tasks in the simulation while being assisted by the virtual codriver. The users' experience with the system is determined via multiple questionnaires which are filled out



(a)

(b)

Figure 1: Driving simulator and simulation software. (a) Setup of the driving simulator. (b) Lane-Change-Task (the signs indicate that a lane change is to be performed).

after the experiment. The next section describes the setup and procedure of the Wizard-of-Oz (WoZ) experiment. Section 4.2 presents the findings of the survey.

4.1. Wizard-of-Oz Setup. In order to create a realistic driving scenario in a safe and controllable environment, a driving simulator was used. It consists of half the body of a real BMW 5 series vehicle in front of a large screen (see Figure 1(a)). The car's controls (i.e., accelerator and brake) and the steering wheel are used to control the simulation-software, which renders a 3D animation of a road and scenery on to the projected screen. The lab is equipped with two sound systems, one outside of the car to play back the engine and environment sounds, the other inside the car, which is used by the operator to give instructions to the subjects, to play music from the on-board entertainment system, and to output the voice of the virtual in-car assistant. The Lane Change Task [83] was used to simulate a primary driving task. Thereby, a test person has to drive a street of three lanes and switch to the lane, which is signalled by signs on the side of the test track (see Figure 1(b)). Additionally, a simple driver-information system is implemented, which can be controlled with a multifunctional device—the iDrive controller placed in the driving simulator. The following functions are implemented into this system:

(i) Input of navigation destination.

(ii) Switching between three alternative options of navigation routes.

(iii) Number dialling on the car phone.

(iv) Viewing and editing calender entries.

In order to simulate as much use-case customised support by the system as possible, the supervisor was able to fully remotely control the driver information system. The virtual in-car assistant's voice is simulated by the Wizard-of-Oz operator. Therefore, the operator's voice is recorded by a microphone in the control room, and after applying online effects, simultaneously played back via the car's centre speaker. Instructions are given to the test persons on what task to perform next (it was made clear to the subjects that these instructions did not belong to the virtual driver assistance system). The instructions have been prerecorded to ensure the same test conditions for all subjects. The following instructions were used for the tasks described in the following paragraphs.

(i) Drive straight with moderate velocity.

(ii) Drive straight with high velocity.

(iii) Enter *Schrobenhausen* as destination.

(iv) Scan the calender for today's appointments.

(v) Call your office to inform them of your late arrival.

The experiment was a first-contact situation for the test subjects, and they did not receive instructions on the capabilities of the driver-information system. Subjects were asked to imagine that is was an ordinary Monday morning and they were starting their usual drive to work.

Welcome Dialogue. After taking a seat in the car, the driver is greeted by the car with a short dialogue. Thereby the user is asked whether he or she is driving the usual way to work, or he or she requires navigational assistance, and whether he or she would like to listen to music.

Driving Only. The driver is now asked to start driving a simulated test track with low speed to get used to the primary driving task in the simulator environment. This test track includes the Lane Change Task (see above). Next, the participant is asked to drive the track at a higher velocity, which induces a higher load due to the primary driving task. During this situation the following use-cases are simulated.

Detection and Countersteering High Workload. The operator instructs the subject to enter a navigation destination in parallel. Now, the system (in our WoZ case simulated by the operator) will detect decreased attentiveness in the primary task, ask the driver to pay more attention to his primary driving task via speech output, inactivate the display elements, and offer the user the option of speech-based input of the destination.

Assisting Confused Drivers. Next, a road congestion is simulated. The user is now instructed to inform his office of his delay via his cell phone. The dialling does not work, however, due to a simulated bad network. The reason is not immediately apparent to the user, who only realises that his call is not being connected. The system detects the induced confusion and offers to connect the call once the network is available again. The wizard was instructed to act once he recognised confusion because the user was hesitating or expressing his confusion verbally.

Obtaining Information from the Internet. The subject is now instructed to scan his calendar for appointments. An appointment in a distant city is scheduled for the next day. A comment that a hotel room must be reserved is attached to the appointment. If the subject does not ask the system for available hotels by himself, after a defined timeout the system will ask the user if hotel information is to be obtained from the internet now. The system will guide the user through a hotel reservation process.

Handling Incoming Calls. After finishing the hotel reservation, an incoming phone call is simulated. However, there is no way apparent to the user to answer the call. Again, the system will detect that the user is not answering the call and will ask for the reason while at the same time offering help, that is, to either ignore the call or to accept it.

Smalltalk. Now the system initiates a dialogue, where it comments on the driver's busy day, and the bad weather, and asks the driver whether a different radio station would be preferred. Finally, an updated traffic report is received with the information that the congestion has not yet cleared. This report is automatically interpreted by the system, and the user is given the option to select three alternative routes from the system display, which will bring him directly to the location of his appointment, instead of his office.

Adapting to the Driver's Behavior. All the use-cases described so far are fixed, and thus common for all subjects. In addition to these planned scenarios the operator was trained to react individually to the subjects responses and comments, adapt his output voice to the user's state (thereby changing his tone of voice to match the user's tone of voice in the current situation), and especially react to nonlinguistic behavior such as laughing, sighing, or hesitation, where it seems appropriate.

4.2. Evaluation and Results. After finishing the experiment, every test subject was asked to fill out a questionnaire, which consists of four parts: The System-Usability-Scale (SUS) [84], and the SEA-Scale (Subjectively Experienced Effort) [85] for rating specific scenarios, and the Attrak-Diff system (http://www.attrakdiff.de/), a questionnaire composed of a semantic differential, for rating the complete system. Since Attrak-Diff is a general system for rating product attractivity, additionally a set of extra questions concerning our specific setup was used.

Table 1: Results for the System Usability Scale (SUS) and Subjectively Experienced Effort (SEA) scale for four selected tasks. SUS: maximum score is 100 (best usability), SEA: maximum score 120 (highest workload: worst), except for results marked with *, where the maximum score of 120 indicates the maximum perceived *decrease* in workload (thus, 120 is best).

Scenario	SUS [0–100]	SEA [0–120]
Intelligent virtual agent support in stress situations	72.2	67.8*
Assisting confused drivers	76.0	66.9*
Smalltalk with virtual agent	76.2	22.2
Adaption of agent speech to the driver's emotional state	70.0	24.3

4.2.1. Description of Participants. Thirteen subjects (twelve male and one female) took part in the experiment. The average age is 27.8 years with a standard deviation of 3.1 years. All of them had a driver's license and were interested in new and innovative technical products. The average yearly mileage of each subject is approximatively 14 000 kilometers.

4.2.2. System Usability and Subjectively Experienced Effort Scales. The analysis of the System Usability Scale (SUS) was performed with the method proposed in [84]. For each use-case, a total score was determined. This score reflects the user's impression of the system for the respective scenario. The maximum assignable score is 100, which corresponds to a completely positive evaluation of the system.

Table 1 shows results for the SUS and SEA scales for four selected tasks. Considering the early prototypical stage of the test system (i.e., with respect to look and feel, and range of functionality), the obtained SUS scores are a promising basis for further system improvements, since ratings above 50 are generally in favour of the system in question. The best scores were obtained for assisting confused subjects and smalltalk with the in-car agent. While the first is to be expected, it is not quite obvious that smalltalk does enhance the system's usability feeling.

The SEA scale describes the subjectively experienced workload for each particular scenario. A high score (maximum value 120) indicates a high perceived workload. Thus, lower values indicate better performance with respect to reducing the driver's workload and keeping his or her focus on the road. For the first two scenarios, "stress", and "confusion", however, a modified scale was used, where a high value (again maximum of 120) indicates the subjectively perceived *decrease* of workload. The result can also be found in Table 1.

Concluding, every scenario is evaluated positively on average. Both the SUS and the SEA scale show good results regarding the use of the system in spite of the prototypical system setup. The subjectively perceived workload decreased noticeably, if the car gave support to the test person ("stress", and "confusion" scenarios on the SEA scale). This is a good basis for further development of such driver state-aware functionalities.

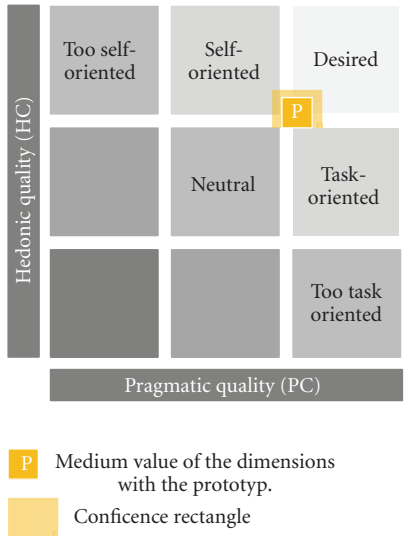


Figure 2: Results Attrakdiff-Portfolio Representation.

4.2.3. Attrak-Diff Rating. With Attrak-Diff a product is evaluated with respect to the following four dimensions.

(i) Pragmatic Quality (PQ): usability of the product.

(ii) Hedonic Quality-Stimulation (HQ-S): support of needs in terms of novel interesting and stimulating functions.

(iii) Hedonic Quality-Identity (HQ-I): identification with the product.

(iv) Attractiveness (ATT): global value of the product based on the quality perception.

The so-called portfolio representation (the result in the 2-D space spanned by PQ and HQ) determines in which character-zone the product can be classified. For this study, the Attrak-Diff was evaluated for the entire system with all its new ideas and concepts. Individual features were not evaluated separately. The resulting portfolio presentation is shown in Figure 2.

The system is rated as "rather desired". However, the classification is neither clearly "pragmatic" nor "hedonic", because the confidence interval overlaps into other character-zones. So there is room for improvement in terms of usability (PQ and HQ). The small confidence rectangle indicates a high agreement among the test subjects.

4.2.4. System Specific Questionnaire. A questionnaire composed of eleven questions using a five point scale was used ("Strongly Agree" (value = 1) to "Strongly Disagree" (value = 5)) for the custom evaluation of the system as a whole. For each question, the mean value of the ratings and the standard deviation (σ) were calculated. The results are summarised in Table 2 and briefly discussed in the following paragraphs.

Nearly every test one thinks that a talking car is reasonable (mean 1.4, $\sigma = 0.4$) and feels rather not observed (mean 3.8, $\sigma = 1.1$) by the car. The question, if the car was disturbing, was evaluated as "moderate" with a light trend to "rather not" (mean 3.5, $\sigma = 0.3$). The test persons would rely with a mean value of 2.7 ($\sigma = 0.8$) on suggestions given by the car. The question, whether the users felt the car would help them to handle difficult driving situations more easily, gave a clear positive result, with a mean value of 1.7 ($\sigma = 0.7$).

Table 2: Results of system specific questionnaire composed of eleven questions. Mean (μ) and standard deviation (σ) of ratings on a five point scale ("Strongly Agree" (value = 1) to "Strongly Disagree" (value = 5)).

Question	μ	σ
A talking car is reasonable	1.4	0.4
I feel observed by a talking car	3.8	1.1
I feel disturbed by a talking car	3.5	0.3
I would rely on suggestions given by the car	2.7	0.8
I feel the car helps to handle difficult driving situations	1.7	0.7
A car should react to my emotion	3.2	1.1
Starting self-initiated dialogs is desired	2.8	1.3
Automatic evaluation of my stress level and emotional state is desired	2.8	1.6
It is helpful if the car can support me in confusing situations	1.3	0.2
I like the ability to request information from the internet via natural speech input	1.5	0.8
It should be possible to mute the car's voice	1.3	0.4

Unclear are the results of the questions, whether a car should react to the driver's emotion (mean 3.2, $\sigma = 1.1$) or should determine the stress-state of the driver (mean 2.8, $\sigma = 1.6$), and whether the car should start talking initiated by itself (mean 2.8, $\sigma = 1.3$). The high standard deviations of the answers to these questions indicate that the individual subjects do have quite clear preferences, thus no unifying conclusion for all users can be drawn. Likewise, the recommendation based on this study would be to provide easy ways to disable such functionalities or have them disabled by default and let the users decide to enable them.

The last four questions show all positive results. The test subjects agreed on the fact that it would help, if the car was able to support the driver in confusing situations (mean 1.3, $\sigma = 0.2$), which is in line with the SUS and SEA scale evaluations. Moreover, they liked the car's ability to request information from the internet via natural speech input (mean 1.5, $\sigma = 0.8$).

Overall, the test persons stated that a talking car—as simulated via the Wizard-of-Oz—makes sense, and they do not feel observed or disturbed. This is an indicator for a good acceptance of such a product.

However, the driver wants to be the master of the situation and makes her or his own decisions, because not all test persons would rely on the car's suggestions, and high standard deviations were observed for the driver state monitoring questions. Virtually all subjects wish to have a functionality, which allows the user to mute the car's voice (mean 1.3, $\sigma = 0.4$). From this point of view, the evaluation of the smalltalk-feature gets more comprehensible. Some

subjects commented that this feature would be disturbing, if it happened too many times.

The unclear results regarding the recognition of emotions and stress-states may also relate to this as well as the fact that these functions could not be implemented consistently enough in this experiment (Wizard-of-Oz). Such a consistent evaluation would require ways of reliably and reproducibly inducing emotional and stress-states, which is a very difficult task that can never be performed perfectly. Thus a very large number of subjects is required for these evaluations.

5. Driver Distraction Detection

Driver inattention is one of the major factors in traffic accidents. The US National Highway Traffic Safety Administration estimates that in 25% of all crashes some form of inattention is involved [86]. Distraction (besides drowsiness) as one form of driver inattention may be characterised as: "any activity that takes a driver's attention away from the task of driving" [87].

In this section we show how reliably driver distraction can be detected using adequate machine learning techniques. The motivation for detecting whether a driver is distracted or not could be adaptive driver assistant systems, for example, lane keeping assistance systems. These systems track the lane markings in front of the vehicle and compute the time until the vehicle will cross the marking. If the driver does not show an intended lane change by using the indicator to signal the change, the systems will use directed steering torques on the steering wheel to guide the car to the middle of the lane.

One problem with lane keeping assistance systems is that they can be annoying in some circumstances [88] since they do not yet respond to the driver's state or her or his intent but to lane markings and the car's speed. If it was possible to recognise a driver's state reliably, the system would give just as much assistance as the driver needed. This would allow for a greater safety margin without annoying the driver with false alarms in normal driving situations.

Our system for online driver distraction detection is based on modeling long-range contextual information in driving and head tracking data. It applies Long Short-Term Memory (LSTM) recurrent neural networks [89, 90] which are able to capture the temporal evolution of low-level data sequences via so-called *memory blocks*. Long Short-Term Memory networks have shown excellent performance in a variety of pattern recognition tasks including emotion recognition from speech [91].

5.1. Database. In order to train and evaluate our system we used data that was recorded during an experiment in which drivers had to fulfil certain "distracting" tasks while driving. The resulting database consists of 32 participants (13 female and 19 male). The car (an Audi A6) was equipped with the "Audi Multimedia System" and an interface to measure CAN-Bus data. Additionally, a head tracking system was installed, which was able to measure head position and head rotation. Head-tracking systems are not common in vehicles today, but the promising research in such cameras for driver state detection will lead to a higher installation rate in serial cars in the near future. So we decided to use head-tracking information in our approach as well.

Eight typical tasks (performed haptically) on the Multimedia Interface were chosen as distraction conditions:

(i) adjusting the radio sound settings,
(ii) skipping to a specific song,
(iii) searching for a name in the phone book,
(iv) searching for a nearby gas station,
(v) dialling a specific phone number,
(vi) entering a city in the navigation device,
(vii) switching the TV mode,
(viii) adjusting the volume of navigation announcements.

The procedure for the experiment was as follows: after a training to become familiar with the car, each participant drove down the same road eight times while performing secondary tasks on the in-vehicle information system. On another two runs the drivers had to drive down the road with full attention on the roadway. In order to account for sequential effects, the order in which the conditions were presented was randomised for each participant. Overall, 53 runs while driving attentively and 314 runs while the drivers were distracted could be measured. The "attentive" runs lasted 3 134.6 seconds altogether, while 9 145.8 seconds of "distracted" driving were logged (see Table 3 for experimental conditions).

An analysis of the influence on lane keeping of the different in-vehicle information system tasks [92] confirmed the tasks to be distracting. Thus, all these tasks were labeled as "distracted" compared to driving down the road with full attention (ground truth: "attentive"). Thereby we labeled runs during which a task had to be completed as *completely* "distracted" since the drivers were engaged with the task during the complete run.

Six signals were chosen for a first analysis:

(i) steering wheel angle,
(ii) throttle position,
(iii) speed,
(iv) heading angle,
(v) lateral deviation,
(vi) head rotation.

Steering wheel angle, throttle position, and speed are direct indicators of the driver behavior. Many studies prove the fact that visually distracted drivers steer their car in a different way than do attentive drivers. The same applies for throttle use and speed (an overview can be found in [93]). The car's heading angle and its lateral deviation in the lane rely on the amount of attention the driver is allocating to the roadway and may hence give useful information about distraction. Head rotation of the driver is an indicator of the driver's visual focus. While using the Multimedia Interface, which is located in the middle console just below

Table 3: Experimental conditions for driving data collection.

Experimental conditions	
Num. participants	32 (13 f, 19 m)
Age of participants	29 to 59
Driving experience	>10,000 km per year
Car	Audi A6 quattro 2.6 TDI
Road	Ayinger Str.
Num. "attentive" runs	53
Num. "distracted" runs	314

the dashboard, the main rotation of the head is to the right. So the head rotation is the most promising indicator of the head-tracking signals. Note that a trivial way of determining driver distraction due to the operation of the Multimedia Interface would be to simply detect, for example, the touching of the Multimedia Interface buttons. However, we decided to use signals that serve as general indicators of driver distraction in order to be able to also detect distraction which is not caused by the operation of the Multimedia Interface.

5.2. Experiments and Results. The database collected as described above was split into a training, a validation, and a test set. For training we randomly chose 21 drivers. The validation set consists of three randomly chosen drivers, while the system was evaluated on the remaining eight drivers. Thus, our evaluations are completely driver independent, that is, the results indicate the performance of the system for a driver which is not known to the system (the system was not optimised for a specific driver's style). The training set consists of 35 baseline runs (i.e., runs during which the driver was attentive) and 146 runs during which the driver was distracted. The test set contains 13 baseline and 51 "distracted" runs.

We evaluated the performance for different numbers of memory blocks (70 to 150) in the hidden layer of the LSTM neural network. The number of memory blocks is correlated to the complexity of the network, that is, the number of parameters which are used to describe the relation between inputs and outputs (see, e.g., [94] for a detailed description of the LSTM memory block principle).

Table 4 shows the results for *sample-wise* classification (i.e., quasi-time-continuous prediction every 10 ms) of driver distraction using the two classes "attentive" (baseline runs) and "distracted" (runs during which the driver was involved in a task at the Multimedia System). A total of 286 000 such samples (frames) is contained in the test set. The best F1-measure could be achieved with an LSTM network consisting of 110 memory blocks. Note that due to the imbalance in the class distribution, the F1-measure is a more adequate performance measure than accuracy. Thereby F1-measure is the harmonic mean of unweighted recall and unweighted precision. For the two-class problem, LSTM networks achieve an F1-measure of up to 88.7%. In Table 5 the classification of *complete runs* is evaluated by averaging the sample-wise LSTM predictions over an entire run. With the best LSTM configuration, an accuracy of 92.9% can be obtained.

By analysing the obtainable classification performance when using only single signals, we can get an impression of the relevance of the individual data streams. The best "single stream" performance can be obtained when using exclusively head rotation, followed by exclusive usage of steering wheel angle, heading angle, throttle position, speed, and lateral deviation, respectively.

Trying to get an impression of the accuracy of distraction detection when driver distraction is not caused by the Multimedia Interface, we tested the system on data that was recorded while the driver had to fulfil tasks like eating a chocolate bar or reading a letter. We found that the obtained F1-measure is only slightly worse for this scenario (83.2%).

Tables 4 and 5 reveal that driver distraction can be detected with relatively high reliability by modeling the temporal evolution of driving and head tracking data. Thus, an adaption of lane-keeping assistance systems which is based on sensor data already available in modern vehicles seems to be a viable and promising approach.

6. Conclusion and Outlook

Summarising all aspects discussed in the past sections, it becomes clear that emotions will be a key issue not only in general oncoming human-computer interaction, but also in the in-car communication.

As we have discussed, emotions affect many cognitive processes, highly relevant to driving, such as categorisation, goal generation, evaluation and decision-making, focus and attention, motivation and performance, intention, communication and learning. There is a need for controlling the driver's emotional state: the high relevance of an emotionally high valence was documented by a substantial body of literature—"happy drivers are the better drivers". This control of the emotional state will thus ensure a safer and more pleasant driving experience. At the same time too high arousal may lead to aggressive driving behaviour. For optimal driving performance, a compromise between too high and too low arousal must therefore be found.

Apart from externally induced states of intoxication (alcohol, drugs, medication) or pain, we had found anger, aggressiveness, fatigue, stress, confusion, nervousness, sadness, and boredom as main negative emotions and mental driver states of interest, and happiness as positive factor.

As basic strategies to control emotion, countersteering emotions was found next to adapting car functionalities to driver emotion. The in-car driver interface can thereby influence users' emotional states in several ways. To provide only few examples, angry drivers could be calmed down and could be made aware of their state, fatigued drivers could be stopped from falling asleep by engagement in a discussion with control of potential boredom for topic-switching, and confused drivers could be offered assistance regarding the current traffic situation.

The growing complexity of in-car electronics demands for new interfaces that do not disturb the drivers' focus on

Table 4: *Sample-wise* classification of driver distraction (attentive vs. distracted) using LSTM networks: accuracy, unweighted recall, unweighted precision, and F1-measure for different numbers of memory blocks.

# Memory blocks	Accuracy (%)	Recall (%)	Precision (%)	F_1 (%)
70	87.2	78.8	85.6	82.1
90	89.0	82.6	87.0	84.8
100	91.1	88.7	88.0	88.4
110	**91.3**	**89.5**	**87.9**	**88.7**
130	89.0	83.3	86.4	84.9
150	86.2	77.2	84.2	80.6

Table 5: Classification of driver distraction (attentive vs. distracted) for *complete runs* using LSTM networks: correctly classified runs, baseline runs correctly classified as "attentive", runs with task correctly classified as "distracted" (in percent) for different numbers of memory blocks.

# Memory blocks	Correctly class. runs (%)	Accuracy baseline runs (%)	Accuracy runs with task (%)
70	88.4	62.6	98.0
90	90.9	72.3	97.8
100	92.9	84.5	96.0
110	**92.9**	**86.7**	**95.2**
130	90.7	74.1	96.9
150	87.8	61.8	97.6

the road or annoy the driver because they are so difficult to use. Natural, human-like interfaces that quickly and tolerantly comprehend drivers' intentions are the key. In Section 4 we evaluated an intelligent driver assistance system, with which users were able to communicate naturally via speech. The evaluation suggests that such a system will generally be accepted by users, as long as they have full control over the system and can mute the system at any time. The driver will expectantly feel more comfortable and safe in such a car because he or she does not need to worry about not knowing how to use the system. The car can also serve as virtual codriver for single drivers, engaging the drivers in conversation and making them feel like having company and not being alone. Further possibilities of increasing driving pleasure are to offer personal settings, personalised conversation (greetings, small talk, etc.) and personalised in-car entertainment and environment customisation. Drivers simply prefer cars where they experience greater pleasure while driving and will therefore likely want to have "socially competent" interfaces in their cars. Further, drivers in the future are expected even more to use the time while driving productively, for example, listen to e-mails (with speech synthesis), make reservations or obtain information about the destination. In order to not interfere with the main task of driving, the driver interface must be operable in hands-free mode and quickly understand the user's intentions, without the user having to utter predefined commands. In this respect, future cars have to become more "socially competent", that is, be able to better understand their drivers' intentions adding the increasingly mandatory intelligent interpretation of multiple modalities such as speech, face and driving behaviour by incorporation of judgement of emotional and affective states.

As an example for the feasibility of driver state recognition, we presented an automated system for detection of driver distraction, which can be implemented in a car with the technology available today. Using Long Short-Term Memory recurrent neural nets, it is possible to continuously predict the driver's state based on driving and head tracking data. The strategy is able to detect inattention with an accuracy of up to 91.3% not dependt of the driver, and can be seen as a basis for adaptive lane-keeping assistance.

The presented paper shows the need, the acceptance, the feasibility and doability of intelligent and affective in-car interfaces. Yet, substantially more work is required to develop products which can be manufactured in series and which are robust enough for the end-user market. In this respect, more usability studies with a broader range of users in even more realistic driving situations (e.g., "out in the wild") are required. Further, implementations of actual prototype systems—instead of the presented Wizard-of-Oz approach—must be built and evaluated by drivers under realistic conditions. Therefore, before implementing such prototypes, more evaluations of, for example, the vocal and visual modalities are required with respect to robustness in the in-car environment and user acceptance.

Naturally, people talk, they talk different from today's command and control-oriented and in the near future oncoming rudimentary natural language-based in-car interaction, and engineers will have to listen [95]. At the same time, engines might soon observe our affective behaviour patterns—for our safety, comfort, and pleasure.

References

[1] PROSPER, "Prosper final report, project for research on speed adaptation policies on european roads," Tech. Rep. Project no. GRD2200030217, May 2006.

[2] K. Takeda, H. Erdogan, J. H. L. Hansen, and H. Abut, Eds., *In-Vehicle Corpus and Signal Processing for Driver Behavior*, Springer, Berlin, Germany, 2009.

[3] H. Bubb, "Fahrerassistenz primär ein Beitrag zum Komfort oder für die Sicherheit?" *VDI Berichte*, no. 1768, pp. 25–44, 2003.

[4] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, "A taxonomy of applications that utilize emotional awareness," in *Proceedings of the 1st International Language Technologies Conference (IS-LTC '06)*, pp. 246–250, Ljubljana, Slovenia, 2006.

[5] C. M. Jones and I. M. Jonsson, "Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses," in *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online (OZCHI '05)*, vol. 122, pp. 1–10, Canberra, Australia, 2005.

[6] G. H. Bower, "Mood and memory," *American Psychologist*, vol. 36, no. 2, pp. 129–148, 1981.

[7] J. A. Singer and P. Salovey, "Mood and memory: evaluating the network theory of affect," *Clinical Psychology Review*, vol. 8, no. 2, pp. 211–251, 1988.

[8] R. B. Zajonc, "On the primacy of affect," *American Psychologist*, vol. 39, no. 2, pp. 117–123, 1984.

[9] A. R. Damasio, *Descartes' Error: Emotion, Reason and the Human Brain*, Avon Books, New York, NY, USA, 1994.

[10] J. E. LeDoux, "Brain mechanisms of emotion and emotional learning," *Current Opinion in Neurobiology*, vol. 2, no. 2, pp. 191–197, 1992.

[11] D. Derryberry and D. M. Tucker, "Neural mechanisms of emotion," *Journal of Consulting and Clinical Psychology*, vol. 60, no. 3, pp. 329–338, 1992.

[12] J. A. Colquitt, J. A. LePine, and R. A. Noe, "Toward an integrative theory of training motivation: a meta-analytic path analysis of 20 years of research," *Journal of Applied Psychology*, vol. 85, no. 5, pp. 678–707, 2000.

[13] N. Frijda, *The Emotions*, Cambridge University Press, New York, NY, USA, 1986.

[14] R. Birdwhistle, *Kinesics and Context: Essays on Body Motion and Communication*, University of Pennsylvania Press, Philadelphia, Pa, USA, 1970.

[15] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*, Prentice Hall, Englewood Cliffs, NJ, USA, 1975.

[16] N. Chovil, "Discourse-oriented facial displays in conversation," *Research on Language and Social Interaction*, vol. 25, pp. 163–194, 1991.

[17] D. Goleman, *Emotional Intelligence*, Bantam Books, New York, NY, USA, 1995.

[18] J. A. Groeger, *Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task*, Frontiers of Cognitive Science, Routledge Chapman & Hall, 2000.

[19] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr, "On the necessity and feasibility of detecting a driver's emotional state while driving," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, vol. 4738 of *Lecture Notes in Computer Science*, pp. 126–138, Springer, Lisbon, Portugal, 2007.

[20] L. James and D. Nahl, *Road Rage and Agressive Driving*, Prometheus Books, Amherst, NY, USA, 2000.

[21] E. Wells-Parker, J. Ceminsky, V. Hallberg, R. W. Snow, G. Dunaway, S. Guiling, M. Williams, and B. Anderson, "An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers," *Accident Analysis and Prevention*, vol. 34, no. 3, pp. 271–278, 2002.

[22] H. Cai, Y. Lin, and R. R. Mourant, "Study on driver emotion in driver-vehicle-environment systems using multiple networked driving simulators," in *Proceedings of the Driving Simulation Conference*, May 2007.

[23] D. A. Hennessy and D. L. Wiesenthal, "The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours," *Ergonomics*, vol. 40, no. 3, pp. 348–361, 1997.

[24] J. L. Deffenbacher, E. R. Oetting, and R. S. Lynch, "Development of a driving anger scale," *Psychological Reports*, vol. 74, no. 1, pp. 83–91, 1994.

[25] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, pp. 459–482, 1908.

[26] J. M. Lyznicki, T. C. Doege, R. M. Davis, and M. A. Williams, "Sleepiness, driving, and motor vehicle crashes," *Journal of the American Medical Association*, vol. 279, no. 23, pp. 1908–1913, 1998.

[27] E. R. Braver, C. W. Preusser, D. F. Preusser, H. M. Baum, R. Beilock, and R. Ulmer, "Long hours and fatigue: a survey of tractor-trailer drivers," *Journal of Public Health Policy*, vol. 13, no. 3, pp. 341–366, 1992.

[28] M. T. Corfitsen, "Tiredness and visual reaction time among young male nighttime drivers: a roadside survey," *Accident Analysis and Prevention*, vol. 26, no. 5, pp. 617–624, 1994.

[29] C. L. Marcus and G. M. Loughlin, "Effect of sleep deprivation on driving safety in housestaff," *Sleep*, vol. 19, no. 10, pp. 763–766, 1996.

[30] A. T. McCartt, S. A. Ribner, A. I. Pack, and M. C. Hammer, "The scope and nature of the drowsy driving problem in New York state," *Accident Analysis and Prevention*, vol. 28, no. 4, pp. 511–517, 1996.

[31] G. Matthews, L. Dorn, T. W. Hoyes, D. R. Davies, A. I. Glendon, and R. G. Taylor, "Driver stress and performance on a driving simulator," *Human Factors*, vol. 40, no. 1, pp. 136–149, 1998.

[32] M. E. Rakauskas, L. J. Gugerty, and N. J. Ward, "Effects of naturalistic cell phone conversations on driving performance," *Journal of Safety Research*, vol. 35, no. 4, pp. 453–464, 2004.

[33] R. Banuls and L. Montoro, "Motivational and emotional aspects involved in driving," in *Traffic Psychology Today*, P.-E. Barjone, Ed., chapter 8, pp. 138–318, Springer, Berlin, Germany, 2001.

[34] K. Ball and G. Rebok, "Evaluating the driving ability of older adults," *Journal of Applied Gerontology*, vol. 13, no. 1, pp. 20–38, 1994.

[35] S. MacDonald, R. Mann, M. Chipman, et al., "Driving behaviour under the influence of cannabis and cocaine," *Traffic Injury Prevention*, vol. 9, no. 3, pp. 190–194, 2008.

[36] X. Li and Q. Ji, "Active affective state detection and user assistance with dynamic bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 35, no. 1, pp. 93–105, 2005.

[37] C. S. Dula and E. S. Geller, "Risky, aggressive, or emotional driving: addressing the need for consistent communication in

research," *Journal of Safety Research*, vol. 34, no. 5, pp. 559–566, 2003.

[38] G. Stecklov and J. R. Goldstein, "Terror attacks influence driving behavior in Israel," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 40, pp. 14551–14556, 2004.

[39] J. Healey and R. Picard, "Smart car: detecting driver stress," in *Proceedings of the 15th International Conference on Pattern Recognition (ICPR '00)*, vol. 4, pp. 218–221, 2000.

[40] M. Grimm, K. Kroschel, B. Schuller, G. Rigoll, and T. Moosmayr, "Acoustic emotion recognition in car environment using a 3d emotion space approach," in *Proceedings of the 33rd Deutsche Jahrestagung für Akustik (DAGA '07)*, pp. 313–314, DEGA, Stuttgart, Germany, March 2007.

[41] I. M. Jonsson, C. Nass, H. Harris, and L. Takayama, "Matching in-car voice with drivers state: impact on attitude and driving performance," in *Proceedings of the 3rd International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pp. 173–181, Rockport, Me, USA, June 2005.

[42] C. Nass, I.-M. Jonsson, H. Harris, et al., "Improving automotive safety by pairing driver emotion and car voice emotion," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '05)*, pp. 1973–1976, ACM, Portland, Ore, USA, 2005.

[43] Y. Uchiyama, S. Kojima, T. Hongo, R. Terashima, and T. Wakita, "Voice information system that adapts to driver's mental workload," *R&D Review of Toyota CRDL*, vol. 39, no. 1, pp. 16–22, 2004.

[44] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 2, pp. 1085–1088, Philadelphia, Pa, USA, March 2005.

[45] C. M. Jones and I. M. Jonsson, "Performance analysis of acoustic emotion recognition for in-car conversational interfaces," in *Universal Access in Human-Computer Interaction. Ambient Interaction, Part II*, vol. 4555 of *Lecture Notes in Computer Science*, pp. 411–420, Springer, Berlin, Germany, 2007.

[46] I. M. Jonsson, M. Zajicek, H. Harris, and C. Nass, "Thank you, i did not see that: in-car speech based information systems for older adults," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '05)*, pp. 1953–1956, Portland, Ore, USA, 2005.

[47] I. M. Jonsson, C. Nass, J. Endo, et al., "Don't blame me i am only the driver: impact of blame attribution on attitudes and attention to driving task," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '04)*, pp. 1219–1222, Vienna, Austria, 2004.

[48] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *Proceedings of the International Conference on Speech Prosody*, ISCA, Dresden, Germany, 2006.

[49] S. Furui., "Robust methods in automatic speech recognition and understanding," in *Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 1993–1998, ISCA, Geneva, Switzerland, September 2003.

[50] T. Athanaselis, S. Bakamidis, and I. Dologlou, "Automatic recognition of emotionally coloured speech," in *Proceedings of the Transactions on Engineering, Computing and Technology (ICCS '06)*, pp. 274–277, Vienna, Austria, March 2006.

[51] Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 424–428, 2007.

[52] R. Fernandez, *A computational model for the automatic recognition of affect in speech*, Ph.D. thesis, MIT Media Arts and Science, 2004.

[53] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 941–944, Honolulu, Hawaii, USA, 2007.

[54] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction—a data-driven approach towards analysing emotion in speech," *User Modelling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 175–206, 2008.

[55] S. D. Kollias and K. Karpouzis, "Multimodal emotion recognition and expressivity analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 779–783, Amsterdam, The Netherlands.

[56] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *Proceedings the 4th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI '07)*, vol. 247 of *IFIP International Federation for Information Processing*, pp. 375–388, 2007.

[57] J. Kim and E. André, "Emotion recognition using physiological and speech signal in short-term observation," in *Proceedings of the International Tutorial and Research Workshop on Perception and Interactive Technologies (PIT '06)*, vol. 4021 of *Lecture Notes in Computer Science*, pp. 53–64, Kloster Irsee, Germany, June 2006.

[58] F. Hönig, J. Wagner, A. Batliner, and E. Nöth, "Classification of user states with physiological signals: on-line generic features vs. specialized," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO '09)*, B. Stewart and S. Weiss, Eds., vol. 15, pp. 2357–2316, Glasgow, UK, 2009.

[59] E. André and J. Y. Chai, "Workshop: eye gaze in intelligent human machine interaction," in *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 431–432, Hong Kong, 2010.

[60] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[61] S. Steidl, B. Schuller, A. Batliner, and D. Seppi, "The hinterland of emotions: facing the open-microphone challenge," in *Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII '09)*, vol. 1, pp. 690–697, IEEE, Amsterdam, The Netherlands, 2009.

[62] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, pp. 205–211, State College, Pa, USA, October 2004.

[63] I. Cohen, G. Ashutosh, and T. S. Huang, "Emotion recognition from facial expressions using multilevel hmm," in *Neural Information Processing Systems*, 2000.

[64] B. Schuller, R. Müller, F. Eyben, et al., "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[65] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, "Toward emotion recognition in car-racing drivers: a biosignal

processing approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 38, no. 3, pp. 502–512, 2008.

[66] T. Paul-Stueve, Ed., *Driver Activity Recognition from Sitting Postures*, Bauhaus-Universität Weimar, Weimar, Germany, 2007.

[67] N. O. Bernsen and L. Dybkjær, "A multimodal virtual codriver's problems with the driver," in *Proceedings of the ISCA Tutorial and Research Workshop on Spoken Dialogue in Mobile Environments*, p. 15, ISCA, Bonn, Germany, 2002.

[68] K. Bachfischer, S. Henze, and C. Wäller, "Emotional user interfaces in the car," in *Proceedings of the 1st Workshop on Emotion and Computing—Current Research and Future Impact*, D. Reichardt, P. Levi, and J.-J. C. Meyer, Eds., pp. 55–59, Bremen, Germany, June 2006.

[69] T. Sasayuki, K. Shin, and Y. Shin, "A driver-adaptive driver assistance system. A case of human-centered its view aid system," *Nippon Kikai Gakkai Robotikusu, Mekatoronikusu Koenkai Koen Ronbunshu*, pp. 2A1–L2-4, 2004, (CD-ROM).

[70] M. A. Tischler, C. Perter, A. Batliner, M. Wimmer, and J. Voskamp, "Application of emotion recognition methods in automotive research," in *Proceedings of the 2nd Workshop on Emotion and Computing—Current Research and Future Impact*, vol. 1, pp. 55–60, Osnabrück, Germany, 2007.

[71] M. Ablassmeier, T. Poitschke, and G. Rigoll, "A new approach of a context-adaptive search agent for automotive environments," in *Extended Abstracts on Human Factors in Computing Systems (CHI '06)*, pp. 1613–1618, ACM Press, New York, NY, USA, April 2006.

[72] S. B. Wang, D. Demirdjian, T. Darrell, and H. Kjellström, "Multimodal communication error detection for driver-car interaction," in *Proceedings of the 4th International Conference on Informatics in Control, Automation and Robotics (ICINCO '07)*, pp. 365–371, Angers, France, May 2007.

[73] F. Althoff, G. McGlaun, P. Schuller, M. Lang, and G. Rigoll, "Evaluating misinterpretations during human-machine communication in automotive environments," in *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI '02)*, N. Callaos, J. H. Porter, and N. Rishe, Eds., vol. 7, pp. 13–17, Orlando, Fla, USA, 2002.

[74] A. Amditis, A. Polychronopoulos, E. Bekiaris, and P. C. Antonello, "System architecture of a driver's monitoring and hypovigilance warning system," in *Proceedings of the Intelligent Vehicle Symposium*, vol. 2, pp. 527–532, IEEE, Athens, Greece, March 2003.

[75] C. L. Lisetti and F. Nasoz, "Affective intelligent car interfaces with emotion recognition," in *Proceedings of the of the 11th International Conference on Human Computer Interaction (HCI '00)*, vol. 2, Las Vegas, Nev, USA, July 2005.

[76] V. Kostov and S. Fukuda, "Emotion in user interface, voice interaction system," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 798–803, Nashville, Tenn, USA, 2000.

[77] F. Angelucci, E. Ricci, L. Padua, A. Sabino, and P. A. Tonali, "Music exposure differentially alters the levels of brain-derived neurotrophic factor and nerve growth factor in the mouse hypothalamus," *Neuroscience Letters*, vol. 429, no. 2-3, pp. 152–155, 2007.

[78] N. Sheehy, "User-interface requirements for elderly people," Tech. Rep., School of Psychology, The Queen's University of Belfast, 2000.

[79] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, no. 5, pp. 1–19, 1990.

[80] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, no. 4, pp. 369–390, 1995.

[81] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant synthesis," in *Proceedings of the ICSA Workshop on Speech and Emotion*, pp. 151–156, ISCA, Belfast, Northern Ireland, 2000.

[82] S. J. L. Mozziconacci, *Speech variability and emotion: production and perception*, Ph.D. thesis, echnical University Eindhoven, 1998.

[83] F. Kuhn, "Methode zur Bewertung der Fahrerablenkung durch Fahrerinformationssysteme," April 2010, http://www.gui-design.de/download/wud_LCT_2005-11_Stuttgart.pdf.

[84] J. Brooke, "SUS—a quick and dirty usability scale," Tech. Rep., Redhatch Consulting, 1986.

[85] K. Eilers, F. Nachreiner, and K. Hänecke, "Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung," *Zeitschrift für Arbeitswissenscahft*, vol. 40, pp. 215–224, 1986.

[86] J. Wang, R. Knipling, and M. Goodman, "The role of driver inattention in crashes; new statistics from the 1995 crashworthiness data system (CDS)," in *Proceedings of the Annual Conference of the Association for the Advancement of Automotive Medicine*, Des Plaines, Ill, USA, 1996.

[87] T. Ranney, E. Mazzae, R. Garrott, and M. Goodman, "NHTSA driver distraction research: past, present and future," Tech. Rep., National Highway Traffic Safety Administration, Washington, DC, USA, 2000.

[88] T. P. Alkim, G. Bootsma, and S. P. Hoogendoorn, "Field operational test "the assisted driver"," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1198–1203, Instanbul, Turkey, 2007.

[89] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[90] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[91] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 1595–1598, Brighton, UK, September 2009.

[92] C. Blaschke, F. Breyer, B. Färber, J. Freyer, and R. Limbacher, "Driver distraction based lane-keeping assistance," *Transportation Research Part F*, vol. 12, no. 4, pp. 288–299, 2009.

[93] K. Young, M. Regan, and M. Hammer, "Driver distraction: a review of literature," Tech. Rep., Monash University Accident Research Center, 2003.

[94] A. Graves, *Supervised sequence labelling with recurrent neural network*, Ph.D. thesis, Technische Universität München, 2008.

[95] E. E. Shriberg, "Spontaneous speech: how people really talk and why engineers should care," in *Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH '05)*, pp. 1781–1784, ISCA, 2005.

Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 209801, 13 pages
doi:10.1155/2010/209801

Research Article

EmoHeart: Conveying Emotions in Second Life Based on Affect Sensing from Text

Alena Neviarouskaya,[1] Helmut Prendinger,[2] and Mitsuru Ishizuka[1]

[1] *Department of Information and Communication Engineering, University of Tokyo, R. 111C1/111D2, Engineering Building 2, 11th floor, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*
[2] *Digital Content and Media Sciences Research Division, National Institute of Informatics, 1613-1B, 16 floor, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*

Correspondence should be addressed to Alena Neviarouskaya, lena@mi.ci.i.u-tokyo.ac.jp

Received 1 April 2009; Accepted 28 September 2009

Academic Editor: Kostas Karpouzis

The 3D virtual world of "Second Life" imitates a form of real life by providing a space for rich interactions and social events. Second Life encourages people to establish or strengthen interpersonal relations, to share ideas, to gain new experiences, and to feel genuine emotions accompanying all adventures of virtual reality. Undoubtedly, emotions play a powerful role in communication. However, to trigger visual display of user's affective state in a virtual world, user has to manually assign appropriate facial expression or gesture to own avatar. Affect sensing from text, which enables automatic expression of emotions in the virtual environment, is a method to avoid manual control by the user and to enrich remote communications effortlessly. In this paper, we describe a lexical rule-based approach to recognition of emotions from text and an application of the developed Affect Analysis Model in Second Life. Based on the result of the Affect Analysis Model, the developed EmoHeart ("object" in Second Life) triggers animations of avatar facial expressions and visualizes emotion by heart-shaped textures.

1. Introduction and Motivation

> *Emotion is what gives communication life. A conversation between emotionally involved partners is bright and lively, but a meeting without feeling is deadly dull.*
>
> Sally Planalp [1]

Emotions play the role of a sensitive catalyst, which fosters lively interactions between human beings and assists in the development and regulation of interpersonal relationships. The expression of emotions shapes social interactions by providing observers a rich channel of information about the conversation partner [2] or his social intentions [3], by evoking positive or negative responses in others [4], and by stimulating other's social behaviour. Keltner et al. [5] highlight that "facial expressions are more than just markers of internal states," which also serve unique functions in a social environment. By accentuating the functional role of emotions, Frijda [6, 7] argues that they preserve and enhance life, and Lutz [8] emphasizes their communicative, moral, and cultural purposes.

The richness of emotional communication greatly benefits from the expressiveness of verbal (spoken words, prosody) and nonverbal (gaze, face, gestures, body pose) cues that enable auditory and visual channels of communication [1]. All types of expressive means potentially carry communicative power and promote better understanding [9]. Emotional information can be (1) encoded *lexically* within the actual words (affective predicates, intensifiers, modals, hedges, etc.) of the sentence, *syntactically* by means of subordinate clauses, and *morphologically* through changes in attitudinal shades of word meaning using suffixes (especially, in languages with rich inflectional system, such as Russian or Italian), (2) consciously or unconsciously conveyed through vast repertoire of prosodic features like intonation, stress, pitch, loudness, juncture, and rate of speech, (3) visually reflected through subtle details of contraction of facial muscles, tints of facial skin, eye movements, gestures,

and body postures [10]. The emotional significance of an utterance is accompanied, complemented, and modified by vocal and visual cues.

Nowadays, media for remote online communications and emerging 3D virtual worlds providing new opportunities for social contact grow rapidly, engage people, and gain great popularity among them. The main motivations for "residents" of chat rooms or virtual environments to connect to these media are seeking conversation, experimenting with a new communication media, and initiating relationships with other people. A study conducted by Peris et al. [11] revealed that "relationships developed online are healthy" and considered by people "as real as face-to-face relationships." Findings described in [12] indicate that there is a positive relationship between the amount of online media use and verbal, affective, and social intimacy, and that frequent online conversation actually encourages the desire to meet face-to-face, reinforcing thus personal interaction. To emphasize the realism and significance of social exchanges in such environments, Chayko [13] recently proposed to use the term "sociomental" rather than "virtual". Without doubt, computer-mediated communications that facilitate contact with others have strong potential to affect the nature of social life in terms of both interpersonal relationships and the character of community [14, 15].

To establish a social and friendly atmosphere, people should be able to express emotions. However, media for online communication lack the physical contact and visualization of emotional reactions of partners involved in a remote text-mediated conversation, limiting thus the source of information to text messages and to graphical representations of users (avatars) that are to some degree controlled by a person. Trends show that people often try to enrich their interaction online, introducing affective symbolic conventions or emphases into text (emoticons, capital letters, etc.) [12, 16–19], colouring emotional messages, or manually controlling the expressiveness of avatars in order to supplement the lack of paralinguistic cues. One of the first attempts to study effects of conveying emotional expressions through communication in computer-mediated environment was done by Rivera et al. [20]. The results of their experiment indicated that subjects allowed to use emoticons were more satisfied with the system than those subjects having conversations without these symbolic emotional expressions. The user study of ExMS [21], messaging system allowing its users to concatenate and annotate avatar animations, showed that interplay between pure text and animation significantly improved the expressiveness of messages, and that users felt pride of being identified with their embodied representation.

In this work we address the task of the enhancement of emotional communication in Second Life. This virtual world imitates a form of real life by providing a space for rich interactions and social events. To trigger visual display of a user's affective state in Second Life, the user has to manually assign appropriate facial expression or gesture to his or her avatar, which can distract the user from the communication process. In order to achieve truly natural communication in virtual worlds, we set a twofold focus in our research: (1) recognition of affective content conveyed through text, and (2) automatic visualization of emotional expression of avatars, which allows avoiding manual control by the user and enriching remote communications effortlessly.

The remainder of the paper is structured as follows. In Section 2, we report on related works. Section 3 summarizes the algorithm for recognition of fine-grained emotions from text, and Section 4 presents the results of the evaluation of our method. The application of the developed Affect Analysis Model in Second Life (EmoHeart) and analysis of the EmoHeart log data are described in Section 5 and Section 6, respectively. Finally, Section 7 concludes the paper.

2. Related Work

The emergence of the field of affective computing [22] has greatly inspired research challenging the issues of recognition, interpretation, and representation of affect. The emotional information expressed through a wide range of modalities has been considered, including affect in written language, speech, facial display, posture, and physiological activity. According to Picard [22], "the basic requirement for a computer to have the ability to express emotions is that the machine have channels of communication such as voice or image, and an ability to communicate affective information over those channels . . .".

Physiological biosignals (such as facial electromyograms, the electrocardiogram, the respiration effort, and the electrodermal activity) were analysed by Rigas et al. [23] to define emotional state of a human. Recent studies have begun to illuminate how emotions conveyed through vocal channel can be detected [24, 25]. Visual information also carries valuable emotional content. Considering the fact that eye and mouth expressions are most evident emotional indicators on the face, Maglogiannis et al. [26] developed a method that, based on color images, detects skin, eye, and mouth regions, and recognizes emotions encoded in these clues by detecting edges and evaluating the color gradient. Aimed at the synchronization of the avatar's state in virtual environment with the actual emotional state of the user, Di Fiore et al. [27] realized the automatic extraction of emotion-related metadata (particularly, facial features) from a real-time video stream originating from a webcam. Castellano et al. [28] have proposed a multimodal approach for emotion recognition from facial expressions, body movement, gestures, and speech. After training individual Bayesian classifiers for each modality, researchers fused the data at both feature and decision levels that resulted in the increase of accuracy compared to the unimodal approach.

The most challenging tasks for computational linguists are text classification as subjective or of factual nature, determination of orientation and strength of sentiment, and recognition of attitude type expressed in text at various grammatical levels. A variety of approaches have been proposed to determine the polarity of distinct terms [29, 30], lexical items in synsets [31], phrases/sentences [32], and documents [33, 34]. To analyse contextual sentiment, rule-based approaches [35, 36] and a machine-learning

method using not only lexical but also syntactic features [37] were proposed. Some researchers employed a keyword spotting technique to recognize emotion from text [38, 39]. Advanced methods targeting textual affect recognition at the sentence level are described in [40–43]. The attempts to automatically display emotions inferred from text in a chat, Instant Messenger (IM), or e-mail environment using still images of persons, rough simplified faces, and avatars are described in [38, 40, 41, 44]. The user study conducted on AffectIM [44], affective IM, showed that (1) the IM system with automatic emotion recognition from text was successful at conveying users' emotional states during communication online, thus enriching expressivity and social interactivity of online communications; (2) avatars were helpful in understanding the partner's emotions and giving some sense of physical presence.

The ideal method to accurately sense the emotional state of a person contacting others remotely would be to integrate approaches aiming at detection of affective state communicated through different expressive modalities and to obtain a decision based on the weights assigned to these expressive means. Our research is concerned with recognition of emotions reflected in linguistic utterances. In the paper we describe the application of the emotion recognition algorithm in the 3D virtual world Second Life.

3. Recognition of Fine-Grained Emotions from Text

In this section, we will summarize the main steps of emotion recognition using our Affect Analysis Model, which was introduced in [45].

3.1. Basis for Affective Text Classification. As the purpose of affect recognition in a remote communication system is to relate text to avatar emotional expressions, affect categories should be confined to those that can be visually expressed and easily understood by users. We analysed emotion categorizations proposed by theorists, and as the result of our investigation, for affective text classification, we decided to use the subset of emotional states defined by Izard [46]: "anger", "disgust", "fear", "guilt", "interest", "joy", "sadness", "shame", and "surprise". Izard's [46] theory postulates the existence of discrete fundamental emotions with their motivational, phenomenological properties, and personal meanings. Besides specific or qualitatively distinct affective states, we defined five communicative functions that are frequently observed in online conversations ("greeting", "thanks", "posing a question", "congratulation", and "farewell").

In order to support the handling of abbreviated language and the interpretation of affective features of lexical items, the Affect database was created. The Affect database includes the following tables: Emoticons, Abbreviations, Adjectives, Adverbs, Nouns, Verbs, Interjections, and Modifiers. The affective lexicon was mainly taken from WordNet-Affect [47]. Emotion categories with intensities were manually assigned to the emotion-related entries of the database by three independent annotators. Emotion intensity values range from 0.0 to 1.0. Emoticons and abbreviations were transcribed and related to named affective states (with intensity), whereby each entry was assigned to only one category (e.g., emoticon ":-S" [worry] was related to "fear" with intensity 0.4). Considering the fact that some affective words may express more than one emotional state, annotators could relate words to more than one category (e.g., the final annotation for noun "*enthusiasm*" is "interest:08, joy:0.5"). Two annotators gave coefficients for intensity degree strengthening or weakening (from 0.0 to 2.0) to the adverbs of degree, and the result was averaged (e.g., coeff("*significantly*") = 2.0).

3.2. Affect Analysis Model. While constructing our lexical rule-based approach to affect recognition from text, we took into account linguistic features of text written in a free informal style [48]. Our Affect Analysis Model was designed based on the compositionality principle, according to which we determine the emotional meaning of a sentence by composing the pieces that correspond to lexical units or other linguistic constituent types governed by the rules of aggregation, propagation, domination, neutralization, and intensification, at various grammatical levels. By analysing each sentence in sequential stages (symbolic cue processing, detection and transformation of abbreviations, sentence parsing, and word/phrase/sentence-level analyses), our method is capable of processing sentences of different complexities, including simple, compound, complex (with complement and relative clauses), and complex-compound sentences.

Symbolic Cue Analysis. In the first stage of the Affect Analysis Model, we test the sentence for emoticons, abbreviations, interjections, "?" and "!" marks, repeated punctuation, and capital letters. Several rules are applied to define the dominant emotion in cases when multiple emoticons and emotion-relevant abbreviations occur in a sentence. As interjections are added to sentences to convey emotion (e.g., "*Oh no*", "*wow*"), they are analysed as well. If there are no emotion-relevant emoticons or abbreviations in a sentence, we prepare the sentence for parser processing: emoticons and abbreviations relating to communicative function categories are excluded from the sentence; and nonemotional abbreviations are replaced by their proper transcriptions found in the database (e.g., "*I m* [am] *stressed bc* [because] *i have frequent headaches*"). In such a way, the issue of correct processing of abbreviated text by syntactical parser is resolved.

Syntactical Structure Analysis. The second stage is devoted to the analysis of syntactical structure of sentences, and it is divided into two main subtasks. First, sentence analysis based on the GNU GPL licensed Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml) [49] (which replaces the commercial parser used in our previous work [45]) returns word base forms (lemmas), parts of speech, and dependency functions representing relational

information between words in sentences. Second, parser output processing is performed. When handling the parser output, we represent the sentence as a set of primitive clauses (either independent or dependent). Each clause might include Subject formation (SF), Verb formation (VF), and Object formation (OF), each of which may consist of a main element (subject, verb, or object) and its attributives and complements. For the processing of complex or compound sentences, we build a so-called "relation matrix", which contains information about dependences that the verbs belonging to different clauses have.

Word-Level Analysis. In the third stage, for each word found in our database, the affective features of a word are represented as a vector of emotional state intensities e = [anger, disgust, fear, guilt, interest, joy, sadness, shame, surprise] (e.g., e("*remorsefully*") = $[0, 0, 0, 0.8, 0, 0, 0.5, 0, 0]$). In the case of a modifier, the system identifies its coefficient (e.g., coeff("*barely*") = 0.4). Our model also varies the intensities of emotional vectors of adjectives and adverbs in comparative or superlative forms (e.g., e("*glad*") = [0,0,0,0,0,0.4,0,0,0], e("*gladder*") = [0,0,0,0,0,0.48,0,0,0] and e("*gladdest*") = $[0, 0, 0, 0, 0, 0.56, 0, 0, 0]$).

Phrase-Level Analysis. The purpose of this stage is to detect emotions involved in phrases, and then in Subject, Verb, or Object formations. We have defined rules for processing phrases:

(1) adjective phrase: modify the vector of adjective (e.g., e("*extremely doleful*") = coeff("*extremely*") $*$ e("*doleful*") = 2.0 $*$ [0,0,0,0,0,0,0.4,0,0] = [0,0,0,0,0,0,0.8,0,0]),

(2) noun phrase: output vector with the maximum intensity within each corresponding emotional state in analysing vectors (e.g., e1 = [0..0.7..] and e2 = [0.3..0.5..] yield e3 = [0.3..0.7..]),

(3) verb plus adverbial phrase: output vector with the maximum intensity within each corresponding emotional state in analysing vectors (e.g., e("*shamefully deceive*") = [0,0.4,0,0,0,0,0.5,0.7,0] where e("*shamefully*") = [0,0,0,0,0,0,0,0.7,0] and e("*deceive*") = [0,0.4,0,0,0,0,0.5,0,0]),

(4) verb plus noun phrase: if verb and noun phrase have opposite valences (e.g., "*break favourite vase*", "*enjoy bad weather*"), consider vector of verb as dominant; if valences are the same (e.g., "*like honey*", "*hate crying*"), output vector with maximum intensity in corresponding emotional states,

(5) verb plus adjective phrase (e.g., "*is very kind*", "*feel bad*"): output vector of adjective phrase.

The rules for modifiers are as follows: (1) adverbs of degree multiply or decrease emotional intensity values; (2) negation modifiers such as "*no*", "*never*", "*nothing*" etc. cancel (set to zero) vectors of the related words, that is, "neutralize the emotional content" (e.g., "*Yesterday I went to a party, but nothing exciting happened there*"); (3) prepositions such as "*without*", "*except*", "*against*", and "*despite*" cancel vectors of related words (e.g., "*I climbed the mountain without fear*" is neutralized due to preposition). Statements with prefixed words like "*think*", "*believe*", "*sure*", "*know*", "*doubt*" etc., or with modal operators such as "*can*", "*may*", "*would*" etc. are neutralized by our system. Conditional clause phrases beginning with "*even though*", "*if*", "*unless*", "*whether*", "*when*" and so forth are neutralized as well (e.g., "*I eat when I'm angry, sad, bored...*").

Each of the Subject, Verb, or Object formations may contain words conveying emotional meaning. During this stage, we apply the described rules to phrases detected within formation boundaries. Finally, each formation can be represented as a unified vector encoding its emotional content.

Sentence-Level Analysis. The emotional vector of a simple sentence (or a clause) is generated from Subject, Verb, and Object formation vectors resulting from phrase-level analysis. The main idea here is to first derive the emotion vector of Verb-Object formation relation. It is estimated based on the "verb plus noun phrase" rule described above. In order to apply this rule, we automatically determine valences of Verb and Object formations using their unified emotion vectors (particularly, nonzero-intensity emotion categories). The estimation of the emotion vector of a clause (Subject plus Verb-Object formations) is then performed in the following manner: (1) if valences of Subject formation and Verb formation are opposite (e.g., SF = "*my darling*", VF = "*smashed*", OF "*his guitar*"; or SF = "*troubled period*", VF = "*luckily comes to an end*"), we consider the vector of the Verb-Object formation relation as dominant; (2) otherwise, we output the vector with maximum intensities in corresponding emotional states of vectors of Subject and Verb-Object formations.

To estimate the emotional vector of a compound sentence, first, we evaluate the emotional vectors of its independent clauses. Then, we define the resulting vector of the compound sentence based on two rules: (1) with comma and coordinate connectors "and" and "so" (e.g., "*It is my fault, and I am worrying about consequences*", "*The exotic birds in the park were amazing, so we took nice pictures*"), or with a semicolon with no conjunction: output the vector with the maximum intensity within each corresponding emotional state in the resulting vectors of both clauses; (2) with coordinate connector "but" (e.g., "*They attacked, but we luckily got away!*"): the resulting vector of a clause following after the connector is dominant.

In order to process a complex sentence with a complement clause (e.g., "*I hope that Sam will not yell at my dog*"), first we derive the emotional vector of the complement clause, then create Object formation for the main clause using this vector, and finally estimate the resulting emotional vector of the main clause with added Object formation. In brief, we represent such sentence as a simple one, using the following pattern: "who-subject does-verb what-object", where object is represented as a complement clause. In our algorithm, the complex sentences containing adjective (relative) clauses are analysed in the following manner:

Table 1: The distributions of emotion labels across "gold standard" sentences.

Gold standard	Number/percent of sentences with distinct emotion labels										
	neutral	anger	disgust	fear	guilt	interest	joy	sadness	shame	surprise	total
2-3 annotators agreed	75/11.4	59/9.0	30/4.6	49/7.5	22/3.4	43/6.6	181/27.6	145/22.1	9/1.4	43/6.6	656/100
3 annotators agreed	8/3.2	17/6.8	9/3.6	24/9.6	12/4.8	8/3.2	88/35.3	58/23.3	3/1.2	22/8.8	249/100

Table 2: The results of experiment with Affect Analysis Model employing Stanford Parser.

Gold standard	Measure	Fine-grained emotion categories										Merged labels		
		neut	ang	disg	fear	guilt	inter	joy	sad	sh	sur	Pos	Neg	Neut
2-3 annotators agreed	Averaged accuracy					**0.649**							**0.747**	
	Precision	0.30	0.77	0.64	0.74	0.71	0.61	0.83	0.74	0.50	0.76	0.84	0.91	0.28
	Recall	0.55	0.34	0.70	0.80	0.55	0.81	0.71	0.64	0.67	0.72	0.80	0.75	0.55
	F-score	0.39	0.47	0.67	0.76	0.62	0.70	0.76	0.69	0.57	0.74	0.82	0.82	0.37
3 annotators agreed	Averaged accuracy					**0.751**							**0.814**	
	Precision	0.15	0.92	0.83	0.87	0.80	0.50	0.96	0.88	0.50	0.82	0.94	0.98	0.08
	Recall	0.75	0.65	0.56	0.83	0.67	0.75	0.78	0.74	0.33	0.82	0.85	0.79	0.75
	F-score	0.24	0.76	0.67	0.85	0.73	0.60	0.86	0.80	0.40	0.82	0.89	0.88	0.14

(1) the emotional vector of adjective clause is estimated; (2) this emotional vector is added to the SF or OF of the main clause depending on the role of the word to which the adjective clause relates; (3) the emotional vector of the whole sentence is estimated.

While processing complex-compound sentences (e.g., "*Max broke the china cup, with which Mary was awarded for the best song, so he regretted profoundly*"), first we generate emotional vectors of dependent clauses, and then of complex sentences, and finally, we analyse the compound sentence formed by the independent clauses. It is important to note that our system enables the differentiation of the strength of the resulting emotion depending on the tense of a sentence and availability of first person pronouns. The dominant emotion of the sentence is determined according to the emotion state with the highest intensity within the final emotional vector.

4. Evaluation of the Affect Analysis Model

In order to evaluate the performance of the Affect Analysis Model and to compare out method with related work, we conducted a set of experiments on data sets extracted from blogs.

4.1. Experiments with Our Collection of Sentences Extracted from Blogs. To measure the accuracy of the proposed emotion recognition algorithm with the freely available Stanford parser [49] (rather than the proprietary Connexor parser (Connexor Machinese Syntax: http://www.connexor.eu/technology/machinese/machinesesyntax/ used in our previous work [45]), we extracted 700 sentences from a collection of diary-like blog posts provided by BuzzMetrics (Weblog Data Collection. BuzzMetrics, Inc. http://www.nielsenbuzzmetrics.com/). Particularly, we focused on online diary or personal blog entries, which are typically written in a free style and are rich in emotional colourations. The most noticeable aspects of diary-like text are privacy, naturalism, and honesty in the expression of the author's thoughts and feelings.

Three independent annotators labelled the sentences with one of nine emotion categories (or neutral) and a corresponding intensity value. For the evaluation of algorithm performance, we created two collections of sentences corresponding to different "gold standards": (1) 656 sentences, on which two or three human raters completely agreed (Fleiss' Kappa coefficient is 0.51), and (2) 249 sentences, on which all three human raters completely agreed (Fleiss' Kappa coefficient is 1.0). Table 1 shows the distributions of emotion labels across "gold standard" sentences.

The performance of the Affect Analysis Model (AAM) employing Stanford Parser was evaluated against both sets of sentences related to "gold standards." Averaged accuracy, precision, recall, and F-score are shown in Table 2 for each fine-grained emotion category. Additionally, we provide the results for merged labels (positive emotions including "interest", "joy", and "surprise"; negative emotions including "anger", "disgust", "fear", "guilt", "sadness", and "shame"; and neutral).

We also evaluated the system performance with regard to estimation of emotion intensity. The percentage of emotional sentences (not considering neutral ones), on which the result of our system conformed to the "gold standards", according to the measured distance between intensities given by human raters (averaged values) and those obtained by the Affect Analysis Model is shown in Table 3. As seen from the table, our system achieved satisfactory results for emotion intensity estimation. The samples of sentences along with their annotations from "gold standard" and from the Affect Analysis Model are listed in Table 4.

The analysis of the failures of Affect Analysis Model revealed that common sense or additional context is required for processing some sentences. For example, human annotators agreed on the "sadness" emotion conveyed

Table 3: Percentage of emotional sentences according to the range of intensity difference between human annotations and output of algorithm.

Gold standard	Percentage of sentences according to the range of intensity difference (%)				
	[0.0–0.2]	(0.2–0.4]	(0.4–0.6]	(0.6–0.8]	(0.8–1.0]
2-3 annotators agreed	48.8	30.6	16.6	3.9	0.0
3 annotators agreed	51.4	27.6	17.1	3.9	0.0

Table 4: Examples of sentences and their annotations.

Sentence	Annotations	
	annotator 1/annotator 2/annotator 3	result of AAM
Tomorrow I am going to pay Glenn his money, and then I am going to fire him.	anger:0.6/anger:1.0/neutral:0.0	anger:0.51
I dislike people who talk to cats like annoying ppl [people] talk to babies.	disgust:0.6/disgust:0.7/neutral:0.0	disgust:0.32
The concept of driving in Thailand is one that filled me with dread, as anyone who has been must surely testify.	fear:0.8/fear:0.5/fear:0.9	fear:0.32
The fleas were entirely my fault, as I brought three cats to the house.	guilt:0.7/guilt:0.9/guilt:1.0	guilt:0.77
I have the desire to turn my head and do something creative other than look at pretty pictures!	interest:0.7/anger:1.0/interest:0.8	interest:0.96
I learn to take time to relax and enjoy life, even if things are stressful.	joy:0.4/neutral:0.0/joy:0.8	joy:0.48
Graceful tools of hope were destroyed by the lack of proper direction.	sadness:0.2/sadness:0.2/neutral:0.0	sadness:0.32
She was there for me and I was so ashamed because I ate fast food.	guilt:0.7/shame:0.7/shame:1.0	shame:0.38
And that house was amazingly huge.	surprise:0.8/surprise:1.0/surprise:1.0	surprise:0.4
i hardly become angry unless provoked.	neutral:0.0/neutral:0.0/anger:0.5	neutral:0.0

Table 5: Comparison of accuracy of Affect Analysis Model employing different parsers (Connexor Machinese Syntax versus Stanford Parser).

Measure	Gold standard			
	2-3 annotators agreed		3 annotators agreed	
	Fine-grained emotions	Merged labels	Fine-grained emotions	Merged labels
	656 sentences, Kappa = 0.51	692 sentences, Kappa = 0.60	249 sentences, Kappa = 1.0	447 sentences, Kappa = 1.0
Accuracy of AAM with Connexor Machinese Syntax	0.726	0.816	0.815	0.890
Accuracy of AAM with Stanford Parser	0.649	0.747	0.751	0.814
Difference in %	7.7	6.9	6.4	7.6

Table 6: Distribution of labels across sentences from benchmark used in the experiment.

Labels	Number of sentences
joy	536
sadness	173
anger	179
disgust	172
surprise	115
fear	115
neutral	600

through "*What I hope is that he can understand how much I treasure this friendship*", while our system resulted in erroneous "joy" emotion. In some cases, where system result did not agree with the "gold standard" due to the rule of neutralization of negated phrases (e.g., sentence "*I don't care whether they like me at the cocktail parties, or not*" was annotated by humans as expressing "anger" emotion, and by our system as "neutral"), the solution would be to reverse the valence of a statement (e.g., positive "*care*" with negation should become negative phrase "*don't care*"); however, finding the pairs of opposite emotions might be problematic. Neutralizations due to "cognition-related" ("*assume*", "*know*", "*think*"), modal ("*can*", "*could*", "*may*", "*would*"), and condition ("*if*") words also caused the problematic interpretations (e.g., AAM resulted in "neutral" emotion in sentences "*I tried explaining to him my outlooks on life last night, and I think that I upset him*", "*She knows that she can trust me, I'll never do her wrong*", and "A*nd if he would*

Table 7: Results of AAM compared to machine learning methods proposed by Aman and Szpakowicz [50].

Algorithm	Measure	joy	sadness	anger	disgust	surprise	fear	neutral
AAM	Averaged accuracy				0.770			
	Precision	**0.846**	**0.673**	**0.910**	**0.946**	0.758	0.785	**0.698**
	Recall	**0.858**	**0.763**	**0.564**	**0.506**	**0.652**	**0.730**	**0.862**
	F-score	**0.852**	**0.715**	**0.697**	**0.659**	**0.701**	**0.757**	**0.771**
ML with unigrams	Precision	0.840	0.619	0.634	0.772	**0.813**	**0.889**	0.581
	Recall	0.675	0.301	0.358	0.453	0.339	0.487	0.342
	F-score	0.740	0.405	0.457	0.571	0.479	0.629	0.431
ML with unigrams, RT features, and WNA features	Precision	0.813	0.605	0.650	0.672	0.723	0.868	0.587
	Recall	0.698	0.416	0.436	0.488	0.409	0.513	0.625
	F-score	0.751	0.493	0.522	0.566	0.522	0.645	0.605

laugh when it happens that would only make me more angry and thus blow up at him", while "gold standard" annotations were "sadness", "joy", and "anger", correspondingly). Such results generate a need for more careful analysis of the cases where condition or modal operators are involved in the sentence. Other errors were caused by the lack of relevant terms in Affect database (e.g., emotion in a sentence "*He's just lying*" was not recognized by our system as word "*lie*" was not included in the lexicon), incorrect results from syntactical parser, and sense ambiguity.

It is worth noting, however, that the accuracy of the Affect Analysis Model with the (commercially available) parser (Connexor Machinese Syntax) used in our previous work was higher in 6%–8% on the same sets of sentences (see details of comparison in Table 5). This indicates that Stanford Parser employed for the syntactical structure analysis is less efficient. On the other hand, as we aim to freely distribute and apply our emotion recognition tool to textual messages in a virtual world Second Life, we have to compromise on the performance of the system for the sake of free distribution.

4.2. Experiment with the Emotion Blog Data Set Developed by Aman and Szpakowicz [51]. This emotion blog data set was developed and kindly provided by Aman and Szpakowicz [51]. It includes sentences collected from blogs, which are characterized by rich emotional content and good examples of real-world instances of emotions conveyed through text. To directly compare the Affect Analysis Model with the machine learning methods proposed by Aman and Szpakowicz [50], we considered their benchmark as the "gold standard." Their blog data include sentences annotated by one of six emotions ("happiness", "sadness", "anger", "disgust", "surprise", and "fear"), or neutral, on which two annotators completely agreed. In the description of this experiment we further use label "joy" instead of "happiness". The distribution of labels across sentences from the benchmark used in the experiment is shown in Table 6.

AAM is capable of recognizing nine emotions, whereas the methods described in [50] classify text to six emotions. In order to compare the results of our approaches we decided to reduce the number of our labels by mapping "interest" to "joy", and "guilt" and "shame" to "sadness". The results of experiments are shown in Table 7, where AAM is compared to two machine learning methods: (1) "ML with unigrams", which employs corpus-based features, namely, all unigrams that occur more than three times in the corpus, excluding stopwords; (2) "ML with unigrams, RT features, and WNA features", which combines corpus-based features with features based on the following emotion lexicons: Roget's Thesaurus (RT) [52] and WordNet-Affect (WNA) [47].

The obtained results (precision, recall, and F-score) revealed that our rule-based system outperformed both machine learning methods in automatic recognition of "joy", "sadness", "anger", "disgust", and "neutral". In case of "surprise" and "fear" emotions, "ML with unigrams" resulted in higher precision, but lower recall and F-score than our AAM.

5. EmoHeart

Emotional expression is natural and very important for communication in real life but currently rather cumbersome in the 3D virtual world Second Life, where expressions have to be selected and activated manually. Concretely, a user has to click on the animation gesture in the list or type the predefined command following the symbol "/" in a textual chat entry. In order to breathe emotional life into graphical representations of users (avatars) through the automation of emotional expressiveness, we applied the developed Affect Analysis Model to textual chat in Second Life. The architecture of the system is presented in Figure 1.

The control of the conversation is implemented through the Second Life object called EmoHeart (http://www.prendingerlab.net/globallab/?page_id=22) which is attached to the avatar's chest and is invisible in the case of "neutral" state. The distributor of the EmoHeart object is located inside a (fictitious) Starbucks cafe (Second Life landmark: http://slurl.com/secondlife/NIIsland/213/38/25/) of the Second Life replica of National Center of Sciences building in Tokyo, which also hosts the National Institute of Informatics (NII). Once attached to the avatar, EmoHeart object (1)

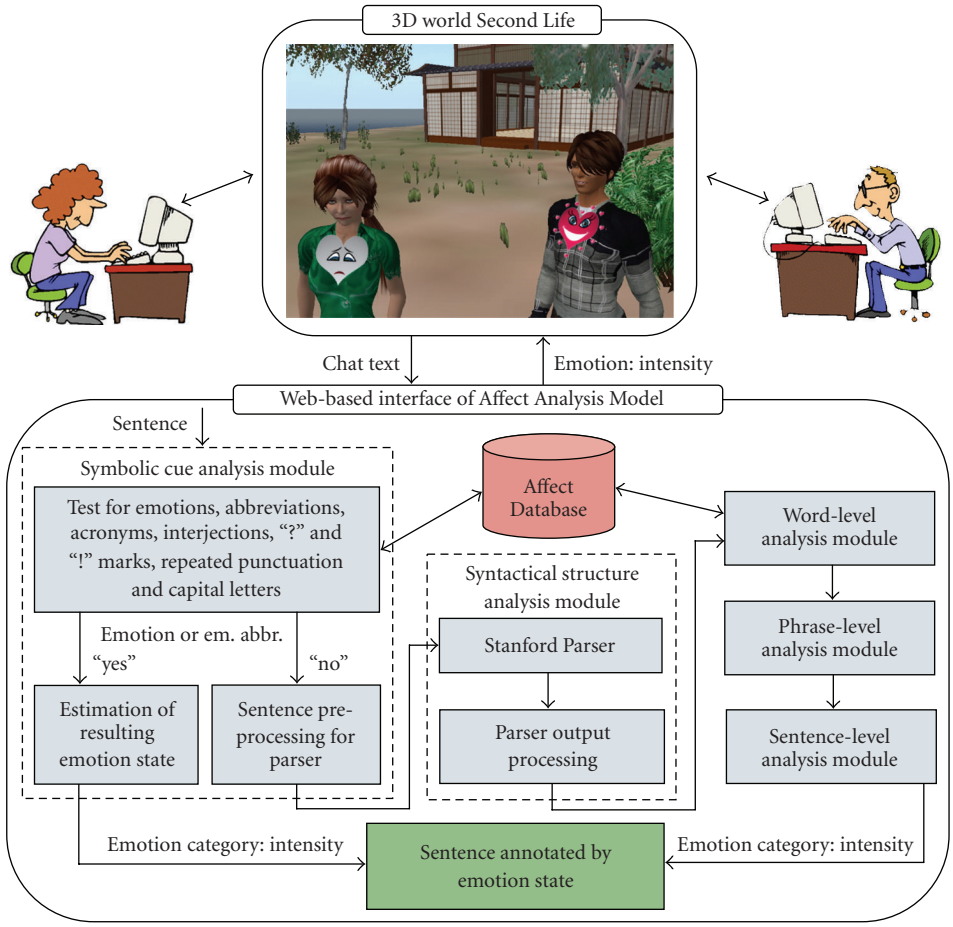


Figure 1: Architecture of the EmoHeart system.

Table 8: Emotional states and relevant expressive means (data partially taken from [53]).

Emotion	Expressive means
Anger	widely open eyes, fixated; pupils contracted; stare gaze; ajar mouth; teeth usually clenched tightly; rigidity of lips and jaw; lips may be tightly compressed, or may be drawn back to expose teeth
Disgust	narrowed eyes, may be partially closed as result of nose being drawn upward; upper lip drawn up; pressed lips; wrinkled nose; turn of the head to the side quasi avoiding something
Fear	widely open eyes; pupils dilated; raised eyebrows; open mouth with crooked lips; trembling chin
Guilt	downcast or glancing gaze; inner corners of eyebrows may be drawn down; lips drawn in, corners depressed; head lowered
Interest	eyes may be exaggeratedly opened and fixed; lower eyelids may be raised as though to sharpen visual focus; increased pupil size; sparkling gaze; mouth slightly smiling; head is slightly inclined to the side
Joy	"smiling" and bright eyes; genuinely smiling mouth
Sadness	eyelids contracted; partially closed eyes; downturning mouth
Shame	downcast gaze; blushing cheeks; head is lowered
Surprise	widely open eyes; slightly raised upper eyelids and eyebrows; the mouth is opened by the jaw drop; the lips are relaxed

listens to each message of its owner, (2) sends it to the web-based interface of the Affect Analysis Model located on the server, (3) receives the result (dominant emotion and intensity), and visually reflects the sensed affective state through the animation of avatar's facial expression, EmoHeart texture (indicating the type of emotion), and size of the texture (indicating the strength of emotion, namely, "low", "middle", or "high"). If no emotion is detected in the text, the EmoHeart remains invisible and the facial expression remains neutral.

Table 9: Statistics on EmoHeart log of 74 users for period December 2008 – January 2009.

Measure	Messages, number	Message length, symbols	Sentences, number
Total	19591 (for all users)	400420 (for all messages)	21396 (for all messages)
Minimal	1 (for user)	1 (for message)	1 (for message)
Maximal	2932 (for user)	634 (for message)	25 (for message)
Average	265 (per user)	20 (per message)	1.09 (per message)

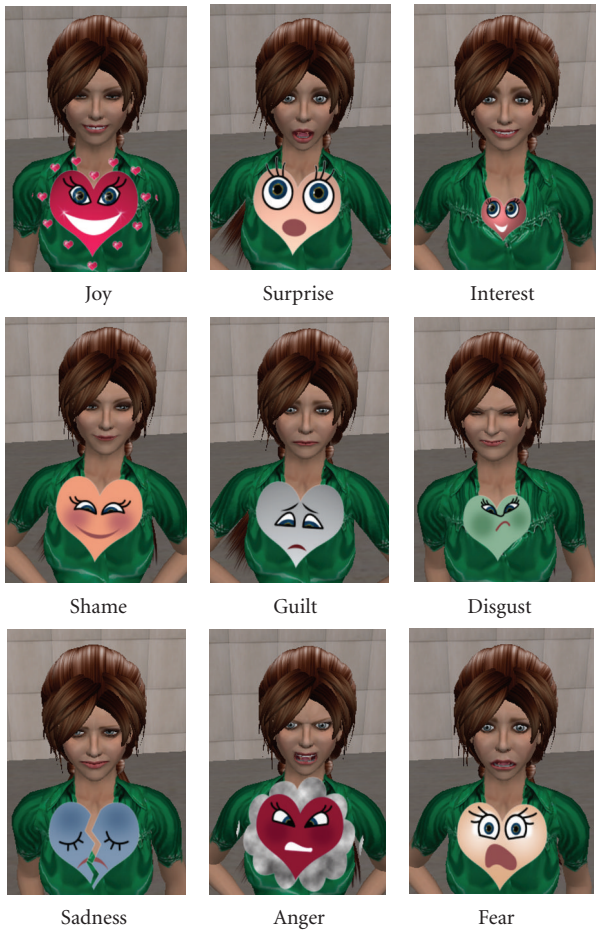


Figure 2: Examples of avatar facial expressions and EmoHeart textures.

Of the bodily organs, the heart plays a particularly important role in our emotional experience. People often characterize personal traits, emotional experiences, or mental states using expressions originating from word "*heart*" (i.e., "*heartfelt*", "*warm-hearted*", "*heartlessness*", "*kind-heartedness*", "*broken-hearted*", "*heart-burning*", "*heart-to-heart*", etc.). The essence of emotional, moral, and spiritual aspects of a human being has long been depicted using heart-shaped symbol. With the heart-shaped object of EmoHeart, we provide an additional channel for visualizing emotions in a vivid and expressive way. The examples of avatar facial expressions and EmoHeart textures are shown in Figure 2.

While designing EmoHeart textures, we followed the description of main characteristic features of expressive means in relation to communicated emotion (Table 8).

6. Analysis of EmoHeart Log

We made EmoHeart available for Second Life users from December 2008. During a two-month period (December 2008 – January 2009), we asked students to promote the EmoHeart object by visiting locations in Second Life and engaging other Second Life residents in social communication. As a result, 89 Second Life users became owners of EmoHeart, and 74 of them actually communicated using it. Text messages along with the results from Affect Analysis Model were stored in an EmoHeart log database. Some general statistics is given in Table 9. As seen from the table, the chat activity of users within two months (from 1 message to 2932 messages per user), as well as the length of a chat message in symbols (from 1 symbol to 634 symbols per message), varied significantly. In average, typical chat message included one sentence.

From all sentences, 20% were categorized as emotional by the Affect Analysis Model and 80% as neutral (Figure 3). We observed that the percentage of sentences annotated by positive emotions ("joy", "interest", "surprise") essentially prevailed (84.6%) over sentences annotated by negative emotions ("anger", "disgust", "fear", "guilt", "sadness", "shame"). We believe that this dominance of positivity expressed through text is due to the nature and purpose of online communication media, which allows people to exchange experiences, share opinions and feelings, and satisfy their social need of interpersonal communication. Harker and Keltner [54] empirically verified that the tendency to express positive emotions creates more harmonious social relationships, which in turn fosters personal growth and well-being.

We analysed the distribution of emotional sentences from EmoHeart log data according to the fine-grained emotion labels from our Affect Analysis Model (Figure 4). We found that the most frequent emotion conveyed through text messages is "joy" (68.8% of all emotional sentences), followed by "surprise", "sadness", and "interest" (9.0%, 8.8%, and 6.9%, resp.). All remaining emotions individually do not exceed the level of 2.1%. The least frequent emotion detected from text messages is "shame" (0.6% of all emotional sentences).

As the Affect Analysis Model also enables detection of five communicative functions (besides nine distinct affective states) that are frequently observed in online conversations,

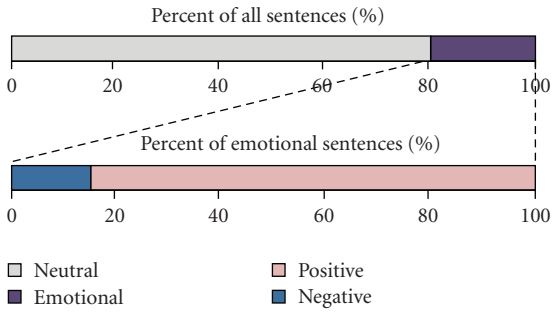


Figure 3: Percentage distribution of emotional (positive or negative) and neutral sentences.

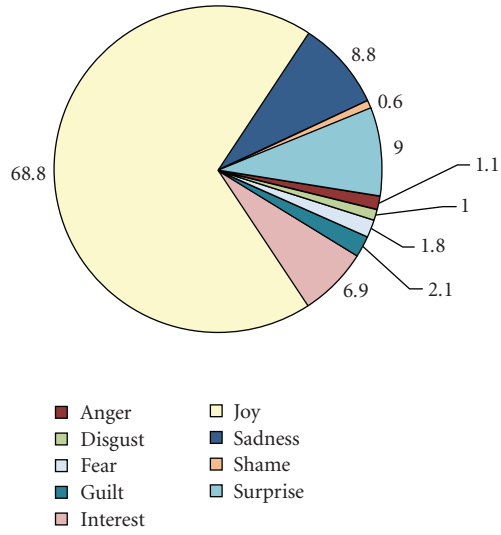


Figure 4: Percentage distribution of sentences with fine-grained emotion annotations.

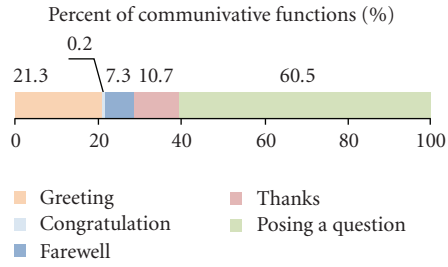


Figure 5: Percentage distribution of five communicative functions.

we analysed the communicative functions identified in the EmoHeart log data as well. The percentage distribution of detected communicative functions is shown in Figure 5. Our observations suggest that during online chatting people often ask each other questions (60.5% of the cases of detected communicative functions), requesting thus for new information, confirmation, or denial. Such social behaviours as "greeting" and "farewell", that are constituent parts of face-to-face communication, were recognized in 21.3% and 7.3% of the cases, respectively. EmoHeart users expressed gratitude in 10.7% and congratulate each other in 0.2% of the cases of detected communicative functions.

7. Conclusion

This paper introduced the integration of the developed emotion recognition module, Affect Analysis Model, into the 3D virtual world Second Life. The proposed lexical rule-based algorithm to affect sensing from text enables analysis of nine emotions at various grammatical levels. For textual input processing, our Affect Analysis Model handles not only correctly written text but also informal messages written in an abbreviated or expressive manner. The salient features of the Affect Analysis Model are the following:

(1) analysis of nine emotions on the level of individual sentences: this is an extensive set of labels if compared to six emotions mainly used in related work,

(2) the ability to handle the evolving language of online communications: to the best of our knowledge, our approach is the first attempt to deal with informal and abbreviated style of writing, often accompanied by the use of emoticons,

(3) foundation in database of affective words (each term in our Affect database was assigned at least one emotion label along with emotion intensity, in contrast to annotations of one emotion label or polarity orientation in competing approaches), interjections, emoticons, abbreviations and acronyms, modifiers (which influence on degrees of emotion states),

(4) vector representation of affective features of words, phrases, clauses, and sentences,

(5) consideration of syntactic relations and semantic dependences between words in a sentence: our rule-based method accurately classifies context-dependent affect expressed in sentences containing emotion-conveying terms, which may play different syntactic and semantic roles,

(6) analysis of negation, modality, and conditionality: most researchers ignore modal expressions and condition prepositions, therefore, their systems show poor performance in classifying neutral sentences, which is, indeed, not easy task,

(7) consideration of relations between clauses in compound, complex, or complex-compound sentences: to our knowledge, AAM is the first system comprehensively processing affect reflected in sentences of different complexity,

(8) emotion intensity estimation: in our work, the strength of emotion is encoded through numerical value in the interval [0.0, 1.0], in contrast to low/middle/high levels detected in some of competing methods.

Our system showed promising results in fine-grained emotion recognition in real examples of online conversation (diary-like blog posts): (1) on data set created by us, averaged

accuracy was 72.6% on sentences where two or three human annotators agreed, and 81.5% on sentences where all three human annotators agreed (nine emotion categories, and neutral); (2) on data set provided by Aman and Szpakowicz [50], averaged accuracy was 77.0% (six emotion categories, and neutral), and our system outperformed the method reported in related work in terms of precision, recall, and F-scores. Currently, the main limitations of the developed affect recognition module are (1) strong dependency on the lexicon resource, Affect database, (2) no disambiguation of word meanings, (3) disregard of contextual information and conversation history, and (4) inability to recognize and process misspelled words in a sentence. In our future study we will investigate those issues and explore the possibilities to overcome the current limitations of the system. As our system is completely lexical and the language of online conversations is "evolving", we are planning to realize a procedure for the automatic updating of the Affect database. With respect to the rules for composition of emotion vectors of terms comprising phrases or clauses, we believe that the approach aiming at learning rules from corpora would be useful.

In Second Life, the Affect Analysis Model serves as the engine behind automatic visualization of emotions conveyed through textual messages. The control of the conversation in Second Life is implemented through the EmoHeart object attached to the avatar's chest. This object communicates with Affect Analysis Model located on the server and visually reflects the sensed affective state through the animation of avatar's facial expression, EmoHeart texture, and size of the texture. In the future, we aim to study cultural differences in perceiving and expressing emotions and to integrate a text-to-speech engine with emotional intonations into textual chat of Second Life.

Acknowledgments

The authors would like to acknowledge and thank Alessandro Valitutti and Dr. Diego Reforgiato for their kind help during the Affect database creation. They wish also to express their gratitude to Dr. Dzmitry Tsetserukou, Dr. Shaikh Mostafa Al Masum, Manuel M. Martinez, Zoya Verzhbitskaya, Hutchatai Chanlekha, and Nararat Ruangchaijatupon who have contributed to annotations of Affect database entries and sentences, for their efforts and time. Special thanks also go to Cui Xiaoke, Tananun Orawiwattanakul, and Farzana Yasmeen for their work on EmoHeart promotion in Second Life. This research was partly supported by a JSPS Encouragement of Young Scientists Grant (FY2005-FY2007), an NII Joint Research Grant with the University of Tokyo (FY2007), and an NII Grand Challenge Grant (FY2008-FY2009).

References

[1] S. Planalp, *Communicating Emotion: Social, Moral, and Cultural Processes*, Cambridge University Press, Cambridge, UK, 1999.

[2] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[3] A. J. Fridlund, "The behavioral ecology and sociality of human faces," *Review of Personality and Social Psychology*, vol. 13, pp. 90–121, 1992.

[4] U. Dimberg and A. Ohman, "Behold the wrath: psychophysiological responses to facial stimuli," *Motivation and Emotion*, vol. 20, no. 2, pp. 149–182, 1996.

[5] D. Keltner, P. Ekman, G. C. Gonzaga, and J. Beer, "Facial expression of emotion," in *Handbook of Affective Science*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds., pp. 415–432, Oxford University Press, New York, NY, USA, 2003.

[6] N. Frijda, *The Emotions*, Studies in Emotion and Social Interaction, Cambridge University Press, Cambridge, UK, 1986.

[7] N. Frijda, "Emotions are functional, most of the time," in *The Nature of Emotion: Fundamental Questions*, P. Ekman and R. J. Davidson, Eds., Oxford University Press, New York, NY, USA, 1994.

[8] C. Lutz, *Unnatural Emotions*, University of Chicago Press, Chicago, Ill, USA, 1988.

[9] J. Allwood, "Bodily communication dimensions of expression and content," in *Multimodality in Language and Speech Systems*, pp. 7–26, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[10] J. Reilly and L. Seibert, "Language and emotion," in *Handbook of Affective Science*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds., pp. 535–559, Oxford University Press, New York, NY, USA, 2003.

[11] R. Peris, M. A. Gimeno, D. Pinazo, et al., "Online chat rooms: virtual spaces of interaction for socially oriented people," *Cyberpsychology and Behavior*, vol. 5, no. 1, pp. 43–51, 2002.

[12] Y. Hu, J. F. Wood, V. Smith, and N. Westbrook, "Friendships through IM: examining the relationship between instant messaging and intimacy," *Journal of Computer-Mediated Communication*, vol. 10, no. 1, article 6, 2004.

[13] M. Chayko, *Connecting: How We Form Social Bonds and Communities in the Internet Age*, State University of New York Press, Albany, NY, USA, 2002.

[14] J. Fernback and B. Thompson, "Virtual Communities: Abort, Retry, Failure?" 2004, http://www.well.com/user/hlr/texts/VCcivil.html.

[15] H. Rheingold, *The Virtual Community*, MIT Press, Cambridge, Mass, USA, 2002.

[16] E. M. Reid, *Electropolis: communication and community on Internet relay chat*, Honours thesis, University of Melbourne, Melbourne, Australia, 1991.

[17] D. Derks, *Exploring the missing wink: emoticons in cyberspace*, Ph.D. dissertation, Open University of Netherlands, Leiderdorp, The Netherlands, 2007.

[18] J. B. Walther and K. P. D'Addario, "The impacts of emoticons on message interpretation in computer-mediated communication," *Social Science Computer Review*, vol. 19, no. 3, pp. 324–347, 2001.

[19] O.T. Yigit, *Emoticon usage in task-oriented and socio-emotional contexts in online discussion board*, M.S. thesis, The Florida State University, Tallahassee, Fla, USA, 2005.

[20] K. Rivera, N. J. Cooke, and J. A. Bauhs, "The effects of emotional icons on remote communication," in *Proceedings of the Conference Companion on Human Factors in Computing Systems (CHI '96)*, pp. 99–100, Vancouver, Canada, April 1996.

[21] P. Persson, "ExMS: an animated and avatar-based messaging system for expressive peer communication," in *Proceedings of the International ACM SIGGROUP Conference on Supporting*

Group Work, pp. 31–39, Sanibel Island, Fla, USA, November 2003.

[22] R. Picard, *Affective Computing*, MIT Press, Cambridge, Mass, USA, 1997.

[23] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis, "A user independent, biosignal based, emotion recognition method," in *Proceedings of the 11th International Conference on User Modeling*, pp. 314–318, Springer, Corfu, Greece, June 2007.

[24] K. Slot, J. Cichosz, and L. Bronakowski, "Emotion recognition with poincare mapping of voiced-speech segments of utterances," in *Proceedings of 9th International Conference on Artificial Intelligence and Soft Computing (ICAISC '08)*, pp. 886–895, Springer, Zakopane, Poland, June 2008.

[25] C.-H. Wu, J.-F. Yeh, and Z.-J. Chuang, "Emotion perception and recognition from speech," in *Affective Information Processing*, J. Tao and T. Tan, Eds., pp. 93–110, Springer, London, UK, 2009.

[26] I. Maglogiannis, D. Vouyioukas, and C. Aggelopoulos, "Face detection and recognition of natural human emotion using Markov random fields," *Personal and Ubiquitous Computing*, vol. 13, no. 1, pp. 95–101, 2009.

[27] F. Di Fiore, P. Quax, C. Vanaken, W. Lamotte, and F. Van Reeth, "Conveying emotions through facially animated avatars in networked virtual environments," in *Proceedings of the 1st International Workshop on Motion in Games (MIG '08)*, vol. 5277 of *Lecture Notes in Computer Science*, pp. 222–233, Springer, Utrecht, The Netherlands, June 2008.

[28] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale, Eds., vol. 4868 of *Lecture Notes in Computer Science*, pp. 92–103, Springer, Heidelberg, Germany, 2008.

[29] P. D. Turney and M. L. Littman, "Measuring praise and criticism: inference of semantic orientation from association," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 315–346, 2003.

[30] A. Andreevskaia and S. Bergler, "Mining WordNet for fuzzy sentiment: sentiment tag extraction from WordNet glosses," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 209–216, Trento, Italy, April 2006.

[31] A. Esuli and F. Sebastiani, "SentiWordNet: a publicly available lexical resource for opinion mining," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 417–422, Genoa, Italy, May 2006.

[32] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367–1373, Geneva, Switzerland, August 2004.

[33] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*, pp. 412–418, Barcelona, Spain, July 2004.

[34] D. Nadeau, C. Sabourin, J. De Koninck, S. Matwin, and P. D. Turney, "Automatic dream sentiment analysis," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI '06)*, Boston, Mass, USA, July 2006.

[35] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in *Proceedings of 2nd International Conference on Knowledge Capture*, pp. 70–77, Sanibel Island, Fla, USA, October 2003.

[36] K. Moilanen and S. Pulman, "Sentiment composition," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP '07)*, pp. 378–382, Borovets, Bulgaria, September 2007.

[37] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP '05)*, pp. 347–354, Vancouver, Canada, October 2005.

[38] J. Olveres, M. Billinghurst, J. Savage, and A. Holden, "Intelligent, expressive avatars," in *Proceedings of the 1st Workshop on Embodied Conversational Characters (WECC '98)*, pp. 47–55, Tahoe City, Calif, USA, October 1998.

[39] C. Strapparava, A. Valitutti, and O. Stock, "Dances with words," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '07)*, pp. 1719–1724, Hyderabad, India, January 2007.

[40] A. C. Boucouvalas, "Real time text-to-emotion engine for expressive Internet communications," in *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environments*, pp. 306–318, IOS Press, Amsterdam, The Netherlands, 2003.

[41] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proceedings of International Conference on Intelligent User Interfaces (IUI '03)*, pp. 125–132, Miami, Fla, USA, 2003.

[42] M. Mulder, A. Nijholt, M. den Uyl, and P. Terpstra, "A lexical grammatical implementation of affect," in *Proceedings of the 7th International Conference on Text, Speech and Dialogue (TSD '04)*, pp. 171–178, Brno, Czech Republic, September 2004.

[43] C. O. Alm, *Affect in text and speech*, Ph.D. dissertation, University of Illinois, Urbana, Ill, USA, 2008.

[44] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "User study of AffectIM, an emotionally intelligent instant messaging system," in *Proceedings of 8th International Conference on Intelligent Virtual Agents (IVA '08)*, pp. 29–36, Tokyo, Japan, September 2008.

[45] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Textual affect sensing for sociable and expressive online communication," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, vol. 4738 of *Lecture Notes in Computer Science*, pp. 218–229, 2007.

[46] C. E. Izard, *Human Emotions*, Plenum Press, New York, NY, USA, 1977.

[47] C. Strapparava and A. Valitutti, "WordNet-Affect: an affective extension of WordNet," in *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC '04)*, pp. 1083–1086, Lisbon, Portugal, May 2004.

[48] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Analysis of affect expressed through the evolving language of online communication," in *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 278–281, ACM Press, Honolulu, Hawaii, USA, January 2007.

[49] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," in *Advances in Neural Information Processing Systems*, vol. 15, pp. 3–10, MIT Press, Cambridge, Mass, USA, 2003.

[50] S. Aman and S. Szpakowicz, "Using Roget's thesaurus for fine-grained emotion recognition," in *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP '08)*, pp. 296–302, Hyderabad, India, January 2008.

[51] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Proceedings of the 10th International*

Conference on Text, Speech and Dialogue, vol. 4629 of *Lecture Notes in Computer Science*, pp. 196–205, Springer, Plzen, Czech Republic, September 2007.

[52] M. Jarmasz and S. Szpakowicz, "The design and implementation of an electronic lexical knowledge base," in *Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI '01)*, pp. 325–333, Ottawa, Canada, June 2001.

[53] C. E. Izard, *The Face of Emotion*, Appleton-Century-Crofts, New York, NY, USA, 1971.

[54] L. Harker and D. Keltner, "Expressions of positive emotion in women's college yearbook pictures and their relationship to personality and life outcomes across adulthood," *Journal of Personality and Social Psychology*, vol. 80, no. 1, pp. 112–124, 2001.

Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2010, Article ID 830759, 23 pages
doi:10.1155/2010/830759

Research Article

Emotional Communication in Finger Braille

Yasuhiro Matsuda,[1] Ichiro Sakuma,[2] Yasuhiko Jimbo,[3] Etsuko Kobayashi,[2] Tatsuhiko Arafune,[4] and Tsuneshi Isomura[1]

[1] *Faculty of Creative Engineering, Kanagawa Institute of Technology, 1030 Shimo-Ogino, Atsugi-Shi, Kanagawa 243-0292, Japan*
[2] *Graduate School of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo 113-8656, Japan*
[3] *Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-Shi, Chiba 277-8563, Japan*
[4] *Institute for Human Science and Biomedical Engineering, National Institute of Advanced Industrial Science and Technology, 1-2-1 Namiki, Tsukuba-Shi, Ibaraki 305-8564, Japan*

Correspondence should be addressed to Yasuhiro Matsuda, yasuhiro@rm.kanagawa-it.ac.jp

Received 29 March 2009; Revised 12 August 2009; Accepted 25 January 2010

Academic Editor: Kostas Karpouzis

We describe analyses of the features of emotions (neutral, joy, sadness, and anger) expressed by Finger Braille interpreters and subsequently examine the effectiveness of emotional expression and emotional communication between people unskilled in Finger Braille. The goal is to develop a Finger Braille system to teach emotional expression and a system to recognize emotion. The results indicate the following features of emotional expression by interpreters. The durations of the code of joy were significantly shorter than the durations of the other emotions, the durations of the code of sadness were significantly longer, and the finger loads of anger were significantly larger. The features of emotional expression by unskilled subjects were very similar to those of the interpreters, and the coincidence ratio of emotional communication was 75.1%. Therefore, it was confirmed that people unskilled in Finger Braille can express and communicate emotions using this communication medium.

1. Introduction

The Deafblind Association of Japan estimates that there are nearly 22,000 deafblind people in Japan (2008). Deafblind people use many different communication media, depending on the age of onset of deafness and blindness and the available resources. "*Yubi-Tenji*" (Finger Braille) is one of the tactual communication media utilized by deafblind individuals (see Figure 1). In two-handed Finger Braille, the index finger, middle finger, and ring finger of both hands function like the keys of a Braille typewriter. A sender dots Braille code on the fingers of a receiver as if typing on a Braille typewriter. The receiver is assumed to be able to recognize the Braille code. In one-handed Finger Braille, the sender first dots the left column of Braille code on the distal interphalangeal (DIP) joints of the three fingers of the receiver and then dots the right column of Braille code on the proximal interphalangeal (PIP) joints. Deafblind people who are skilled in Finger Braille can understand the speech conversation and express various emotions because of the prosody (intonation) of Finger Braille [1]. A rule of Finger Braille is that the sender keeps touching the fingers of the receiver even when not dotting, because receivers feel uneasy in the absence of touching or tactile cues. Because there is such a small number of non-disabled people who are skilled in Finger Braille, deafblind people communicate only through an interpreter.

Various Braille input devices have recently been developed [2, 3], but they require deafblind people to wear gloves or type on a keyboard to input the Finger Braille, or to use actuators to output and convert the speech of non-disabled people to Finger Braille. With these devices, deafblind people are burdened with wearing sensors and actuators, and they must master a new communication system with such support devices.

The objective of this study is the development of a Finger Braille support device which employs the skin-contact communication of deafblind people, because skin

contact is the only form of nonverbal communication for deafblind people. The concept of the Finger Braille support device is shown in Figure 2. The advantages of this device are as follows: both deafblind people and non-disabled people unskilled in Finger Braille can communicate using conventional Finger Braille, and deafblind people are not encumbered by a support device because the non-disabled people operate the support device and wear all of the sensors.

Our support device consists of a Finger Braille teaching system and a Finger Braille recognition system. The teaching system recognizes the speech of a non-disabled person and displays the associated dot pattern of Finger Braille. The non-disabled person can then dot Finger Braille on the fingers of the deafblind person by observing the displayed dot pattern [4, 5]. The recognition system recognizes the dotting of Finger Braille by the deafblind person and synthesizes this tactile communication into speech for the non-disabled person [6]. The intent of the support device is to assist not only verbal communication but also nonverbal (emotional) communication. To assist in emotional communication, we have been developing an emotion teaching system and an emotion recognition system [7]. The emotion teaching system teaches non-disabled people to express emotions by skin contact. The emotion recognition system recognizes the emotions dotted by the deafblind and presents the information as synthesized speech for the non-disabled.

In the present study, emotional expression by Finger Braille interpreters (Experiment I) and emotional communication by people who are unskilled in Finger Braille (Experiment II) were examined. This paper presents analyses of the features of emotional expression by Finger Braille interpreters and the effectiveness of emotional expression and emotional communication between unskilled people.

2. Emotional Expression by Finger Braille Interpreters (Experiment I)

To analyze the features of emotional expression using Finger Braille, emotional expression by interpreters of Finger Braille (Experiment I) was examined [8, 9].

2.1. Methods. The subjects were two Finger Braille interpreters (Subjects A and B) with 20 and 11 years of experience, respectively. Both subjects were right-handed. Both subjects gave informed consent after hearing a description of the study. In the experiment, the subject became a sender and a tester became a receiver. The same tester, who was one of the authors, was used throughout the experiment.

Two dialogues were used. "Rain has fallen" or *Ame futte kita* ⠠⠍⠑ ⠋⠥⠞⠞⠑ ⠅⠊⠞⠁ (7 characters, 3 clauses, "*t*" is a double consonant, "*me*" and "*te*" are characters at the end of a clause) and "It is the time" or *Jikan desuyo* ⠘⠚⠊ ⠅⠁⠝ ⠙⠑⠎⠥⠽⠕ (8 characters, 1 clause, "*Ji*" is a combination of voiced sound and "*si*", "*de*" is a combination of voiced sound and "*te*"). These dialogues did not include words about emotion.

Plutchik proposed a multidimensional model of emotions and defined joy, anticipation, anger, disgust, sadness, surprise, fear, and acceptance as the basic human emotions [10]. Bridges observed emotional development in early infancy and found that distress (sadness), delight (joy) and anger developed from excitement in early infancy [11].

We decided to investigate the expressed emotions of neutral, joy, sadness, and anger, which are basic emotions developed in early infancy of human beings. These emotions were common to those in several previous studies of emotional expression and communication [12–20].

In the previous studies, actors or announcers were selected as the subjects, because of their experiences expressing emotion [16, 18]. In the present study, the subjects were neither actresses nor announcers; so we set six scenes (two dialogues with joy, sadness, and anger) for the subjects to express the emotions easily. The scenes are as follows.

Scenes of "Rain has fallen"

Joy: In the morning of a sport activity, you do not want to participate. If it is raining, the activity would be cancelled. Now rain is falling.

Sadness: On a holiday morning, the weather is fine. You are changing your clothes to go outside. Now rain is falling.

Anger: The probability of rain is 0% today. You have gone out without an umbrella. Now rain is falling.

Scenes of "It is the time"

Joy: A TV program that you want to watch is going to start. You call your family.

Sadness: You have come to an airport to send off your friend who is going abroad to study. Now it is time to depart.

Anger: It is the closing time of your shop, but a customer will not go home.

In this experiment, we adopted two-handed Finger Braille (the Perkins Brailler type, which is the most popular type of Finger Braille) and one-handed (right hand) Finger Braille.

In the two-handed Finger Braille experiment, the variables compose a 4×5 within-subjects design of expressed emotion (neutral, joy, sadness, anger) and position of characters (inner clause, end of clause, end of sentence, voiced sound, double consonant). In the one-handed Finger Braille experiment, the variables compose a 4×6 within-subjects design of expressed emotion (neutral, joy, sadness, anger) and position of characters (inner character, inner clause, end of clause, end of sentence, voiced sound, double consonant). The dependent variables are the duration of code, duration of pause, and finger load. The inner character is the dotting of DIP joints if the character includes the dotting of both DIP and PIP joints.

In our previous work [7, 8], we designed each emotion factor and each character factor and conducted two analyses of variance (ANOVA) for both dialogues (e.g., the 4×7 within-subjects design in "Rain" of the two-handed Finger Braille experiment, and the 4×8 within-subjects design

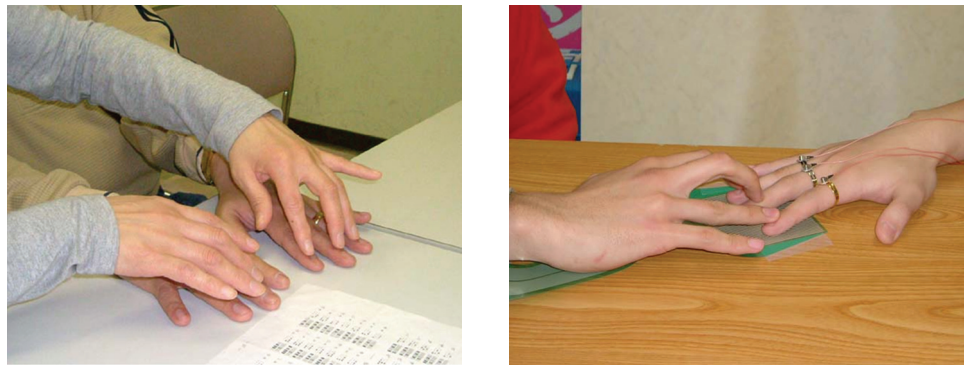


Figure 1: Two-handed Finger Braille and one-handed Finger Braille.

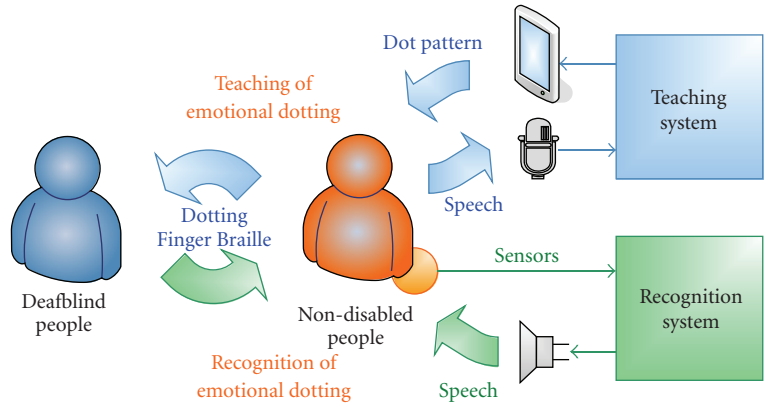


Figure 2: Finger Braille support device.

in "Time" of the two-handed Finger Braille experiment). It was difficult to generalize the features of emotional expression and prosody using the character factor of the specific dialogues. Thus, in the present study, we redesigned this experiment.

The experimental flow is shown in Figure 3. First, the two-handed Finger Braille experiment was conducted followed by the one-handed Finger Braille (right-hand) experiment, with each experiment including one practice session and five experimental sessions. The subject rested after every 2-3 sessions.

In one experimental session, the subjects dotted Finger Braille on the fingers of the tester eight times in the order of "Rain" with neutral, "Time" with neutral, "Rain" with joy, "Time" with joy, "Rain" with sadness, "Time" with sadness, "Rain" with anger, and "Time" with anger. The dialogue and emotion associated with the dot was displayed on the LCD of a notebook PC that was placed in front of the subject. The subject did not hear and transfer the emotional speech of the dialogues but expressed the impressions of emotions using Finger Braille. The subject did not alter the dialogue in any way and only expressed emotions by changing the dotting speed and applied pressure. To eliminate the influence of the previous dotted dialogue with emotion, the subject dotted at intervals (1-2 minutes) and confirmed the next scene of dialogue with emotion.

As is customary in communication using Finger Braille, the subject and tester sat side by side during the two-handed Finger Braille experiment and face to face during the one-handed Finger Braille experiment. The hands of the subject and tester were in constant contact during the two-handed Finger Braille experiment.

The tester put his fingers on a pressure sensor sheet (Tactile Sensor System, Nitta) which measured the change of pressure as a result of dotting. The sampling frequency was 137 Hz, the measurement range was 15–150 kPa, the sensibility was 527 Pa, the size of the sensor sheet was 84 × 84 mm, and the sensor matrix was 44 × 44. The sensor sheet was segmented into three blocks (width: 28 mm) and the tester put his index, middle, and ring fingers into the blocks.

2.2. Results

2.2.1. Calculation of Duration and Finger Load. The load of each finger was determined by the change in pressure applied to the three blocks, and the start, end, and maximum load of dotting were measured. The duration and finger load for each dotting were also calculated. Sample data from the one-handed Finger Braille experiment are shown in Figure 4.

The duration of the code was defined as the time from the start of dotting to the end of dotting. The duration of the pause was defined as the duration from the completion of

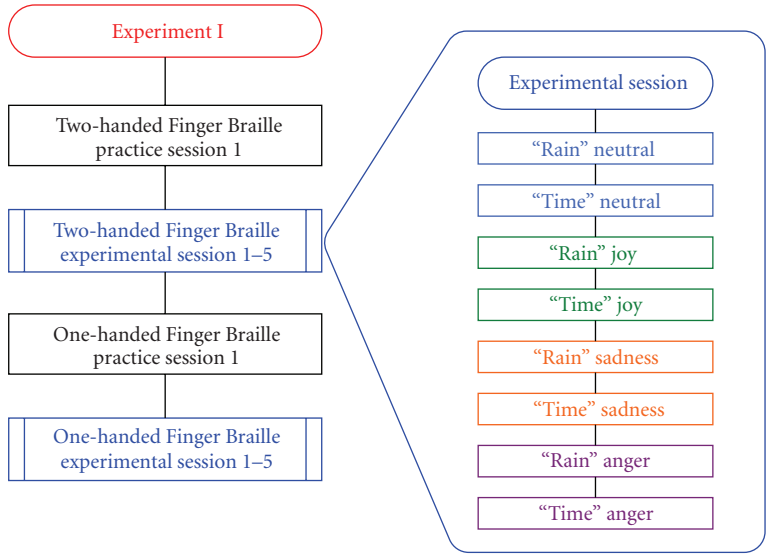


Figure 3: Flow of Experiment I.

dotting to the start of the next dotting sequence. The finger load was defined as the difference between the maximum load and the mean load between the start and end of dotting. If multiple fingers were dotted at once, the finger load was defined as the mean of the finger loads for the dotted fingers.

2.2.2. Results of Two-Handed Finger Braille Experiment

(a) Duration of Code. The mean of the duration of the code as a function of the emotions and position of characters is shown in Figure 5.

A 4 × 5 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,580) = 119.3$, $P < .001$) and the position of characters ($F(4,580) = 121.0$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(12,580) = 15.73$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all the positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of the characters revealed the following: the duration of code of sadness was significantly longer than the other durations in the inner clause ($P < .001$); the duration of code of joy was significantly shorter than the other durations in the inner clause ($P < .001$); the duration of code of anger was significantly shorter than the duration of code of neutral in the inner clause ($P < .001$); the duration of code of sadness was significantly longer than the other durations in the end of clause ($P < .041$); the duration of code of sadness was significantly longer than the other durations in the end of the sentence ($P < .001$); the duration of code of neutral was significantly longer than the durations of code of joy and anger in the end of the sentence ($P < .001$); the duration of code of sadness was significantly longer than the other durations in the voiced sound ($P < .050$); the duration of code of joy was significantly shorter than the other durations in the voiced sound ($P < .001$); the duration of code of joy was significantly shorter than the other durations in the double consonant ($P < .011$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .001$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the durations of code of the end of the sentence were significantly longer than the other durations in neutral ($P < .001$); the durations of code of the voiced sound and double consonant were significantly shorter than the other durations in neutral ($P < .011$); the duration of code of the end of the clause and end of the sentence was significantly longer than the other durations in joy ($P < .020$); the duration of code of the end of the sentence was significantly longer than the other durations in sadness ($P < .001$); the duration of code of the end of the clause was significantly longer than the durations of code of the inner clause, voiced sound, and double consonant in sadness ($P < .003$); the duration of code of the double consonant was significantly shorter than the other durations in anger ($P < .001$).

(b) Duration of Pause. The mean of the duration of the pause as a function of the emotions and position of characters is shown in Figure 6. The duration of pause before the double consonant was shorter than the other durations of pause. So, we defined the duration of pause before the double consonant as the duration of pause of the double consonant and the duration of pause after the double consonant as the duration of pause of the inner clause. There was no duration of pause at the end of sentences.

A 4 × 4 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,504) = 5.615$, $P < .001$) and the position of characters ($F(3,504) = 51.837$, $P < .001$). Scheffe tests on the emotion factor revealed that the durations of pause of sadness and neutral

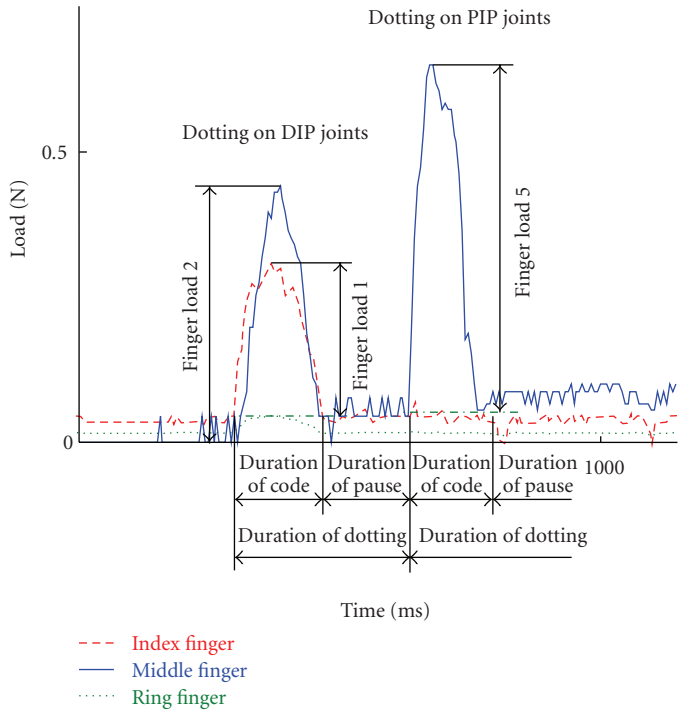


Figure 4: Calculation of durations and finger loads (one-handed Finger Braille).

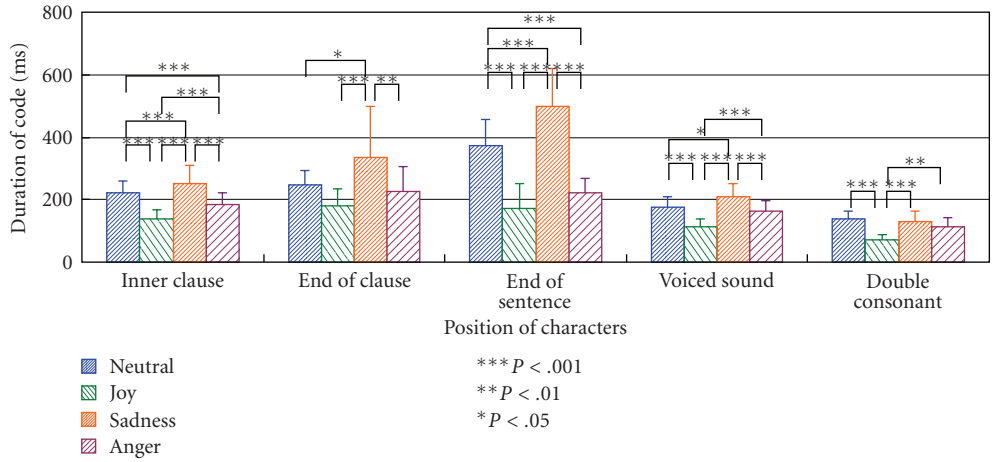


Figure 5: Mean of duration of code as a function of emotions and position of characters for two-handed Finger Braille (error bars represent standard deviations).

were significantly longer than the durations of pause of anger and joy ($P < .003$). Scheffe tests on the factor of the position of characters revealed that the duration of pause of the double consonant was significantly shorter than the other durations ($P < .001$), and the duration of pause of the end of the clause was significantly longer than the other durations ($P < .001$).

(c) Finger Load. The mean of the finger load as a function of the emotions and position of characters is shown in Figure 7.

A 4 × 5 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,580) = 344.6$, $P < .001$) and the position of characters ($F(4,580) = 31.49$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(12,580) = 16.52$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the finger load of anger was significantly larger than the other finger loads in all positions of characters ($P < .001$); the finger load of joy was significantly larger than the finger loads of neutral and sadness in the inner clause ($P < .036$); the finger load of joy was significantly larger than the finger load of sadness in the end of the clause and the voiced sound ($P < .017$).

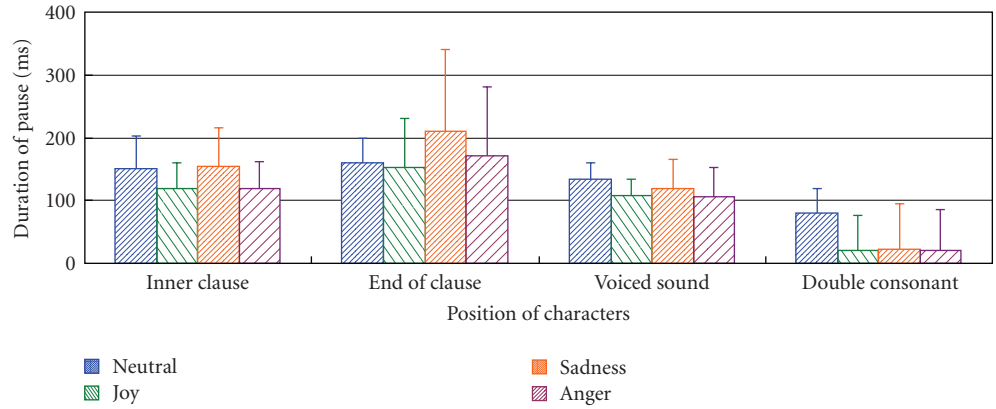


Figure 6: Mean of duration of pause as a function of emotions and position of characters for two-handed Finger Braille (error bars represent standard deviations).

A test of the simple main effect of position of characters revealed a significant effect of position of characters in all emotions ($P < .001$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the finger load of the voiced sound was significantly larger than the finger loads of the end of clauses, inner clause and end of sentence in neutral ($P < .001$); the finger load of the voiced sound was significantly larger than the other finger loads in joy ($P < .002$); the finger load of the voiced sound was significantly larger than the other finger loads in sadness ($P < .006$); the finger loads of the end of the sentence and voiced sound were significantly larger than the other finger loads in anger ($P < .007$).

2.2.3. Results of One-Handed Finger Braille Experiment

(a) Duration of Code. The mean of the duration of the code as a function of the emotions and position of characters is shown in Figure 8.

A 4 × 6 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,1013) = 180.3$, $P < .001$) and the position of characters ($F(5,1013) = 75.36$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(15,1013) = 14.06$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the duration of code of sadness was significantly longer than the other durations in the inner character ($P < .001$); the duration of code of joy was significantly shorter than the other durations in the inner character ($P < .001$); the duration of code of anger was significantly shorter than the duration of code of neutral in the inner character ($P < .008$); the duration of code of sadness was significantly longer than the other durations in the inner clause ($P < .001$); the duration of code of joy was significantly shorter than the other durations in the inner clause ($P < .001$); the duration of code of sadness was significantly longer than the other durations in the end of clause ($P < .003$); the duration of code of joy was significantly shorter than the duration of code of neutral in the end of clause ($P < .048$); the duration of code of sadness was significantly longer than the other durations in the end of sentence ($P < .001$); the duration of code of joy was significantly shorter than the duration of code of neutral in the end of sentence ($P < .003$); the duration of code of sadness was significantly longer than the other durations in the voiced sound ($P < .015$); the duration of code of joy was significantly shorter than the duration of anger in the voiced sound ($P < .024$); the durations of code of sadness and neutral were significantly longer than the durations of code of joy and anger in the double consonant ($P < .019$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .001$). Scheffe tests on the factor of the position of characters in each emotion revealed that the duration of code of the end of sentence was significantly longer than the other durations in all emotions ($P < .030$), and the duration of code of the end of clause was significantly longer than the duration of the double consonant in sadness ($P < .024$).

(b) Duration of Pause. The mean of the duration of the pause as a function of the emotions and position of characters is shown in Figure 9. We defined the duration of pause before the double consonant as the duration of pause of the double consonant and the duration of pause after the double consonant as the duration of pause of the inner clause. These definitions are the same as those for the two-handed Finger Braille experiment. There was no duration of pause at end of the sentences.

A 4 × 5 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,935) = 4.009$, $P < .008$) and the position of characters ($F(4,935) = 4.101$, $P < .003$). Scheffe tests on the emotion factor revealed that the duration of pause of sadness was significantly longer than the other durations ($P < .036$). Scheffe tests on the factor of the position of characters revealed that the duration of pause of the end of the clause was significantly longer than the duration of pause of the inner character ($P < .005$).

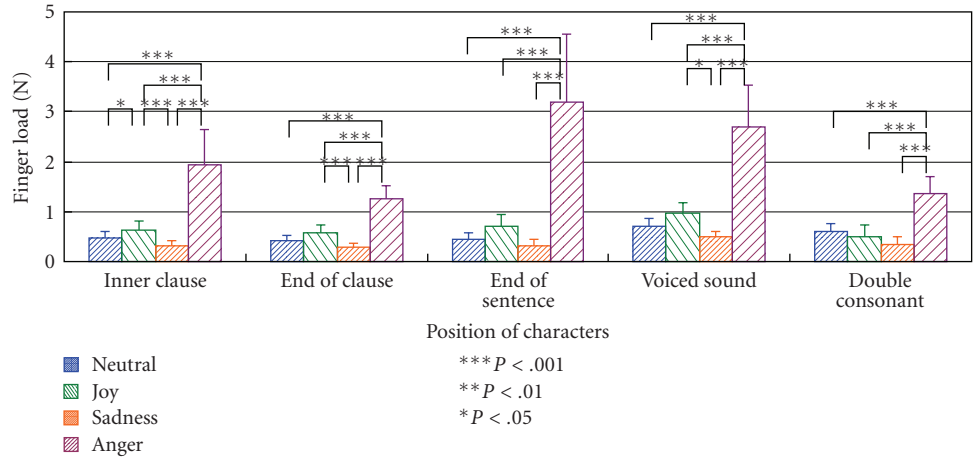


Figure 7: Mean of finger load as a function of emotions and position of characters for two-handed Finger Braille (error bars represent standard deviations).

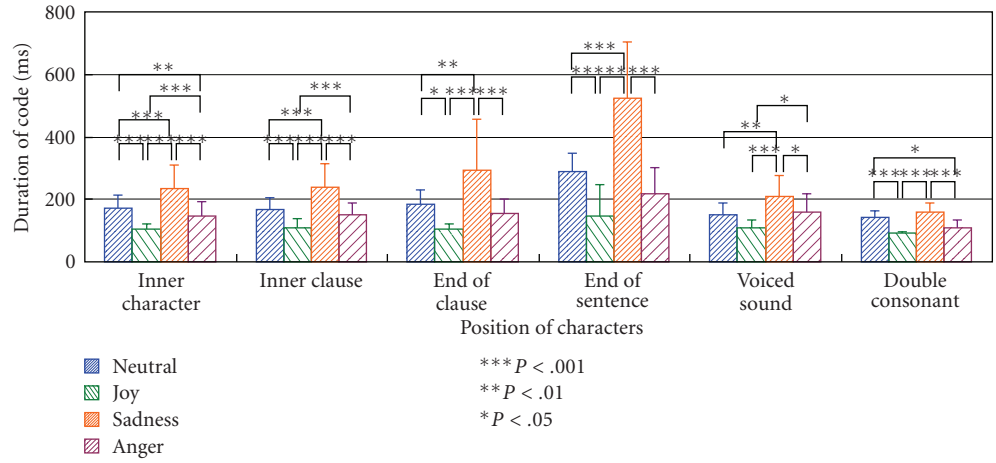


Figure 8: Mean of duration of code as a function of emotions and position of characters for one-handed Finger Braille (error bars represent standard deviations).

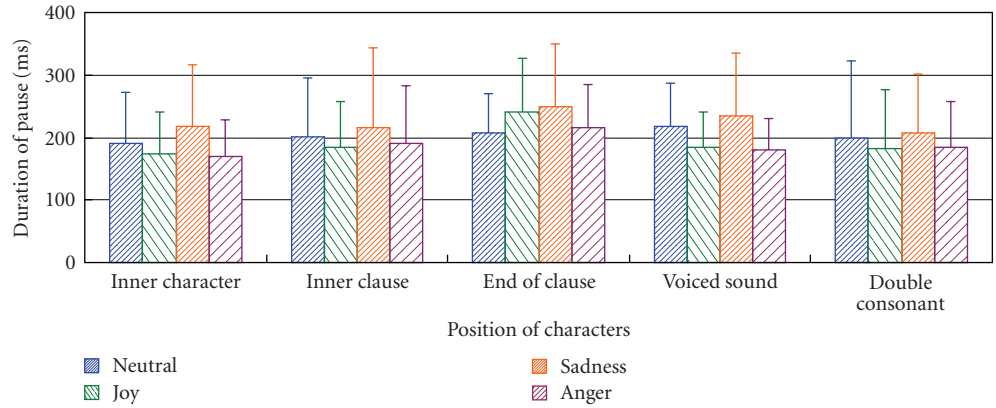


Figure 9: Mean of duration of pause as a function of emotions and position of characters for one-handed Finger Braille (error bars represent standard deviations).

Table 1: Responses of interpreters to questionnaire on expressing emotions.

Emotion	Subject	Answer
Neutral	A	Conveying the meaning exactly to someone in view
	B	Imaging synthesized speech
Joy	A	Rhythmically
	B	Airily
Sadness	A	Uninterrupted dotting, like wrapping someone's hand
	B	Negatively speaking
Anger	A	Like shoving someone away
	B	Like dashing words over someone

(c) Finger Load. The mean of the finger load as a function of the emotions and position of characters is shown in Figure 10.

A 4 × 6 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,1013) = 261.4$, $P < .001$) and the position of characters ($F(5,1013) = 23.11$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(15,1013) = 5.496$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the finger load of anger was significantly larger than the other finger loads in all positions of characters ($P < .001$), and the finger load of joy was significantly larger than the finger loads of neutral and sadness in the inner character, inner clause, and end of clause ($P < .011$); the finger load of joy was significantly larger than the finger load of sadness in the double consonant ($P < .030$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .034$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the finger loads of the double consonant, end of sentence, and inner character were significantly larger than the finger loads of the end of clause and inner clause in neutral ($P < .034$); the finger load of the end of sentence was significantly larger than the finger loads of the end of clause, inner clause, voiced sound, and inner character in anger ($P < .013$).

2.2.4. Questionnaire Results (Interpreters). After the experiment, the subjects were asked how they expressed the emotions, and their responses are presented in Table 1. In addition, the subjects answered that they sometimes hit the hand of the receiver or dotted by claw-like fingers in anger.

Subjects also responded that it was more difficult to express emotions using one-handed Finger Braille with people sitting face to face than using two-handed Finger Braille sitting side by side, because the skin contact of the hands and the distance of each person were considered important for emotional communication.

Table 2: Standardized coefficients of canonical discriminant functions and contribution ratios (two-handed Finger Braille).

Discriminant variable	Function 1	Function 2	Function 3
Duration of code	−0.005	0.990	−0.272
Duration of pause	0.005	0.100	1.015
Finger load	1.000	0.164	0.059
Contribution ratio	84.3%	15.6%	0.1%
Significance probability	$P < .001$	$P < .001$	$P < .237$

Table 3: Standardized coefficients of canonical discriminant functions and contribution ratios (one-handed Finger Braille).

Discriminant variable	Function 1	Function 2	Function 3
Duration of code	−0.560	0.820	−0.161
Duration of pause	0.053	0.248	0.993
Finger load	0.903	0.494	0.058
Contribution ratio	76.0%	23.9%	0.1%
Significance probability	$P < .001$	$P < .001$	$P < .508$

2.3. Discussion

2.3.1. Features of Emotional Expression Using Finger Braille. To analyze the contribution of the variables to emotional expression, a discriminant analysis was conducted on data from both the two-handed and one-handed Finger Braille experiments ($N = 520$ and 955, resp.). The discriminant variables were the duration of code, duration of pause, and finger load. Dotting of the ends of sentences was excluded, because there was no duration of pause.

Standardized coefficients of canonical discriminant functions and contribution ratios were calculated for both the two-handed and one-handed Finger Braille experiments. The results are presented in Tables 2 and 3, respectively.

According to the standardized coefficients, discriminant function 1 represented the finger load, discriminant function 2 represented the duration of code, and discriminant function 3 represented the duration of pause. According to the contribution ratios, the finger load was the variable contributing the most to emotional expression, followed by the duration of code. The duration of pause did not contribute to emotional expression.

The discriminant ratios of emotion for two-handed and one-handed Finger Braille are presented in Tables 4 and 5, respectively. The total discriminant ratios were 68.5% in two-handed Finger Braille and 71.1% in one-handed Finger Braille. The discriminant ratio of joy was the highest (two-handed Finger Braille 86.9%, one-handed Finger Braille 85.2%); the discriminant ratio of anger was second (two-handed Finger Braille 80.8%, one-handed Finger Braille 69.9%). The dottings of neutral and sadness were frequently misdiscriminated with each other.

As indicated by the results of ANOVA (2.2.2–2.2.3), the features of emotional expression were as follows: (1) the duration of the code of joy was significantly shorter than that of the other emotions; (2) the duration of the code of sadness was significantly longer than that of the other emotions; (3)

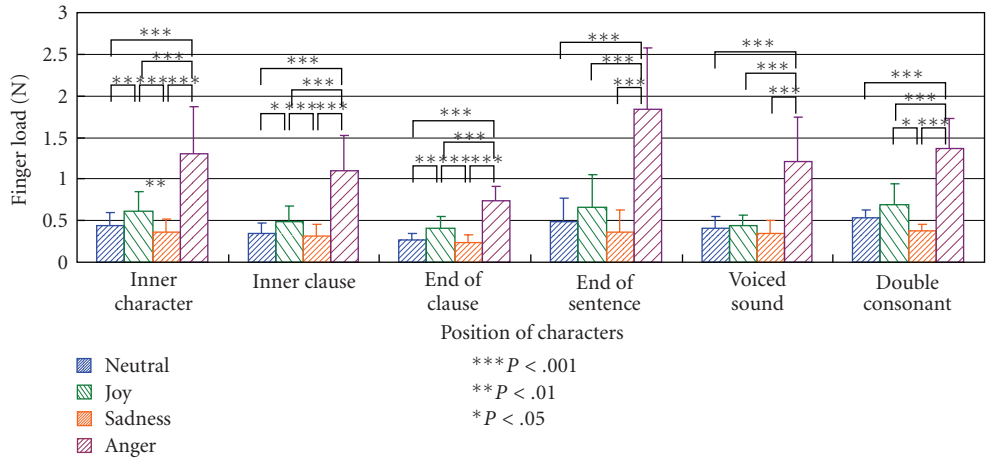


Figure 10: Mean of finger load as a function of emotions and position of characters for one-handed Finger Braille (error bars represent standard deviations).

Table 4: Discriminant ratios of emotion (two-handed Finger Braille).

Expressed emotion	Discriminated emotion				Discriminant ratio
	Neutral	Joy	Sadness	Anger	
Neutral	62	25	43	0	47.7%
Joy	14	113	3	0	86.9%
Sadness	38	16	76	0	58.5%
Anger	14	9	2	105	80.8%
Total	128	163	124	105	68.5%

Table 5: Discriminant ratios of emotion (one-handed Finger Braille).

Expressed emotion	Discriminated emotion				Discriminant ratio
	Neutral	Joy	Sadness	Anger	
Neutral	159	38	40	3	66.3%
Joy	20	201	2	13	85.2%
Sadness	83	5	152	0	66.3%
Anger	13	55	4	167	69.9%
Total	275	299	198	183	71.1%

the finger load of anger was significantly larger than that of the other emotions; (4) the finger load of joy was significantly larger than that of sadness and neutral; (5) the duration of the code of anger was significantly shorter than that of sadness and neutral.

Next we discuss the features of emotional expression by the other communication media. In our previous work [9], we presented a brief discussion of other media. In the present study, we go into more detail.

Laukka analyzed the emotional expression of speech [12]. The results of his analysis were as follows: anger was characterized by a high mean pitch, high voice intensity, and fast speech rate; happiness was characterized by a high mean pitch, medium-high voice intensity, and fast speech rate; sadness was characterized by a low mean pitch, low voice intensity, and slow speech rate.

Bresin and Friberg analyzed the emotional expression of music [13]. The results were as follows: anger was characterized by a very rapid tempo, loud sound level, and mostly nonlegato articulation; happiness was characterized by a fast tempo, moderate or loud sound level, and airy articulation; sadness was characterized by a slow tempo, low sound level, and legato articulation.

Dahl and Friberg analyzed the emotional expression of a musician's body movement in performances on the marimba [14]. The results were the following: anger was characterized by large, fast, uneven, and jerky movements; happiness was characterized by large and somewhat fast movements; sadness was characterized by small, slow, even, and smooth movements.

Clynes and Panksepp proposed a "sentic form" to express the fundamental human emotions. Sentic form can be measured by the patterns of finger pressure [15]. The results were the following: anger was characterized by high pressure and short duration; joy was characterized by short duration; sadness was characterized by low pressure and long duration.

The duration of code in the present study (Experiment I) was similar to the speech rate, tempo of music, speed of movement, and duration of finger pressure. The finger load in Experiment I was similar to the voice intensity, sound level of music, largeness of movement, and strength of finger pressure. Therefore, the features of emotional expression using Finger Braille are similar to the features of the emotional expressions of speech, music, body movement, and finger pressure.

2.3.2. Features of Prosody of Finger Braille. As indicated by the results of ANOVA (2.2.2–2.2.3), the subjects dotted dialogues with prosody (intonation) in addition to emotional expression. The followings were the features of prosody: (1) the duration of the code of the end of the clause was significantly longer; (2) the duration of the pause of the end of the clause was significantly longer; (3) the duration of the code

of the end of the sentence was significantly longer; (4) the finger load of the end of the sentence was significantly larger; (5) the duration of the code of the double consonant was significantly shorter; (6) the duration of the pause before the double consonant was significantly shorter; (7) the duration of the code of the voiced sound was significantly shorter; (8) the finger load of the voiced sound was significantly larger.

Dousaka et al. and Fujimori et al. analyzed the features of prosody of Finger Braille [21, 22]. A comparison of the durations of code, pause, and dotting between the four emotions of two-handed Finger Braille in our study and in previous studies is presented in Figure 11. The values of the standard deviation were not clear in the previous studies, but the durations of code and pause of Dousaka's study were similar to the durations of code and pause of joy and anger in the present study; the duration of dotting of Fujimori's study was similar to the durations of dotting of neutral and sadness. The range of the duration of dotting between Dousaka's study and Fujimori's study was also similar to the range of the duration of dotting of the four emotions in the present study. Thus, it was considered that the features of prosody of Experiment I are natural variances of the dotting of Finger Braille.

Miyagi et al. analyzed the changes of prosody in emphasized clauses [23]. The followings are the features for prosody without emphasized clauses: the characters at the end of the clauses were dotted strongly; the duration of pauses after the ends of the clauses was long; the duration of codes at the ends of the sentences was long. When clauses were emphasized, the characters at the ends of clauses were dotted longer and more strongly; the duration of pauses after the ends of clauses was longer. In the present study, the durations of code at the ends of clauses and sentences were longer for sadness; the finger loads at the end of sentences were stronger for anger (Experiment I). These results indicate that the prosody of Finger Braille must also be similarly changed to express emotions and to emphasize clauses.

3. Additional Experiment of Emotional Expression by Finger Braille Interpreter

To confirm the features of emotional expression using Finger Braille, an additional experiment of emotional expression by the Finger Braille interpreter was conducted.

3.1. Methods. The subject was Subject A in Experiment I. The additional experiment was conducted one year and eight months after Experiment I. The subject gave informed consent after hearing a description of the study.

Two new dialogues were used. "Let's hurry up" or *Hayaku shimasyou* ⠠⠓⠁⠽⠁⠅⠥ ⠠⠩⠓⠊⠍⠁⠎⠽⠕⠥ (8 characters, 2 clauses, "*syo*" is a combination of a diphthong and "*so*", "*ku*" is a character at the end of a clause), and "Is it really?" or *Hontou desuka* ⠠⠓⠕⠝⠞⠕⠥ ⠠⠙⠑⠎⠥⠅⠁ (8 characters, 1 clause, "*de*" is a combination of a voiced sound and "*te*"). These dialogues did not include words about emotion.

The expressed emotions were neutral, joy, sadness, and anger, which are identical to those used in Experiment I. We set six scenes (two dialogues with joy, sadness, and anger) for the subject. The scenes were as follows.

Scenes of "Let's hurry up"

Joy: You are planning to go out with your close friend.

Sadness: Your family member has to leave home because he has a new job. Now you help pack his bag.

Anger: Your work is late but your colleague has not started to work.

Scenes of "Is it really?"

Joy: You have received news that you passed a difficult examination.

Sadness: You have received news that your close friend is in the hospital.

Anger: Suddenly, you have been wrongly accused.

In this experiment, we adopted two-handed Finger Braille (the Perkins Brailler type) and one-handed (right hand) Finger Braille, as in Experiment I.

In the two-handed Finger Braille experiment, the variables compose a 4 × 4 within-subjects design: expressed emotion (neutral, joy, sadness, anger) and position of characters (inner clause, end of clause, end of sentence, voiced sound). In the one-handed Finger Braille experiment, the variables compose a 4 × 5 within-subjects design: expressed emotion (neutral, joy, sadness, anger) and position of characters (inner character, inner clause, end of clause, end of sentence, voiced sound). The diphthong was categorized as a voiced sound. The dependent variables were duration of code, duration of pause, and finger load.

The experimental flow is shown in Figure 12. First, the one-handed Finger Braille (right hand) experiment was conducted, followed by the two-handed Finger Braille. Each experiment included one practice session and five experimental sessions. The subject rested after every 2-3 sessions.

In the experimental session, the subject dotted Finger Braille on the fingers of the tester eight times in the order of "Hurry" with neutral, "Really" with neutral, "Hurry" with joy, "Really" with joy, "Hurry" with sadness, "Really" with sadness, "Hurry" with anger, and "Really" with anger.

The other procedure and conditions were identical to those used in Experiment I.

3.2. Results. We calculated the duration of code, duration of pause, and finger load for each dotting by the change of pressure, as in Experiment I (see Figure 4).

3.2.1. Results of Two-Handed Finger Braille Experiment

(a) Duration of Code. The mean of the duration of the code as a function of the emotions and position of characters is shown in Figure 13.

A 4 × 4 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,283)$ =

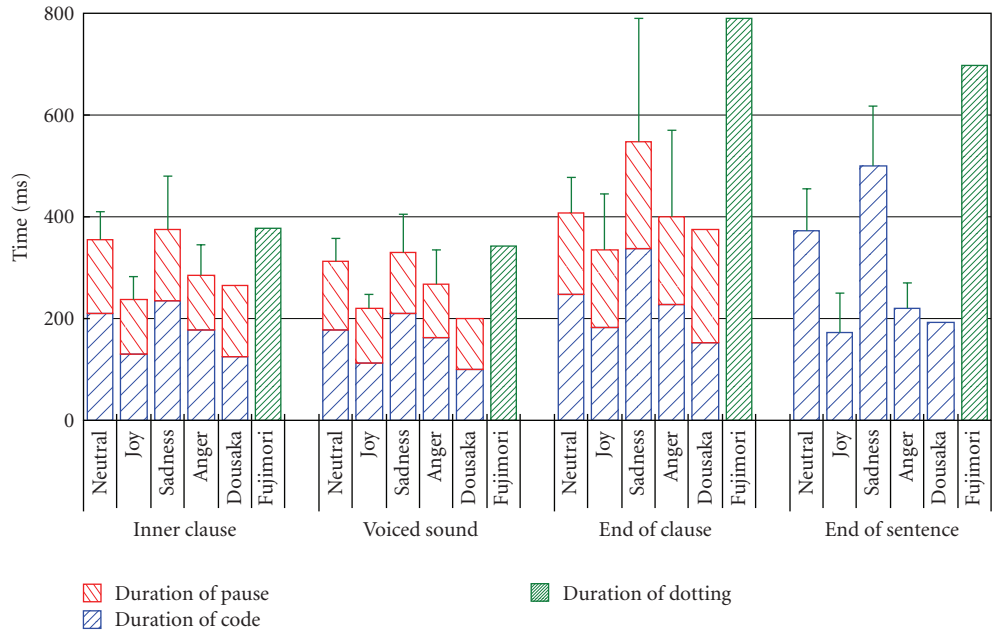


Figure 11: Comparison of durations of code, pause, and dotting between the four emotions in our study and in previous studies [21, 22] for two-handed Finger Braille (error bars represent standard deviations of duration of dotting).

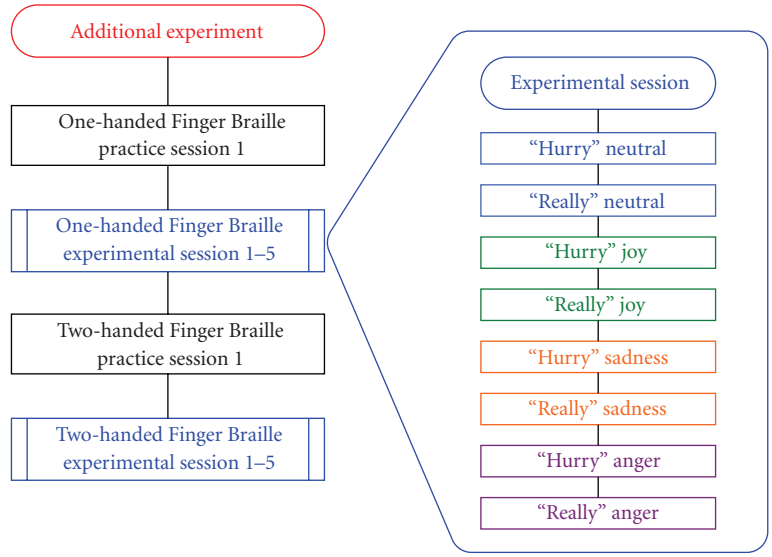


Figure 12: Flow of addition to Experiment I.

152.8, $P < .001$) and the position of characters ($F(3,283) = 56.87$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(9,283) = 9.918$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the duration of code of sadness was significantly longer than the other durations in all positions of characters ($P < .001$, the voiced sound $P < .015$); the duration of code of joy was significantly shorter than the other durations in the inner clause, end of sentence, and voiced sound ($P < .048$, the inner clause $P < .001$); the duration of code of joy was significantly shorter than the durations of code of sadness and anger in the end of clause ($P < .006$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .001$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the duration of code of the end of the sentence was significantly longer than the durations of code of the inner clause and voiced sound in neutral ($P < .003$); the duration of code of the voiced sound was significantly shorter

than the durations of code of the end of the sentence and the inner clause in joy ($P < .021$); the duration of code of the voiced sound was significantly shorter than the other durations in sadness and anger ($P < .006$); the durations of code of the end of the clause and end of the sentence were significantly longer than the other durations in sadness and anger ($P < .004$).

(b) Duration of Pause. The mean of the duration of the pause as a function of the emotions and position of characters is shown in Figure 14.

A 4 × 3 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,247) = 5.353$, $P < .001$) and the position of characters ($F(2,247) = 15.09$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(6,247) = 2.993$, $P < .008$).

A test of the simple main effect of emotion revealed a significant effect of emotion in the end of the clause and voiced sound ($P < .003$). Scheffe tests on the emotion factor in these position of characters revealed the following: the duration of pause of anger was significantly longer than the durations of pause of neutral and joy in the end of the clause ($P < .045$); the duration of pause of sadness was significantly longer than the duration of pause of neutral in the end of clause ($P < .038$); the duration of pause of joy was significantly shorter than the durations of pause of neutral and anger in the voiced sound ($P < .042$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .050$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the duration of pause of the end of the clause was significantly longer than the duration of pause of the inner clause in joy and sadness ($P < .046$); the duration of pause of the end of the clause was significantly longer than the other durations in anger ($P < .002$).

(c) Finger Load. The mean of the finger load as a function of the emotions and position of characters is shown in Figure 15.

A 4 × 4 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,283) = 229.5$, $P < .001$) and the position of characters ($F(3,283) = 50.11$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(9,283) = 6.707$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the finger load of anger was significantly larger than the other finger loads in all positions of characters ($P < .001$); the finger load of sadness was significantly smaller than the other finger loads in the inner clause, end of sentence, and voiced sound ($P < .001$); the finger load of joy was significantly larger than the finger load of neutral in the inner clause and end of clause ($P < .028$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .013$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the finger load of the voiced sound was significantly larger than the finger loads of the end of clauses and inner clause in neutral, joy, and anger ($P < .033$); the finger load of the end of sentence was significantly larger than the finger loads of the end of clauses and inner clause in joy and anger ($P < .041$).

3.2.2. Results of One-Handed Finger Braille Experiment

(a) Duration of Code. The mean of the duration of the code as a function of the emotions and position of characters is shown in Figure 16.

A 4 × 5 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,579) = 109.9$, $P < .001$) and the position of characters ($F(4,579) = 29.90$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(12,579) = 12.27$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$, the end of clause $P < .034$). Scheffe tests on the emotion factor in each position of characters revealed the following: the duration of code of sadness was significantly longer than the other durations in the inner character, inner clause, and end of sentence ($P < .001$); the duration of code of joy was significantly shorter than the other durations in the inner character, inner clause, and end of sentence ($P < .028$); the duration of code of sadness was significantly longer than the durations of code of joy and anger in the voiced sound ($P < .007$); the duration of code of joy was significantly shorter than the duration of neutral in the voiced sound ($P < .024$).

A test of the simple main effect of the position of characters revealed a significant effect of position of characters in all emotions ($P < .001$, neutral $P < .031$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the duration of code of the end of clause was significantly longer than the other durations in joy ($P < .001$); the durations of code of the end of clause and end of sentence were significantly longer than the durations of code of the inner character, inner clause and voiced sound in sadness ($P < .025$); the durations of code of the end of clause and end of sentence were significantly longer than the duration of code of the inner character in anger ($P < .021$).

(b) Duration of Pause. The mean of the duration of the pause as a function of the emotions and position of characters is shown in Figure 17.

A 4 × 4 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,541) = 4.390$, $P < .005$) and the position of characters ($F(3,541) = 7.827$, $P < .001$). Scheffe tests on the emotion factor revealed that the duration of pause of neutral was significantly longer than the other durations ($P < .001$). Scheffe tests on the factor of the position of characters revealed that the duration of pause of the inner character was significantly shorter than the durations of pause of the inner clause and end of clause ($P < .009$).

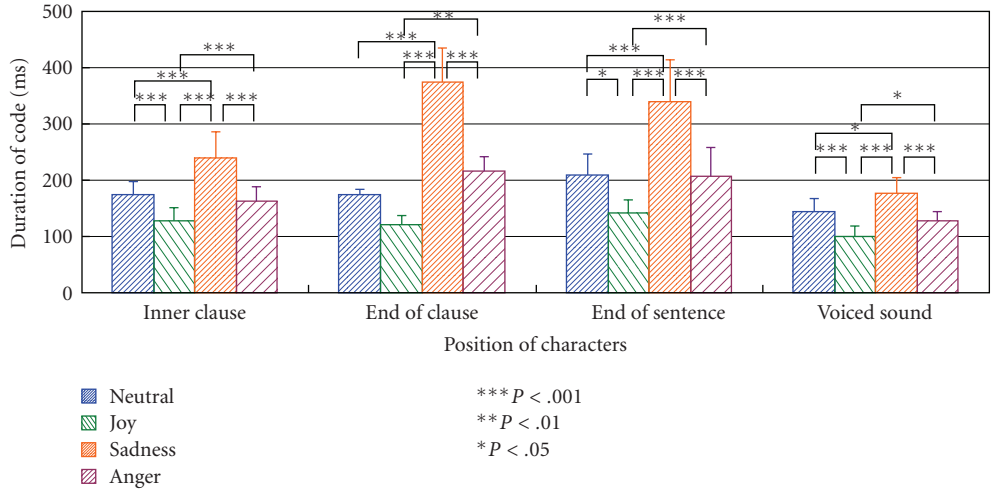


Figure 13: Mean of duration of code as a function of emotions and position of characters for two-handed Finger Braille (error bars represent standard deviations).

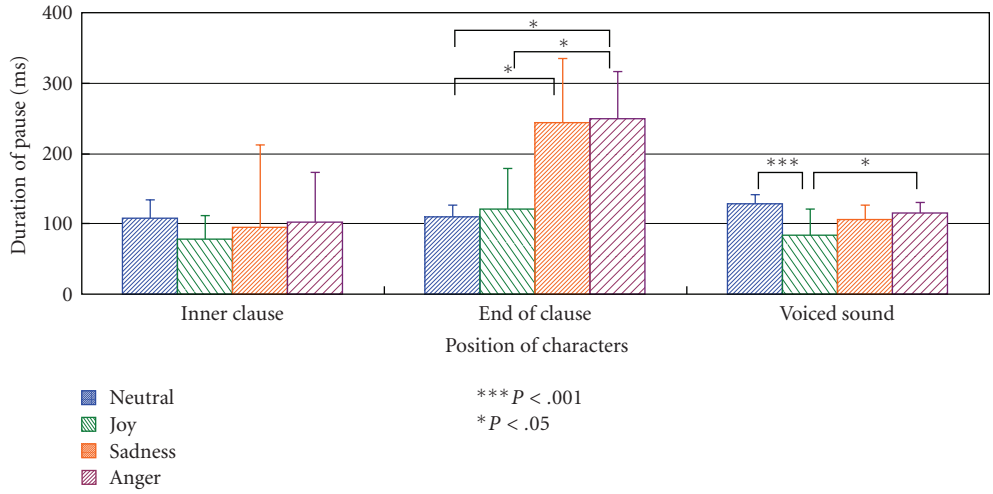


Figure 14: Mean of duration of pause as a function of emotions and position of characters for two-handed Finger Braille (error bars represent standard deviations).

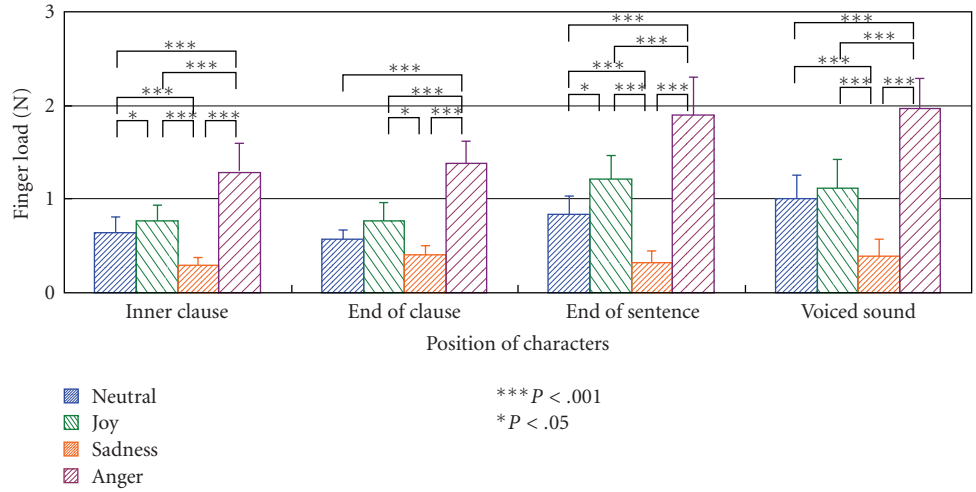


Figure 15: Mean of finger load as a function of emotions and position of characters for two-handed Finger Braille (error bars represent standard deviations).

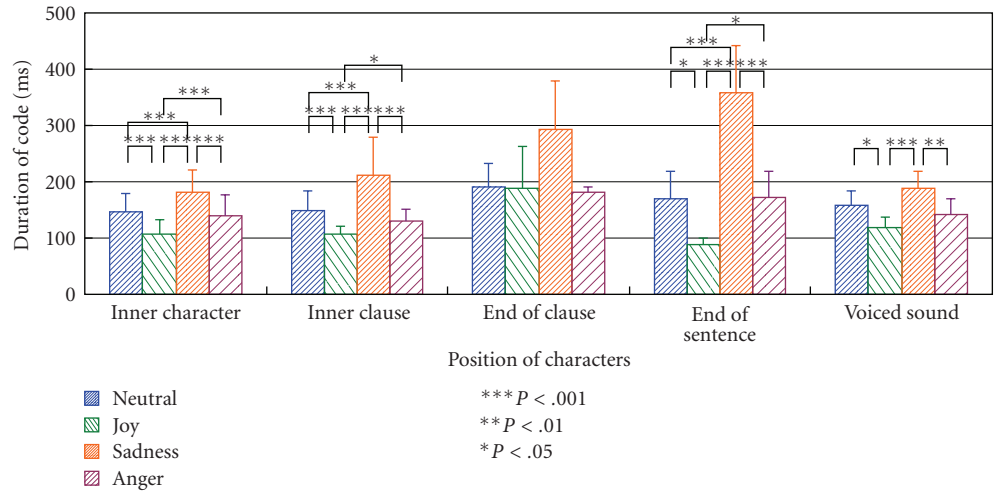


Figure 16: Mean of duration of code as a function of emotions and position of characters for one-handed Finger Braille (error bars represent standard deviations).

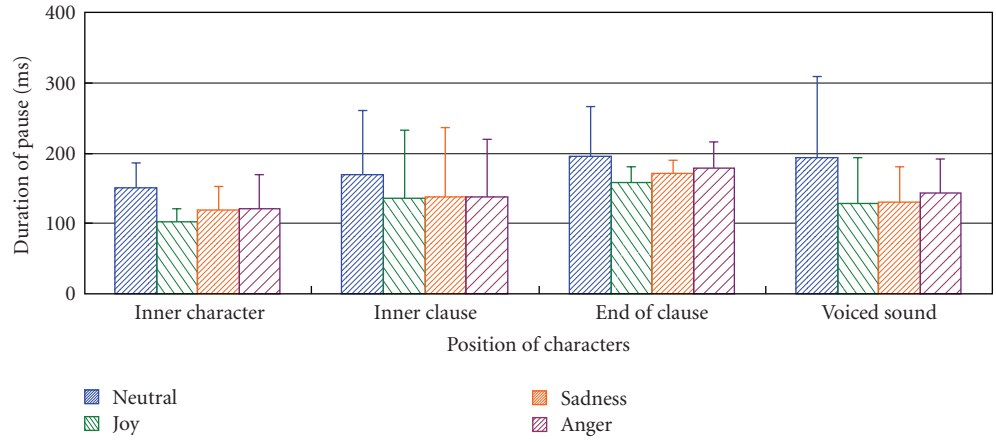


Figure 17: Mean of duration of pause as a function of emotions and position of characters for one-handed Finger Braille (error bars represent standard deviations).

(c) Finger Load. The mean of the finger load as a function of the emotions and position of characters is shown in Figure 18.

A 4 × 5 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,580) = 106.5$, $P < .001$) and the position of characters ($F(4,580) = 36.00$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(12,580) = 2.999$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the finger load of anger was significantly larger than the other finger loads in all positions of characters ($P < .001$); the finger load of joy was significantly larger than the finger loads of neutral and sadness in all positions of characters ($P < .017$); the finger load of sadness was significantly smaller than the other finger loads in all positions of characters ($P < .028$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .001$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the finger load of the end of sentence was significantly larger than the finger load of the inner clause in neutral, joy, and anger ($P < .001$); the finger load of the end of sentence was significantly larger than the other finger loads in sadness ($P < .013$).

3.3. Discussion. As indicated by the results of ANOVA (3.2.1–3.2.2), the features of emotional expression were as follows: (1) the duration of the code of joy was significantly shorter than that of the other emotions; (2) the duration of the code of sadness was significantly longer than that of the other emotions; (3) the finger load of anger was significantly larger than that of the other emotions; (4) the finger load of joy was significantly larger than that of sadness and neutral. These features were very similar to the results of Experiment I.

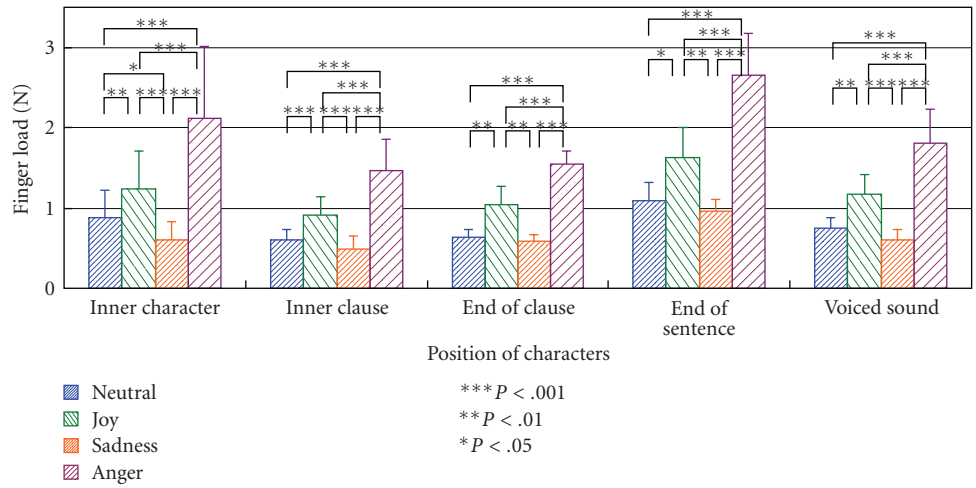


Figure 18: Mean of finger load as a function of emotions and position of characters for one-handed Finger Braille (error bars represent standard deviations).

As indicated by the results of ANOVA (3.2.1–3.2.2), the followings were the features of prosody: (1) the duration of the code of the end of the clause was significantly longer; (2) the duration of the pause of the end of the clause was significantly longer; (3) the duration of the code of the end of the sentence was significantly longer; (4) the finger load of the end of the sentence was significantly larger; (5) the duration of the code of the double consonant was significantly shorter; (6) the duration of the code of the voiced sound was significantly shorter; (7) the finger load of the voiced sound was significantly larger. These features were also very similar to the results of Experiment I.

In addition, there was a tendency for the finger load of the additional experiment to be larger than the finger load of Experiment I. This tendency could be a daily variance of the dotting strength of Finger Braille.

Therefore, we can confirm that the features of emotional expression and prosody are independent of the dotted dialogues.

4. Emotional Communication by the Unskilled (Experiment II)

To analyze the effectiveness of emotional expression and emotional communication between people who are unskilled in Finger Braille, we conducted another experiment of emotional communication using Finger Braille (Experiment II) [24].

4.1. Methods. The subjects were twelve non-disabled college students (ages: 21–26 years old) who were not unskilled in Finger Braille. All subjects were right-handed. All subjects gave informed consent after hearing a description of the study.

The dialogues were "Rain has fallen (*Ame futte kita* ⠠⠍⠑ ⠋⠥⠞⠞⠑ ⠅⠊⠞⠁)" and "It is the time (*Jikan desuyo* ⠚⠊⠅⠁⠝ ⠙⠑⠎⠥⠽⠕)", which were the same dialogues used in Experiment I.

Expressed emotions were neutral, joy, sadness, and anger, which were identical to those used in Experiment I. We set six scenes (two dialogues with joy, sadness, and anger) for subjects, which were also identical to the ones used in Experiment I.

In this experiment, we adopted one-handed Finger Braille (right hand), because one-handed Finger Braille is easier than two-handed Finger Braille for unskilled individuals.

The variables compose a 4 × 6 within-subjects design: expressed emotion (neutral, joy, sadness, anger) and position of characters (inner character, inner clause, end of clause, end of sentence, voiced sound, double consonant). The dependent variables were duration of code, duration of pause, and finger load.

The experimental flow is shown in Figure 19. The subjects were divided into four groups. The experiment was conducted for four days. On day 1, two practice sessions were conducted, followed by two intragroup experiment sessions. On days 2–4, three intergroup sessions were conducted. In the experiment session, a subject became a sender and another subject became a receiver. The experiment consisted of 132 sessions for all pairs of subjects: 11 pairs as sender and 11 pairs as receiver.

The sender and receiver sat side by side during the experiment to promote the emotional expression by the sender. To simulate deafblindness, the receiver wore earplugs, headphones playing white noise, and a night shade. The sender dotted one-handed Finger Braille (right-hand) on the fingers of the receiver eight times per session in a randomized order of dialogues with emotions. To eliminate the influence of the previous dotted dialogue with emotion, the subject dotted at intervals (1-2 minutes) and confirmed the next scene of dialogue with emotion. The sender observed a sheet of dot patterns of the experimental dialogues, and the receiver responded to the dialogue and emotion after the sender dotted one time. Unlike the interpreters in Experiment I, the subjects were not taught the features of emotional expression and prosody. The subjects expressed

emotions freely by changing the strength and speed of dotting.

The receiver put his fingers on the pressure sensor sheet (Grip Force Sensor System, Nitta), which measured the change of pressure as a result of dotting. The sampling frequency was 150 Hz, the measurement range was 20–200 kPa, and the sensibility was 781 Pa. The sensor sheet was divided into 20 blocks, and the size of each block was 16 × 16 mm. The sensor matrix of the block was 4 × 4. The receiver put his index, middle, and ring fingers on the three blocks.

4.2. Results

4.2.1. Results of Emotional Expression. We calculated the duration of code, duration of pause, and finger load for each dotting by the change of pressure, as in Experiment I (see Figure 4).

(a) Duration of Code. The mean of the duration of the code as a function of the emotions and position of characters is shown in Figure 20.

A 4 × 6 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,13689) = 675.6$, $P < .001$) and the position of characters ($F(5,13689) = 39.13$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(15,13689) = 5.153$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed the following: the duration of code of joy was significantly shorter than the other durations in all positions of characters ($P < .001$); the duration of code of anger was significantly longer than the other durations in the inner character, inner clause, end of sentence, voiced sound, and double consonant ($P < .001$, double consonant $P < .016$); the duration of code of anger was significantly longer than the duration of code of neutral in the end of clause ($P < .006$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .001$). Scheffe tests on the factor of the position of characters in each emotion revealed the following: the durations of code of the inner clause and end of clause were significantly longer than the durations of the voiced sound and double consonant in neutral and sadness ($P < .012$); the duration of code of the end of the sentence was significantly shorter than the durations of code of the inner clause, inner character, voiced sound, and end of clause in joy ($P < .017$); the duration of code of the voiced sound was significantly shorter than the durations of code of the inner clause, end of clause, and end of sentence in anger ($P < .016$).

(b) Duration of Pause. The mean of the duration of the pause as a function of the emotions and position of characters is shown in Figure 21. We defined the duration of pause before the double consonant as the duration of pause of the double consonant and the duration of pause after the double consonant as the duration of pause of the inner clause, as in the Experiment I. There was no duration of pause at the ends of sentences.

A 4 × 6 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,12607) = 43.926$, $P < .001$) and the position of characters ($F(5,12607) = 320.3$, $P < .001$). Scheffe tests on the emotion factor revealed that the duration of pause of sadness was significantly longer than the other durations ($P < .001$), and the duration of pause of anger was significantly shorter than the other durations ($P < .001$). Scheffe tests on the factor of the position of characters revealed the following: the duration of pause of the end of the clause was significantly longer than the other durations ($P < .001$); the duration of pause of the inner character was significantly shorter than the other durations ($P < .001$); the duration of pause of the inner clause was significantly longer than the durations of pause of the voiced sound, and double consonant ($P < .001$).

(c) Finger Load. The mean of the finger load as a function of the emotions and the position of characters is shown in Figure 22.

A 4 × 6 within-subjects analysis of variances (ANOVA) revealed two significant main effects: emotion ($F(3,13689) = 728.9$, $P < .001$) and the position of characters ($F(5,13689) = 378.2$, $P < .001$). There was a significant interaction term of the emotion and position of characters ($F(15,13689) = 25.29$, $P < .001$).

A test of the simple main effect of emotion revealed a significant effect of emotion in all positions of characters ($P < .001$). Scheffe tests on the emotion factor in each position of characters revealed that the finger load of anger was significantly larger than the other finger loads in all positions of characters ($P < .001$), and the finger load of sadness was significantly smaller than the other finger loads in all positions of characters ($P < .002$).

A test of the simple main effect of the position of characters revealed a significant effect of the position of characters in all emotions ($P < .034$). Scheffe tests on the factor of the position of characters in each emotion revealed that the finger loads of the inner character and double consonant were significantly larger than the finger loads of the end of clause, end of sentence, inner clause and voiced sound in all emotions ($P < .001$).

4.2.2. Results of Emotional Communication. In this experiment, the receiver answered the recognized dialogue and emotion dotted by the sender after each dotting. As a result, the accuracy of recognition of the dialogues was a total of 96.5%. The relationship between expressed emotions by the senders, recognized emotions by the receivers, and their coincidence ratios is presented in Table 6. The coincidence ratio of anger was the highest (84.5%), and the coincidence ratio of sadness was second (78.4%). If the coincidence ratio was high, the emotions expressed by the sender were communicated well to the receiver. If the coincidence ratio was low, the emotions expressed by the sender were communicated poorly to the receiver. The coincidence ratios

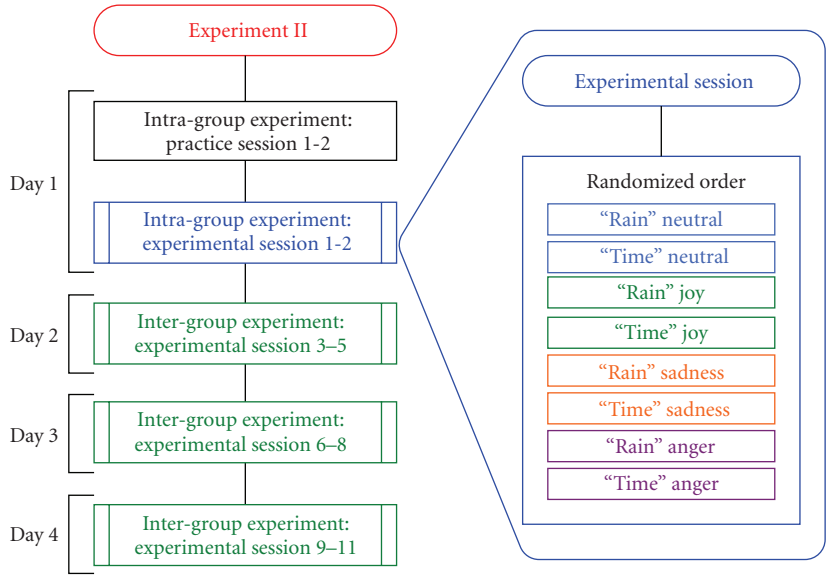


Figure 19: Flow of Experiment II.

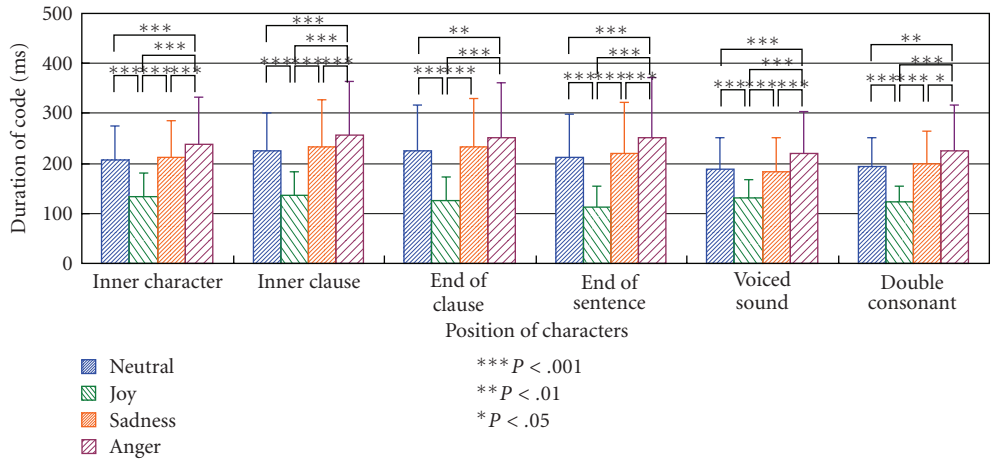


Figure 20: Mean of duration of code as a function of emotions and position of characters (error bars represent standard deviations).

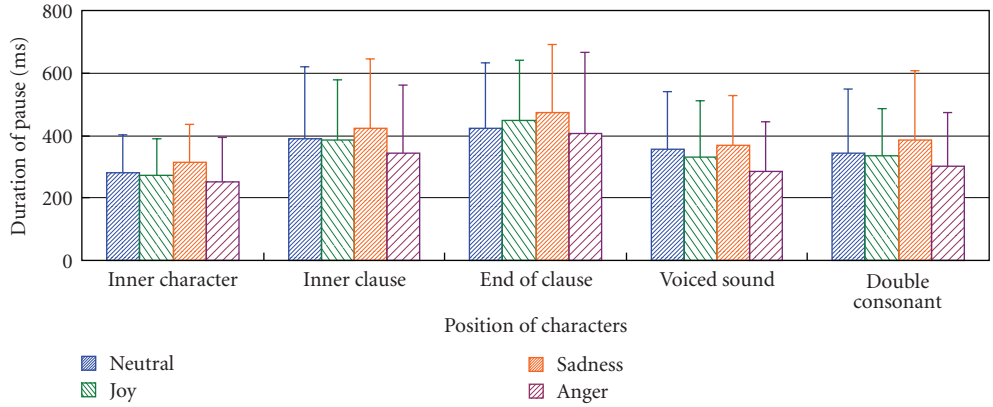


Figure 21: Mean of duration of pause as a function of emotions and position of characters (error bars represent standard deviations).

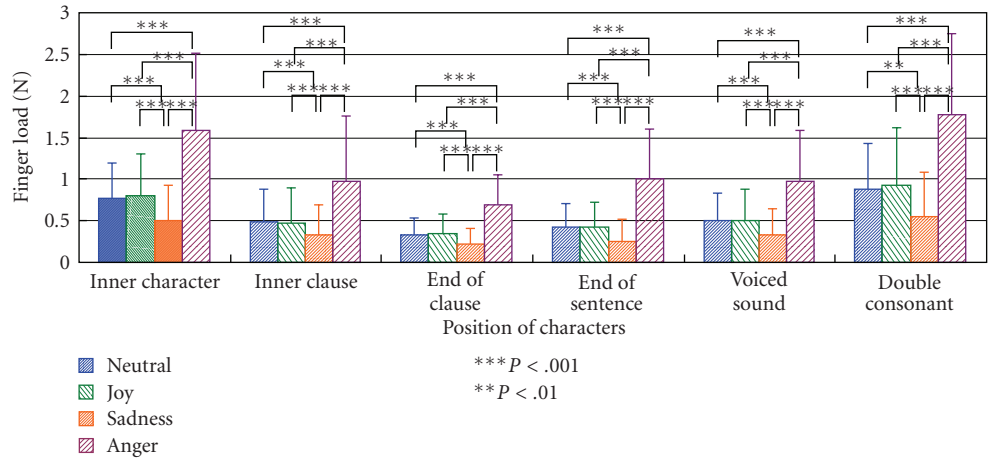


Figure 22: Mean of finger load as a function of emotions and position of characters (error bars represent standard deviations).

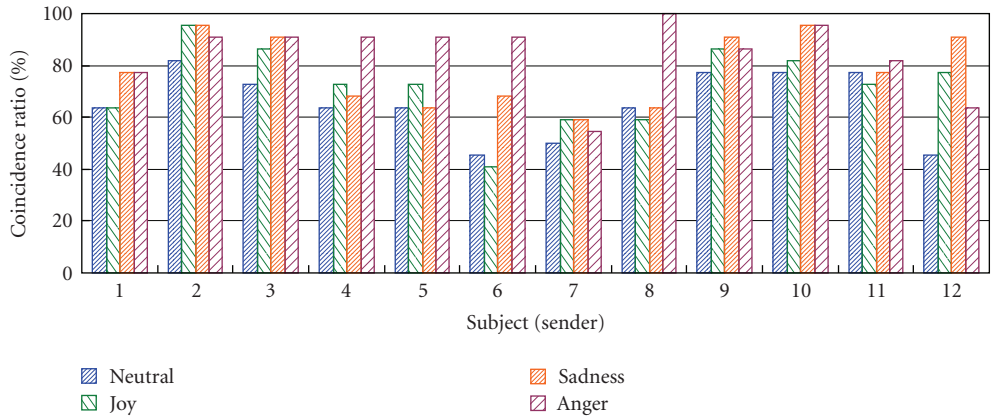


Figure 23: Coincidence ratio of emotion as a function of emotions and subject (sender).

as a function of the emotions and subject (sender) are shown in Figure 23. Subject 2 was the best communicative sender (coincidence ratio 90.9%); Subject 7 was the worst communicative sender (coincidence ratio 55.7%).

The change of coincidence ratios as a function of the emotions and order of dotted dialogue in a session are shown in Figure 24. In dotted dialogues 1 and 2, the coincidence ratios of the four emotions were almost 60%. In dotted dialogues 3–6, the coincidence ratios of anger and sadness were 80%–90%. The coincidence ratio of joy was almost 70%, and the coincidence ratio of neutral was almost 60%. In dotted dialogues 7 and 8, the coincidence ratios of the four emotions were 80% or higher.

4.2.3. Questionnaire Results (Unskilled Subjects). After the experiment, the subjects were asked how they expressed these emotions, and their responses are listed in Table 7. These responses were similar to the ones given by the interpreters in Experiment I.

Six subjects answered that they dotted Finger Braille while sometimes glancing at the sheet of dot patterns. The other six subjects answered that they dotted Finger Braille while continually observing the sheet of dot patterns. As a degree of emotional expression, one subject answered "well expressed"; six subjects answered "almost expressed"; five subjects answered "somewhat expressed."

All subjects recognized the dotted dialogues by the first two characters or the last two characters. Some subjects also answered that it was hard to recognize the expressed emotions if the first emotion in the session was neutral, and two or three dialogues in the session were necessary to recognize the neutral level of the sender.

Table 6: Coincidence ratios of emotion.

Expressed emotion	Recognized emotion				Coincidence ratio
	Neutral	Joy	Sadness	Anger	
Neutral	172	31	44	17	65.2%
Joy	38	191	8	27	72.3%
Sadness	43	11	207	3	78.4%
Anger	21	14	6	223	84.5%
Total	274	247	265	270	75.1%

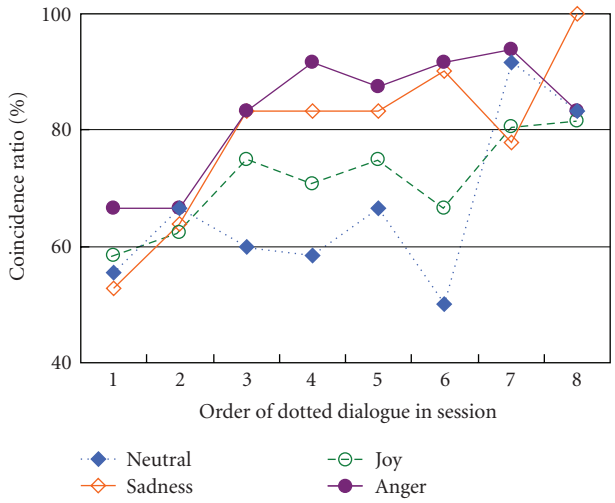


Figure 24: Change of coincidence ratio as a function of emotions and order of dotted dialogue in session.

4.3. Discussion

4.3.1. Comparison of Emotional Expression between Interpreters and the Unskilled. There were similar features of the emotional expression between the interpreters in Experiment I and the unskilled subjects in Experiment II: the duration of joy was significantly shorter than that of the other emotions; the finger load of anger was significantly larger than that of the other emotions. As different features, the duration of the code of anger by the interpreters was slightly shorter, but the duration of the code of anger by the unskilled subjects was significantly longer. In a previous study, we found that the unskilled subjects faced difficulties in dotting strongly in a short time [25]. Therefore, the duration of code of strong dotting by the unskilled subjects had a tendency to be long.

Since the unskilled subjects were not taught the prosody of Finger Braille, they could not express its metrical structure. The unskilled subjects dotted Finger Braille by glancing at or continually observing the sheet of dot patterns, and then many of them paused at the end of a clause and checked the patterns.

According to the questionnaire results, the impressions of emotional expression by the unskilled subjects were similar to those by the interpreters.

4.3.2. Discriminant Analysis of Emotional Expression. To evaluate emotional expression by the subjects in Experiment II, twelve discriminant analyses on the data of each subject were conducted (N = 12,627 total). Discriminant variables were the duration of the code, duration of the pause, and the finger load. The dotting of the end of the sentence was excluded, because there was no duration of pause.

The total discriminant ratios of emotion are listed in Table 8. The total discriminant ratio was 60.6%. The discriminant ratio of joy was the highest (80.9%), and the discriminant ratio of anger was second (63.6%). The dottings of neutral and sadness were frequently misdiscriminated between each other. The discriminant ratio as a function

Table 7: Responses of unskilled senders to questionnaire on expressing emotions.

Emotion	Subject	Answer
Neutral	1	As the standard of the other emotions
	2	Constant tempo, medium duration
	3	Mechanically, medium strength, medium speed
	4	Constant rhythm
	5	Constant speed, medium strength
	6	Somewhat weaker than angry
	7	Normally
	8	Not one of the other emotions
	9	Characterless, only considered dotting a sentence
	10	Flat rhythm, medium strength, somewhat long
	11	Somewhat slowly, medium strength
	12	Carefully, slowly
Joy	1	Quick tempo, airily
	2	Rhythmically, short duration, airily
	3	Rhythmically, medium strength
	4	Rhythmically, quick tempo
	5	Quickly, airily
	6	Quickly
	7	Rhythmically
	8	Weakly, rhythmically
	9	Short duration, somewhat quickly
	10	Quick rhythm, medium strength
	11	Quicker tempo than neutral, like flicking fingers
	12	Airily, quickly
Sadness	1	Considered to avoid recognition as neutral
	2	Slow tempo, long duration, softly
	3	Weakly, slowly
	4	Slowly, airily
	5	Slowly, weakly
	6	Weakly, slowly
	7	Slowly, airily
	8	Weakly, slowly
	9	Long duration, weakly, slowly
	10	Slow rhythm, quite weakly
	11	Quite weakly, airily
	12	Softly
Anger	1	Strongly, heavily
	2	Somewhat quickly, medium duration, strongly
	3	Strongly, quickly
	4	Strongly, slowly
	5	Strongly
	6	Strongly
	7	Strongly, somewhat slowly
	8	Strongly, like pushing fingers
	9	Same duration and speed as neutral, strongly
	10	Rhythm between joy and sadness, quite strongly
	11	Quite strongly, like pushing fingers
	12	Claw-like fingers

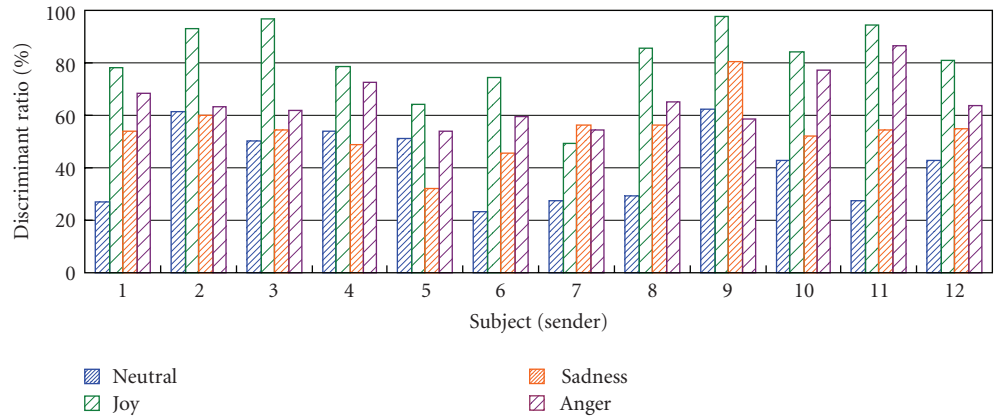


Figure 25: Discriminant ratio as a function of emotions and subject (sender).

Table 8: Total discriminant ratios of emotion.

Expressed emotion	Discriminated emotion				Discriminant ratio
	Neutral	Joy	Sadness	Anger	
Neutral	1,356	601	763	436	43.0%
Joy	174	2,551	220	207	80.9%
Sadness	678	571	1,727	182	54.7%
Anger	500	368	278	2,015	63.7%
Total	2,708	4,091	2,988	2,840	60.6%

of the emotions and subject is presented in Figure 25. If the discriminant ratio is high, the subject expressed emotions well; if the discriminant ratio is low, the subject expressed emotions poorly. Subject 9 was the best expressed subject (discriminant ratio 74.9%); Subject 7 was the worst expressed subject (discriminant ratio 47.0%).

The represented variables for discriminant functions, contribution ratios, and significance probability are listed in Table 9. For Subjects 2, 4, 8, 9, and 11, discriminant function 1 represented the duration of the code, discriminant function 2 represented the finger load, and discriminant function 3 represented the duration of the pause (group "CLP"). For Subjects 1, 3, 5, 7, 10, and 12, discriminant function 1 represented the finger load, discriminant function 2 represented the duration of the code, and discriminant function 3 represented the duration of the pause (group "LCP"). For Subject 6, discriminant function 1 represented the duration of the code, discriminant function 2 represented the duration of the pause, and discriminant function 3 represented the finger load (group "CPL"). For Subjects 2, 4, 9, and 12, discriminant functions 1, 2, and 3 contributed to emotional expression. For Subjects 1, 3, 5, 6, 7, 8, 10, and 11, discriminant functions 1 and 2 contributed to emotional expression. Subjects of the group CLP and LCP could control both the duration of the code and the finger load, and subjects of the group CLP could control the duration of the code well.

The mean of the discriminant ratio of group CLP was 66.5% (S.D. = 6.0), and the mean of the discriminant ratio of group LCP and group CPL was 56.4% (S.D. = 7.3). A t-test revealed a significant difference between group CLP and groups LCP and CPL ($t(10) = 2.54$, $P < .030$). To discuss the level of learning of emotional expression using Finger Braille, we calculated the discriminant ratio of experimental sessions 1–4 and 5–11. As for the experimental sessions 1–4, the mean of the discriminant ratio of group CLP was 59.0% (S.D. = 11.6), and the mean of the discriminant ratio of group LCP and group CPL was 49.3% (S.D. = 10.3). A t-test revealed no significant difference between group CLP and groups LCP and CPL ($t(10) = 1.55$, $P < .153$). As for the experimental sessions 5–11, the mean of the discriminant ratio of group CLP was 70.7% (S.D. = 4.9), and the mean of the discriminant ratio of group LCP and group CPL was 60.4% (S.D. = 7.3). A t-test revealed a significant difference between group CLP and groups LCP and CPL ($t(10) = 2.73$, $P < .021$). These results

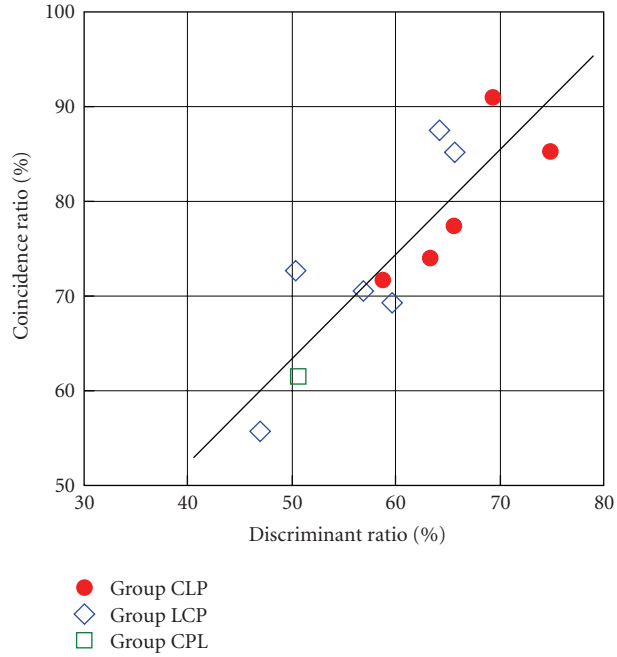


Figure 26: Relationship between discriminant ratio and coincidence ratio of each subject.

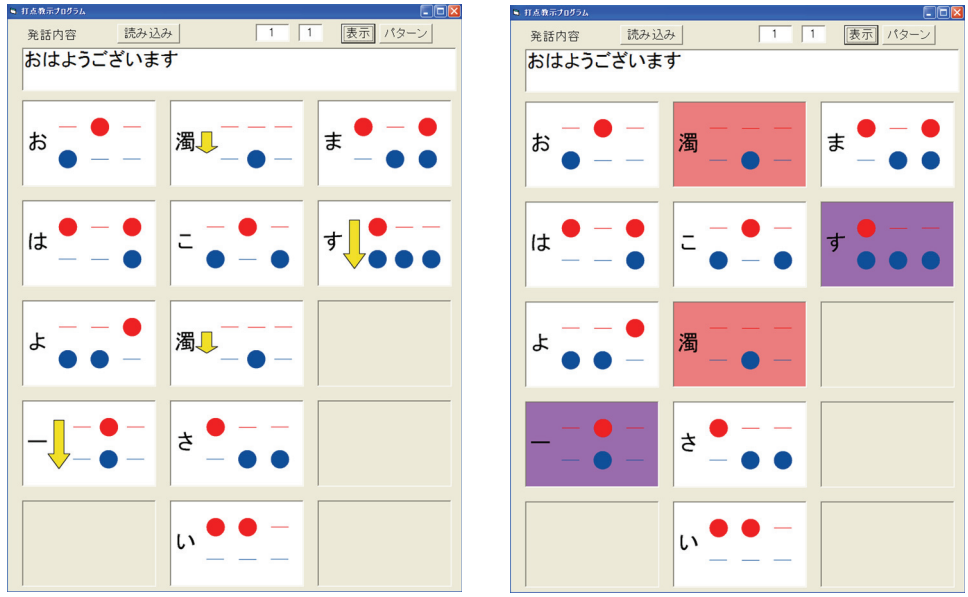


Figure 27: Examples of patterns for teaching unskilled people how to express emotions using Finger Braille.

Table 9: Represented variables for discriminant functions, contribution ratios, and significance probability (***$P <$ **.001**).

Group	Subject	Discriminant function 1	Discriminant function 2	Discriminant function 3
		Duration of code	Finger load	Duration of pause
CLP	2	53.9%***	44.6%***	1.5%***
	4	84.3%***	14.7%***	1.0%***
	8	62.1%***	37.9%***	0%
	9	75.4%***	21.2%***	3.3%***
	11	90.8%***	9.1%***	0.1%
		Finger load	Duration of code	Duration of pause
LCP	1	78.2%***	21.8%***	0%
	3	75.7%***	24.1%***	0.1%
	5	64.2%***	35.6%***	0.2%
	7	78.3%***	21.5%***	0.1%
	10	81.8%***	18.2%***	0%
	12	65.3%***	29.2%***	5.5%***
		Duration of code	Duration of pause	Finger load
CPL	6	83.3%***	16.4%***	0.3%

indicate the following: the discriminant ratio of group CLP was significantly higher in experimental sessions 5–11; controlling the dotting strength (finger load) was easier for the unskilled people; in addition, controlling the dotting speed (duration of code) was a concern about expressing emotions well using Finger Braille. Group CLP had learned the emotional expression well during experimental sessions 1–4. Although we had not checked the dexterity of subjects concerning dotting Finger Braille (e.g., playing piano or typing on a keyboard) in the present study, we surmise that group CLP was more dexterous than groups LCP and CPL.

The relationship between the discriminant ratio and coincidence ratio of each subject (sender) is shown in Figure 26. As a result of regression analysis ($N = 12$), the correlation coefficient R was 0.862 ($P < .001$). Therefore, if the sender could express emotions well, the expressed emotions could be recognized accurately by the receiver.

4.3.3. Features of Emotional Communication. As shown in Figure 24, the receivers could recognize anger and sadness in dotted dialogue 3 and neutral and joy in dotted dialogue 7. These results indicate that the receivers first recognized the strength of dotting, and after the receivers better comprehended the level of neutral by the senders, they could recognize the duration of dotting.

Now, we discuss the features of emotional communication by other communication media.

Shirasawa et al., Iida et al., and Kinjou et al. analyzed emotional communication by speech [16–18].

In Shirasawa's experiment [16], one speaker, who had experience as an actor and announcer, recorded eight phrases with neutral, joy, sadness, anger, surprise, and disgust. Then, 55 subjects heard the phrases and identified the emotions. The coincidence ratio was 51.5%. The coincidence ratio of sadness was the highest (64%).

In Iida's experiment [17], two speakers recorded four phrases with neutral, joy, sadness, anger, surprise and disgust. Then, seven subjects heard the phrases and identified the emotions. The coincidence ratio was 50.0%. The coincidence ratio of sadness was the highest (91.1%).

In Kinjou's experiment [18], six speakers, of whom three had experience as actors, recorded 24 words with neutral, joy, sadness, and anger. Then, 55 subjects heard the words and identified the emotions. The coincidence ratio was 77.1%. The coincidence ratio of neutral was the highest (83.3%).

Yoshikawa and Morisaki analyzed emotional communication by facial expression [19]. One subject expressed emotions by facial expression and another subject identified the emotions. Two situations with joy, sadness, anger, surprise, disgust, fear, and annoyance were prepared. The subjects were nine acquainted pairs and nine unacquainted pairs. The coincidence ratio was 61.6% for the acquainted pairs. The coincidence ratio of joy was the highest (97%) for the acquainted pairs. The coincidence ratio was 58.8% for the unacquainted pairs. The coincidence ratio of joy was the highest (93%) for the unacquainted pairs.

Ohgushi and Hattori analyzed emotional communication by vocal performance [20]. Three female singers sang the vowel sound "a" with joy, sadness, anger, fear, and neutral emotion. Fifty-one subjects heard the vocal performance and identified the emotions. The coincidence ratio was 52.8%. The highest coincidence ratio was anger (66%).

The results of the present study showed that the coincidence ratios of Experiment II were equal to or more than the coincidence ratios of these previous studies. Thus, it was considered that the unskilled people can express and communicate emotions using Finger Braille, which is a new tactile communication medium.

However, some subjects expressed emotions poorly, and it was difficult for the receivers to recognize the dotted emotions. Thus, some assistance in emotional communication was needed for the non-disabled people. One of the methods to teach emotional expression to unskilled people is to teach the impression of emotional expression by applying the similarities of the impression of emotional expression between the interpreters and the unskilled people (e.g., "rhythmically" for joy, "weakly and slowly" for sadness, "strongly and quickly" for anger). Another method of teaching emotional expression is to introduce a dot pattern with some symbols directly representing dotting speed and strength (e.g., a dot pattern with a long or short array for dotting speed, and a dot pattern with a colored background for dotting strength). Some example patterns for teaching emotional expression to people unskilled in Finger Braille are shown in Figure 27. We are planning to develop and implement these methods of teaching emotional expression in future studies.

5. Conclusions

In the present study, emotional expression by interpreters using Finger Braille (Experiment I) was first examined, and the features of emotional expression using Finger Braille were analyzed. The resulting features were as follows: (1) the duration of the code of joy was significantly shorter than that of the other emotions (neutral, anger, sadness); (2) the duration of the code of sadness was significantly longer than that of the other emotions; (3) the finger load of anger was significantly larger than that of the other emotions; (4) the finger load of joy was significantly larger than that of sadness and neutral; (5) the duration of the code of anger was significantly shorter than that of sadness and neutral. As shown by the results of the additional experiment, these features of emotional expression are independent of the dotted dialogue.

Next, emotional communication by people unskilled in Finger Braille (Experiment II) was examined, and the effectiveness of emotional expression and emotional communication between these individuals was analyzed. The results indicate that the features and the impression of emotional expression by the unskilled people were very similar to those by the interpreters, and the coincidence ratio of emotional communication was 75.1%. Therefore, it was confirmed that unskilled people can express and communicate emotions using Finger Braille.

The followings are plans for future studies: (1) expansion of the emotions expressed and the number of subjects used in Finger Braille experiments, (2) development of methods to teach non-disabled people emotional expression using Finger Braille, (3) analysis of the relationship between emotional expression using Finger Braille and speech, and (4) development of an emotion recognition algorithm.

Acknowledgments

The authors greatly thank Ms. Satoko Mishina and Ms. Megumi Fukuma (interpreters of Finger Braille) for their support. This study was supported by the Japan Society for the Promotion of Science under a Grant-in-Aid for Scientific Research (no. 21500522) and the Ministry of Education, Culture, Sports, Science and Technology of Japan under a Grant-in-Aid for Scientific Research (no. 16700430). This study was partly supported by Kanagawa Academy of Science and Technology (KAST) under a research grant.

References

[1] S. Fukushima, *Person with Deafblind and Normalization*, Akashi Shoten, 1997.

[2] S. S. An, J. W. Jeon, S. Lee, H. Choi, and H.-G. Choi, "A pair of wireless braille-based chording gloves," in *Proceedings of the 9th International Conference on Computers Helping People with Special Needs*, vol. 3118 of *Lecture Notes in Computer Science*, pp. 490–497, 2004.

[3] T. Amemiya, K. Hirota, and M. Hirose, "OBOE: Oboe-like Braille interface for outdoor environment," in *Proceedings of the 9th International Conference on Computers Helping People with Special Needs*, vol. 3118 of *Lecture Notes in Computer Science*, pp. 498–505, 2004.

[4] Y. Matsuda, T. Isomura, I. Sakuma, E. Kobayashi, Y. Jimbo, and T. Arafune, "Finger Braille teaching system for people who communicate with deafblind people," in *Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA '07)*, pp. 3202–3207, Harbin, China, August 2007.

[5] Y. Matsuda, I. Sakuma, Y. Jimbo, E. Kobayashi, T. Arafune, and T. Isomura, "Development of Finger Braille teaching system—teaching of dotting finger and position using speech recognition," *Journal of the Society of Life Support Technology*, vol. 19, no. 3, pp. 105–116, 2007.

[6] Y. Matsuda, I. Sakuma, Y. Jimbo, E. Kobayashi, T. Arafune, and T. Isomura, "Finger Braille recognition system for people who communicate with deafblind people," in *Proceedings of IEEE International Conference on Mechatronics and Automation (ICMA '08)*, pp. 268–273, Takamatsu, Japan, August 2008, WE2-2.

[7] Y. Matsuda, T. Isomura, I. Sakuma, Y. Jimbo, E. Kobayashi, and T. Arafune, "Emotion recognition of Finger Braille," in *Proceedings of the 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '08)*, pp. 1409–1411, Harbin, China, August 2008.

[8] Y. Matsuda, I. Sakuma, Y. Jimbo, E. Kobayashi, T. Arafune, and T. Isomura, "Analysis of emotional expression of Finger Braille," in *Proceedings of the 7th Asian-Pacific Conference on Medical and Biological Engineering*, vol. 19 of *IFMBE Proceedings*, pp. 484–487, 2008.

[9] Y. Matsuda, I. Sakuma, Y. Jimbo, E. Kobayashi, T. Arafune, and T. Isomura, "Study on emotional communication in skin contact—emotional expression experiment by Finger Braille interpreters," *Transaction of Human Interface Society*, vol. 10, no. 4, pp. 417–426, 2008.

[10] R. Plutchik, *Emotions and Life*, American Psychological Association, Washington, DC, USA, 2002.

[11] K. M. B. Bridges, "Emotional development in early infancy," *Child Development*, vol. 3, pp. 324–341, 1932.

[12] P. Laukka, *Vocal expression of emotion*, Ph.D. thesis, Faculty of Social Sciences, Uppsala University Library, 2004.

[13] R. Bresin and A. Friberg, "Emotional coloring of computer-controlled music performances," *Computer Music Journal*, vol. 24, no. 4, pp. 44–63, 2000.

[14] S. Dahl and A. Friberg, "Expressiveness of musician's body movements in performances on marimba," in *Proceedings of the 5th International Gesture Workshop (GW '04)*, pp. 479–486, Genova, Italy, April 2004.

[15] M. Clynes and J. Panksepp, *Emotions and Psychopathology*, Plenum Press, New York, NY, USA, 1988.

[16] T. Shirasawa, Y. Kato, and N. Ohnishi, "Analysis of kansei information perceived from speech," *Technical Report of IEICE*, vol. 96, no. 115, pp. 47–52, 1996.

[17] A. Iida, S. Iga, and M. Yasumura, "Study of emotion in speech: fnidings from perceptual experiments," *Information Processing Society of Japan*, vol. 97, no. 16, pp. 113–118, 1997.

[18] Y. Kinjou, Y. Tsuchimoto, and I. Nagayama, "Feeling recognition of spoken words for advanced communication with emotional information processing," *Technical Report of IEICE*, vol. 101, no. 594, pp. 49–54, 2002.

[19] S. Yoshikawa and A. Morisaki, "Sending emotional message by facial expressions," *Technical Report of IEICE*, vol. 98, no. 503, pp. 31–38, 1999.

[20] K. Ohgushi and M. Hattori, "Emotional expression in vocal music performances and transmission to the audience," in *Proceedings of the Autumn Meeting of the Acoustical Society of Japan*, pp. 219–222, 2000.

[21] T. Dousaka, M. Aoki, H. Fukasawa, and Y. Nagashima, "Analysis of prosodic features in Finger Braille," *Technical Report of IEICE*, vol. 99, no. 2, pp. 5–8, 2000.

[22] Y. Fujimori, M. Miyagi, K. Ikegami, Y. Horiuchi, and A. Ichikawa, "Time structure analysis and investigation of prosody rules of Finger Braille," *Technical Report of IEICE*, no. 1, pp. 29–35, 2000.

[23] M. Miyagi, K. Miyazawa, A. Ueno, et al., "Analysis of prosody in strength and time structure of Finger Braile," *Technical Report of IEICE*, vol. 107, no. 61, pp. 25–28, 2007.

[24] Y. Matsuda and T. Isomura, "Study on emotional communication in skin contact—emotional communication experiment in Finger Braille," *Transaction of Human Interface Society*, vol. 5, no. 2, pp. 163–170, 2003.

[25] Y. Matsuda, I. Sakuma, Y. Jimbo, E. Kobayashi, T. Arafune, and T. Isomura, "Study on teaching of the way to dot of Finger Braille—teaching of dotting finger and position of monosyllable," *Transaction of Human Interface Society*, vol. 7, no. 3, pp. 379–390, 2005.