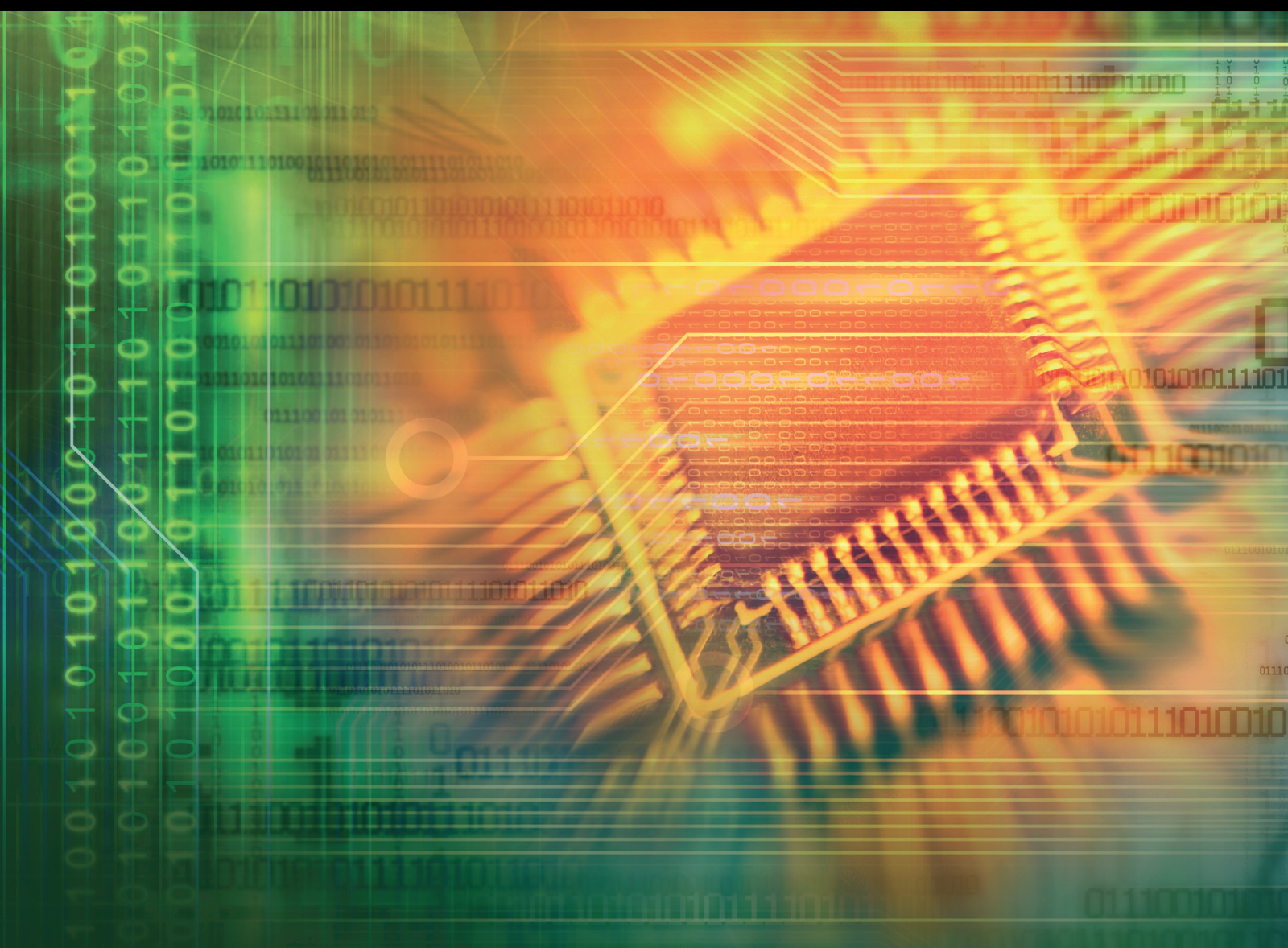
Journal of Electrical and Computer Engineering

Signal Processing Platforms and Algorithms for Real-life Communications and Listening to Digital Audio

Lead Guest Editor: Alexander Petrovsky
Guest Editors: Wanggen Wan, Manuel R. Zurera, and Alexey Karpov



Signal Processing Platforms and Algorithms for Real-life Communications and Listening to Digital Audio

Journal of Electrical and Computer Engineering

Signal Processing Platforms and Algorithms for Real-life Communications and Listening to Digital Audio

Lead Guest Editor: Alexander Petrovsky
Guest Editors: Wanggen Wan, Manuel R. Zurera, and Alexey Karpov

Editorial Board

Contents

Hindawi
Journal of Electrical and Computer Engineering
Volume 2017, Article ID 2913236, 2 pages
https://doi.org/10.1155/2017/2913236

Editorial

Signal Processing Platforms and Algorithms for Real-Life Communications and Listening to Digital Audio

Alexander Petrovsky,[1] Wanggen Wan,[2] Manuel Rosa-Zurera,[3] and Alexey Karpov[4]

[1]*Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus*
[2]*Shanghai University, Shanghai, China*
[3]*University of Alcalá, Madrid, Spain*
[4]*St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia*

Correspondence should be addressed to Alexander Petrovsky; palex@bsuir.by

Received 2 April 2017; Accepted 2 April 2017; Published 19 July 2017

Design of modern electronic communication systems involves diversified scientific areas including algorithms, architectures, and hardware development. Variety of existent multimedia devices gives rise to development of platform-dependent signal processing algorithms. Their integration into existent digital environment is an urgent problem for application engineers. Considering a wide range of applications including hearing aids, real-life communications, and listening to digital audio, the following research areas are of particular importance: advanced time-frequency representations, audio user interfaces, audio and speech enhancement, assisted listening, and perception and phonation modeling.

This special issue aims at publishing papers presenting novel methodologies and techniques (including theoretical methods, algorithms, and software) correspondent to these research areas. It includes high-quality papers dealing with applications in speech recognition, emotion recognition in speech signals, or informed source separation in orchestral recordings. The problems addressed in the accepted papers are among the top trends in the signal processing and communications research community.

Human interaction with computers through voice-user interfaces is the issue one of the papers deals with. It is based on Automatic Speech Recognition, and nowadays the objective is to solve the problem of spontaneous speech recognition. Spontaneous speech is characterized by hesitations, disfluencies, and changes that convey information about the speaker. The paper "Experiments on Detection of Voiced Hesitations in Russian Spontaneous Speech," by V. Verkhodanova and V. Shapranov, addresses the issue of voiced hesitations (filled pauses and sound lengthenings) detection in Russian spontaneous speech by utilizing different machine learning techniques: from grid search and gradient descent in rule-based approaches to such data-driven ones as ELM and SVM based on the automatically extracted acoustic features. Experimental results on the mixed and quality diverse corpus of spontaneous Russian speech indicate the efficiency of the techniques for the task in question, with SVM outperforming other methods.

The need for understanding business trends, ensuring public security, and improving the quality of customer service has caused a sustained development of speech analytics systems which transform speech data into a measurable and searchable index of words, phrases, and paralinguistic markers. Keyword spotting technology makes a substantial part of such systems. In the article entitled "A Russian Keyword Spotting System Based on Large Vocabulary Continuous Speech Recognition and Linguistic Knowledge" by V. Smirnov et al., the authors present an automatic system for keyword spotting in continuous speech. This system uses high-level linguistic knowledge and models for Russian speech and language; it has been implemented as software and applied in real-life telecommunication tasks for continuous speech processing.

In recent years, more attention is paid to the study of emotion recognition. Speech, as one of the most important ways of communication in human daily life, contains rich emotional information. Speech emotion recognition because of its wide application significance and research value in intelligence and naturalness of human-computer interaction aspects has got more and more attention from the researchers in recent years. The authors Z. Cairong et al. of the article "A Novel DBN Feature Fusion Model for Cross-Corpus Speech Emotion Recognition" present an automatic system for speaker emotion recognition by speech analysis. This system uses Deep Belief Nets for feature fusion and selection; it has been studied in cross-corpus experiments for emotion recognition tasks using emotional Chinese and German speech databases.

Speech intelligibility and speech recognition are important and trending topics of research in various fields of science: Linguistics, Medicine, Electrical Engineering, and Information Technology. Speech recognition process is investigated from different sides, as only an integrated approach could lead to a better understanding of this process. One of these research areas is intelligibility improvement of synthesized speech in noise, which has demanded much attention in recent years. In the article entitled "Crosslinguistic Intelligibility of Russian and German Speech in Noisy Environment" by R. Potapova and M. Grigorieva, the authors present quantitative results of experimental research on speech perception and intelligibility of spoken utterances and words by different listeners in various noisy conditions. Multiple experiments have been made using Russian and German speech data with addition of white and pink acoustic noises with different intensity and signal-to-noise ratios.

Audio source-separation is also a challenging task when we have sources corresponding to different instrument sections, which are strongly correlated in time and frequency. Without any previous knowledge, it is difficult to separate two sections which play, for instance, consonant notes simultaneously. One way to tackle this problem is to introduce into the separation framework information about the characteristics of the signals, such as a well-aligned score. The paper entitled "Score-Informed Source Separation for Multichannel Orchestral Recordings" by M. Miron et al. proposes and evaluates a system for score-informed audio source separation for multichannel orchestral recordings. The given article aims at adapting and extending score-informed audio source separation to the inherent complexity of orchestral music. This scenario involves not only challenges, like changes in dynamics and tempo, a large variety of instruments, high reverberance, and simultaneous melodic lines, but also opportunities as multichannel recordings. Results show that it is possible to align the original score with the audio of the performance and separate the sources corresponding to the instrument sections. In addition, authors derive applications, which allow for multiperspective audio enhancement, as acoustic scene rendering, and integrate them into an online repository which allows the distribution of the generated audio content.

Acknowledgments

We thank all the authors who made submissions to this special issue and the reviewers for their support and detailed reviews in making this special issue possible.

Alexander Petrovsky
Wanggen Wan
Manuel Rosa-Zurera
Alexey Karpov

Hindawi
Journal of Electrical and Computer Engineering
Volume 2017, Article ID 1831856, 9 pages
https://doi.org/10.1155/2017/1831856

Research Article

Crosslinguistic Intelligibility of Russian and German Speech in Noisy Environment

Rodmonga Potapova and Maria Grigorieva

Institute of Applied and Mathematical Linguistics, Moscow State Linguistic University, Ostozhenka 38, Moscow 119034, Russia

Correspondence should be addressed to Maria Grigorieva; ivanovam112@mail.ru

Received 19 June 2016; Revised 26 January 2017; Accepted 26 February 2017; Published 13 July 2017

Academic Editor: Alexey Karpov

This paper discusses the results of the pilot experimental research dedicated to speech recognition and perception of the semantic content of the utterances in noisy environment. The experiment included perceptual-auditory analysis of words and phrases in Russian and German (in comparison) in the same noisy environment: various (pink and white) types of noise with various levels of signal-to-noise ratio. The statistical analysis showed that intelligibility and perception of the speech in noisy environment are influenced not only by noise type and its signal-to-noise ratio, but also by some linguistic and extralinguistic factors, such as the existing redundancy of a particular language at various levels of linguistic structure, changes in the acoustic characteristics of the speaker while switching from one language to another one, the level of speaker and listener's proficiency in a specific language, and acoustic characteristics of the speaker's voice.

1. Introduction

Speech intelligibility and speech recognition are important and trending topics of research in various fields of science: Linguistics, Medicine, Electrical Engineering, and Information Technology. Speech recognition process is investigated from different sides, as only an integrated approach could lead to a better understanding of this process. One of research directions is study of biological and neurologic mechanisms of speech perception [1–5]. Another research area is intelligibility of synthesized speech in noise [6, 7]. Much attention is paid to development of algorithms improving speech intelligibility in noise [8–10].

Study of listener's specifics on the process of speech recognition showed that music training affects positively speech-in-noise perception [11, 12]. Another important and rather controversial topic is perception of accented speech in noise by native and nonnative listeners; thus the research [13] revealed that native listeners were able to percept the test material at the same level regardless of the accent of the speaker while previous studies [14, 15] showed that speech of native speakers was generally found by native listeners more intelligible than speech of nonnative ones; nonnative listeners showed a trend of better perception of speech produced by speakers from the same language environment as themselves, that is, those having familiar accent [16, 17].

In Russia studies of speech characteristics and speech intelligibility and recognition in noise environment started in the middle of 20th century ([18–20], etc.). Experiments [19, 20] on word intelligibility for Russian speech against white and pink noise with levels of signal-to-noise ratio of 0 dB and lower showed different results: according to [19] the results for these two types of noise were very similar, while according to [20] word intelligibility for speech against white noise was higher. At the end of 20th century a major study of speech perception was carried out in white noise with various levels of signal-to-noise ratio [21], which investigated influence of different factors on Russian speech recognition: natural versus synthesized speech, different parts of speech, number of syllables in the word, place of stress, different types of phonemes, and so forth, which resulted in the range of the factors, which actually have influence on speech perception and recognition in noise environment and the level of influence of each of these factors. Another study [22] investigated changes of speaker characteristics (such as voice pitch, tempo of speech, voice strength) in the conditions of

switching from a native (Russian) language to a foreign one (English). The latest researches were focused on cognitive mechanisms of semantic content decoding of Russian speech in noise [23, 24], which demonstrated that dialogs have better intelligibility than monologs and reading, and such factors as background knowledge of listeners about the topic of the conversation and their general interest, as well as emotional level of the speaker, also influenced the process of speech perception.

The current research studies perception of native (Russian) and nonnative (German) speech in noisy environment (pink and white types of noise were chosen for the experiment) focuses on the following aims:

(1) To identify the effect of the tested types of noise with various levels of signal-to-noise ratio (in comparison) on speech perception;

(2) To identify effects of linguistic and extralinguistic factors on speech perception in noisy environment.

2. Method and Experiment

Our pilot research included perceptual-auditory analysis at various levels of linguistic structure of speech utterances in Russian and German (in comparison) in the same noisy environment: various (pink and white) types of noise with various levels of signal-to-noise ratio as well as effects of linguistic and extralinguistic factors on speech perception in noisy environment.

The research material of the study was a specially composed (according to the method of Potapova [25, 26]) ad hoc corpus of words and phrases in Russian and German in realizations of Russian and German native speakers, which were mixed with various (pink and white) types of noise with various levels of signal-to-noise ratio (0 dB, −3 dB, −6 dB, −9 dB, and −12 dB). The material allowed analyzing of protection and intelligibility degrees at acoustic, phonetic, syntactic, and lexical levels of linguistic structure.

The following requirements for development of the ad hoc research material were stated:

(1) Test phrases should be grammatically and semantically linked and consist of words which exist in both languages;

(2) Various types of consonants and vowels should be represented in the test phrases;

(3) Acoustic realization of the chosen types of consonants and vowels should be similar and comparable in both languages;

(4) Comparable (by place and manner of articulation) consonants and vowels should be in identical positions in a syllable (for all test words and phrases in Russian and German);

(5) The rhythmic scheme of test words and phrases should be identical for both languages;

(6) Combinations with various types of vowels in stressed position in the first syllable (with regard to unilateral distribution) should be tested for each type of consonants.

Table 1: Test phrases for experiment.

N	Analyzed languages	
	Russian	German
1	**Б**а**б**а **б**и**л**а **Б**о**р**ю ['babə 'bilə 'borju] *(Eng.: Woman beat Borya)*	**Bar**by **bit**tet **Bo**ris ['barbi 'bitət 'boːris] *(Eng.: Barby asks Boris)*
2	**П**а**п**а **п**и**ш**ет **П**о**л**е ['papə 'piʃət 'polje] *(Eng.: Father writes to Polya)*	**Pa**pa **spiel**te **Po**ker ['paːpa 'ʃpiːltə 'poːkɐ] *(Eng.: Father played poker)*
3	**З**а**н**ят **З**и**н**ин **зон**тик ['zanjat 'zinin 'zontik] *(Eng.: Zina's umbrella is booked)*	**Sa**ra **sitzt** am **So**fa ['zaːra 'zitst am 'zoːfa] *(Eng.: Sara is sitting on the sofa)*
4	**С**а**ш**а **с**и**т**о **с**у**ш**ит ['saʃə 'sitə 'suʃit] *(Eng.: Sasha dries sieve)*	**Sa**scha **schick**te **Schu**he ['saːʃa 'ʃiːktə 'ʃuːə] *(Eng.: Sasha sent shoes)*
5	**М**а**м**а **М**и**л**у **м**о**е**т ['mamə 'milu 'mojet] *(Eng.: Mother washes Mila)*	**Ma**ma **miet**et **Moll**e ['maːma 'miːtət 'mɔlə] *(Eng.: Mother helps Molle)*

According to the requirements of the research material the following test phrases were formulated (see Table 1: analyzed syllables are bold in the table; hereafter the system of IPA was used for transcriptions).

Each phrase consists of 3 words, having in the first stressed syllable a combination of the tested type of consonant with one of the tested types of vowels. Since pronunciation norms of the German language require voicing the voiceless fricative consonant "s" preceding a vowel, speakers were instructed to pronounce this sound as a voiceless one in the word "Sascha."

Speakers, who took part in the study, were native speakers of the literary Russian language without prominent dialectal features of pronunciation, speaking also German (50%) and being native speakers of the literary German language without prominent dialectal features of pronunciation, speaking also Russian (50%). The level of knowledge of foreign language of all speakers was the same, B2-C1, which was tested according to the system developed by the Council of Europe [27]. 50% of speakers were women (native Russian speakers and native German speakers) and 50% were men (native Russian speakers and native German speakers).

Each speaker read aloud test phrases and isolated words from phrases three times each. Thus, the total number of obtained realizations of the test words and phrases totaled 480; the total number of realizations for a single speaker was 120 (60 in Russian and 60 in German).

All test words and phrases were combined into 2 tables (in Russian and in German). The order of words and phrases was random and differed for different speakers. All test words and phrases were read with the intonation of the completed narrative, followed by a pause.

Audio recording of the research material was conducted in a specially equipped room, preventing foreign interference and noise: an anechoic chamber of the Institute of Applied

Table 2: Answer table for listeners.

# of phonogram	Do you hear a speech signal? (mark the corresponding cell with «x»)		What is the utterance language? (mark the corresponding cell with «x»)			Write down everything, you have heard (in Russian)
	Yes	No	Russian	German	Do not know	
1						
2						
⋮						
n						

and Mathematical Linguistics of Moscow State Linguistic University.

Two samples of noise (white and pink) were generated using the program Cool Edit Pro 2.0. for mixing them with audio records of the test words and phrases realizations.

Each speech segment was mixed with white and pink noise with various levels of signal-to-noise ratio: 0 dB, −3 dB, −6 dB, −9 dB, and −12 dB. Thus, for each spoken realization of the test material 10 variants of mixed signals were obtained, 4800 samples in total, plus 150 phonograms containing only noise (75 with white noise and 75 with pink noise). The total number of phonograms for the experiment was 4950.

The number of listeners was 21: 6 males and 15 females, 19–21 y.o., native speakers of the Russian language without prominent dialectal features of pronunciation with normal hearing, who have a command of English at level B2-C1 (which was tested on the system developed by the Council of Europe [27]). Some of them (12 listeners) are proficient in German at level B1-B2 and some (9 listeners) do not know German at all.

All phonograms were numbered randomly for presentation to each listener: total number of rotation variants was 11.

Listeners have to listen to phonograms according to their sequence numbers in the proposed rotation variant and to write down the answers for each of them in the table (see example of the answer table in the Table 2).

They could listen to each phonogram as many times they wanted. Each half an hour there were short breaks. Work time per day did not exceed 4 hours. The perception test run during 2 days: total work time for each listener was 8 hours.

The total number of played phonograms was 23085. The total number of played phonograms, which contained only noise, was 727.

The total number of played phonograms, containing speech signal (test words and phrases mixed with noise), was 22358 (the share of phonograms with white and pink noise types made up 50% each).

These calculations indicate that the size of the base of played phonograms is sufficient to ensure reliable and stable quality of the data.

A summary table with quantitative description of the experiment is presented in Table 3.

We calculated statistical sampling error for the findings to prove the observed tendencies statistically. The statistical sampling error was calculated using the following formula [28]:

$$E = Z_{\alpha/2}\sqrt{\frac{pq}{n}}, \quad (1)$$

Table 3: Quantitative description of the experiment.

N	Characteristics	Number
1	Research material: number of words and phrases in Russian	20
2	Research material: number of words and phrases in German	20
3	Research material: number of words and phrases in Russian and German	40
4	Number of speakers	4
5	Number of realizations of test words and phrases in Russian and German per one speaker	120
6	Total number of realizations (phonograms) of test words and phrases in Russian and German	480
7	Number of tested types of noise	2
8	Number of tested levels of signal-to-noise ratio for each type of noise	5
9	Number of phonograms for perceptual-auditory analysis, which contain speech signal (mixes of test words and phrases with noise)	4800
10	Number of phonograms for perceptual-auditory analysis, which contain only noise	150
11	Total number of phonograms for perceptual-auditory analysis	4950
12	Number of listeners	21
13	Number of auditions, containing only noise	727
14	Total number of auditions, containing speech signal (mixes of test words and phrases with noise)	22358
15	Total number of auditions	23085

where E is the statistical sampling error, $Z_{\alpha/2}$ is a z-value with a given confidence probability: in our case $Z_{\alpha/2} = 1.96$ or a confidence level of 95%, pq is dispersion of an alternative characteristic (dispersion of the sample share): in our case, as the sample share is unknown, the maximum value $pq = 0{,}25$ is taken [29], and n is the sample size.

3. Discussion

The working hypothesis of the study was as follows: speech recognition (detection of speech in noise and identification of the utterance language) and perception of the semantic content of the utterance in the variable noisy environment are influenced by the type of noise and the signal-to-noise ratio, as well as by some of linguistic and extralinguistic factors. These factors are the existing redundancy of a particular

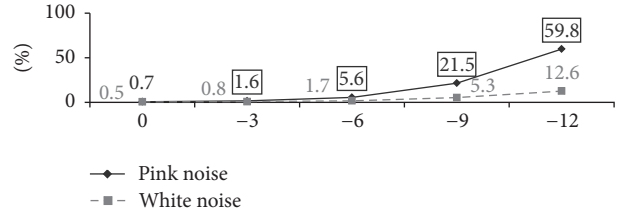


Figure 1: Influence of the signal-to-noise ratio on detection of speech signal in noise.

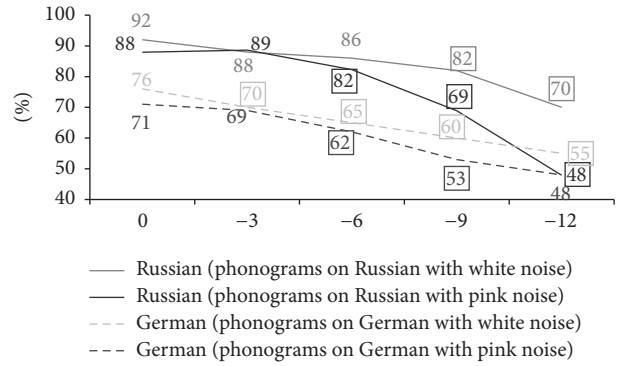


Figure 2: Recognition of the utterance language: comparison of Russian and German in noisy environment (with various types of noise and various levels of signal-to-noise ratio).

language at various levels of linguistic structure, changes in the acoustic characteristics of the speaker while switching from one language to another one, the speaker and listener's level of proficiency in a specific language, and acoustic characteristics of the speaker's voice.

The experiment showed that within the corpus of the research material pink noise provides better protection of the utterance than white noise at equal integral level of signal-to-noise ratio (for all tested levels) in terms of the following indicators: detection of speech signal in noise (see Figure 1), correct identification of the utterance language (see Figure 2), and correct perception of the semantic content of the utterance. On Figures 1 and 2 statistically significantly higher (at the level of 95%) values in relation to the previous higher level of signal-to-noise ratio are marked with a frame.

The lower the level of signal-to-noise ratio (the higher the level of noise over the level of the desired signal), the higher the difference of efficiency degree between pink and white types of noise, reaching its maximum at the lowest tested signal-to-noise ratio (−12 dB): while assessing detection of speech signal in noise, the efficiency of white noise masking is ~4.7 times lower as compared to pink noise. This result corresponds to findings observed in [20].

Detection of the utterance in noise also depends on level of speaker and listener's proficiency in a specific language, as well as on utterance language. Thus, for listeners, who are native Russian speakers, a higher score of detection of utterance in noise was shown for utterances in German, pronounced by native German speakers, than for utterances in Russian, pronounced by native Russian speakers (see

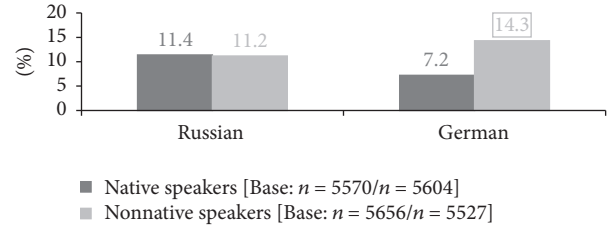


Figure 3: Utterance miss (false negative error) for listeners, who are native Russian speakers, depending on the characteristics of the speaker.

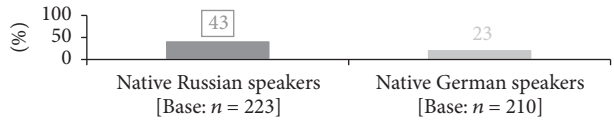


Figure 4: Recognition of the word «Поле» (['papə'piʃət 'polje], Eng.: «Papa writes to Polya»), depending on the characteristics of the speaker.

Table 4: Main substitutes (4% and more) for phonogram «била» (['bilə], *Eng.*: «*beat*»): native Russian speakers at signal-to-noise ratios = −9 and −12 dB ([Base: $n = 90$]).

NN	Substitutes	English transcription	Answer share (in %)
1	била	['bilə]	42
2	Мила	['milə]	6
3	мыла	['mɨlə]	4
4	хиар	['hiar]	4
5	26 substitutes with share 1%–3%		43

Figure 3: a statistically significantly higher (at the level of 95%) value of utterance miss (false negative error) for German (depending on the characteristics of the speaker) is marked with a frame), which demonstrates the effect of the language (and its acoustic characteristics) of the utterance on detection and recognition of the utterance in noise.

Listeners, who are native speakers of the language of the utterance, are able to detect native speech in noisy environment regardless the speaker's level of proficiency in this language of (see Figure 3); however a foreign accent of the speaker significantly reduces the score of recognition of the utterance language (see Figure 4: a statistically significantly higher (at the level of 95%) value of recognition (depending on the characteristics of the speaker) of the word «Поле» is marked with a frame), and recognition of the semantic content of the utterance, which confirms findings of experiments [14, 15].

Besides, for a number of phonograms with low scores of recognition, the list of these substituting words also differed for speech of native and nonnative speakers: Tables 4 and 5 show most frequent substitutes for the word «била» (['bilə'], *Eng.:* «*beat*») for phonograms of native and nonnative speakers.

From Tables 4 and 5 we can also see that quite a very wide range of substituting words with low answer shares

Table 5: Main substitutes (4% and more) for phonogram «била» (['bilə], *Eng.:* «*beat*»): native German speakers at signal-to-noise ratios = −9 and −12 dB ([Base: n = 107]).

NN	Substitutes	English transcription	Answer share (in %)
1	била	['bilə]	34
2	Зина	['zinə]	13
3	Мила	['milə]	13
4	Дима	['dimə]	6
5	28 substitutes with share 1%–3%		35

Figure 5: «Word cloud» for phonogram «била»: native Russian speakers at signal-to-noise ratios = −9 and −12 dB [Base: n = 90].

Figure 6: «Word cloud» for phonogram «била»: native German speakers at signal-to-noise ratios = −9 and −12 dB [Base: n = 107].

(1%–3%) was observed for these phonograms: 26 substitutes for phonograms of Russian native speakers and 28 substitutes for phonograms German native speakers and these lists were very different. These varieties of words with low answer shares are visible on the «tag clouds» for these two phonograms (e.g., on Figures 5 and 6). The «tag cloud» or «word cloud» technique was used for analysis of the substituting word frequencies, in which the font size for a particular word was used depending on frequency of its occurrence: the higher the frequency, the greater the font size for a specific word. All the "word clouds" for this experiment were drawn using the online program Wordle™ [30]. Thus on Figures 5 and 6 we can see lots of small words which are different for both phonograms.

At the acoustic level the degree of the utterance protection also depends on the fundamental frequency of the speaker's voice: within the corpus of the research words and phrases in Russian and German utterances in realizations of males are more concealed from detection against noise (i.e., they demonstrated higher score of utterance miss (false negative error) against noise) than utterances in realizations of females against both tested types of noise with all tested levels of signal-to-noise ratio (see Figures 7 and 8).

At the phonetic level various sounds have various degrees of intelligibility depending on their acoustic nature. Thus, the most resistant (among those tested within the research material of the experiment) to recognition in the stressed syllable are consonants [s] and [m], as well as vowel [a], while

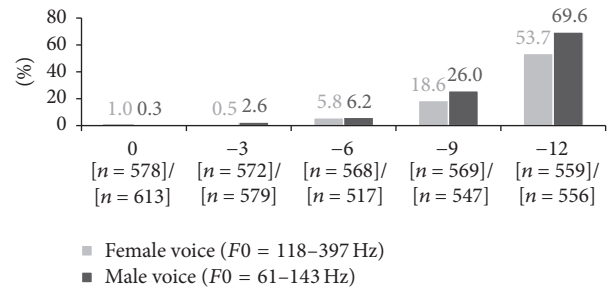


Figure 7: Utterance miss (false negative error) for utterances in Russian against pink noise: comparison of male and female voices.

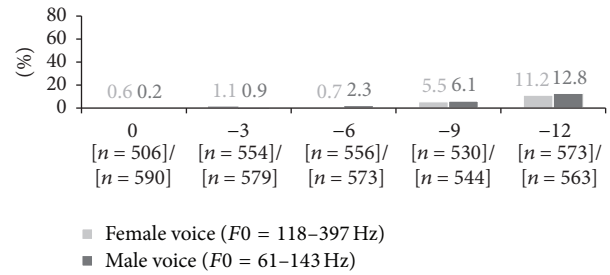


Figure 8: Utterance miss (false negative error) for utterances in Russian against white noise: comparison of male and female voices.

the most masked are consonants [b] and [p] and vowel [i] (see Figures 9–12).

At the syntactic level correct recognition of words depends on the context: the scores of correct recognition of words, which functioned as a part of a phrase, were higher than the scores for the same words in an isolated position against both tested (pink and white) types of noise at all tested levels of signal-to-noise ratios (see Figures 13 and 14: the data in Figure 13 is sorted in descending order and formatted in MS Excel from the dark colour to the light one).

In the phrases the best recognized parts were subjects (which always came first in the phonogram corpus), with predicates showing the second highest score (see Figures 15 and 16).

At the lexical level recognition of utterances in Russian is influenced by the occurrence of words in the language: within the research corpus of Russian words the highest score of recognition was shown by words «мама», «Саша», and «папа», which have the highest frequency of occurrence in the Russian language among all the tested words, while the lowest scores of recognition were shown by words «Поле», «Зинин», «Борю», and «Милу», which are proper names and less common than «Саша» (see Table 6 index; ipm represents number of occurrences of lemma per one million words in the corpus; here the score is presented according to [31]).

4. Conclusion

According to the results of the experiment the following factors influence intelligibility and perception of the speech in noisy environment:

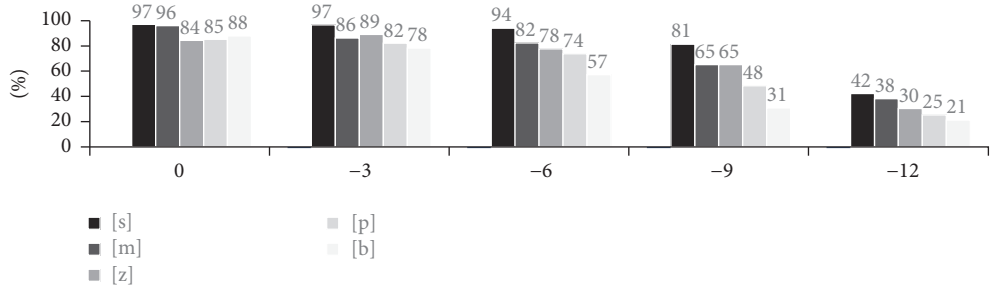


Figure 9: Correct recognition of consonants in first stressed syllable against pink noise.

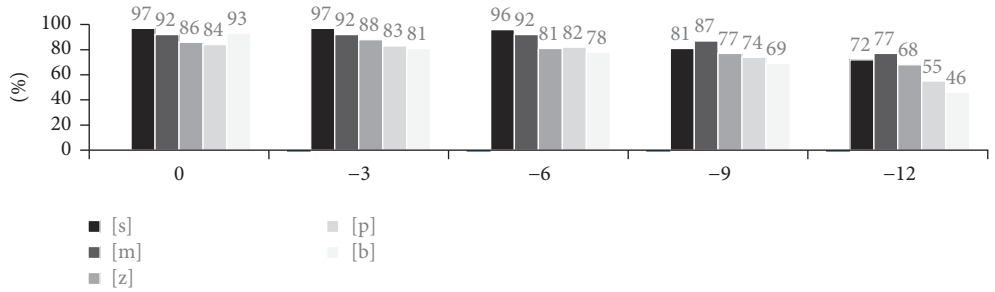


Figure 10: Correct recognition of consonants in first stressed syllable against white noise.

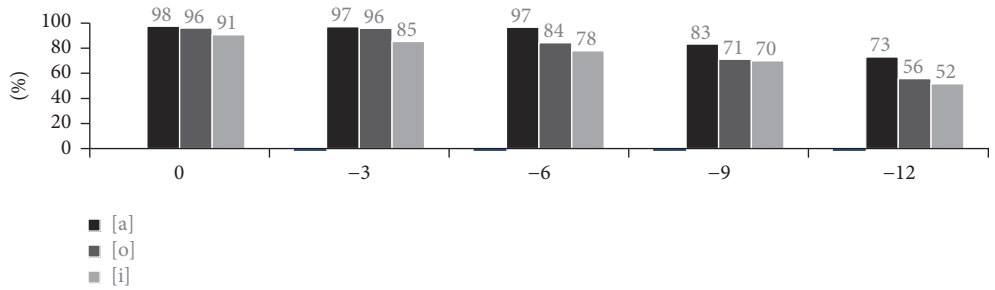


Figure 11: Correct recognition of vowels in first stressed syllable against pink noise.

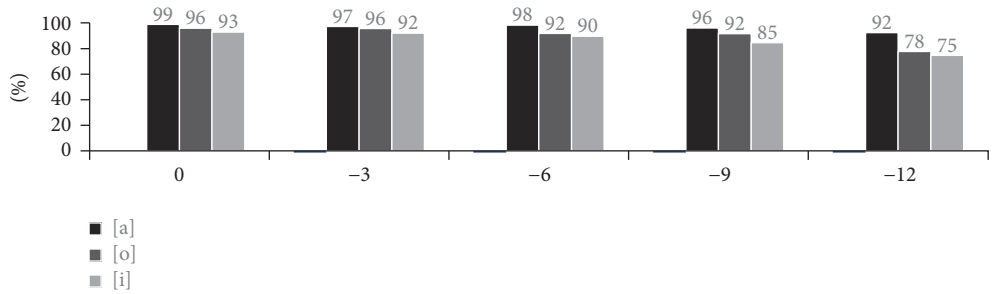


Figure 12: Correct recognition of vowels in first stressed syllable against white noise.

(1) Type of noise and signal-to-noise ratio (pink noise provides better protection of the utterance than white noise at equal integral level of signal-to-noise ratio (for all tested levels) in terms of the following indicators: detection of speech signal in noise, correct identification of the utterance language, and correct perception of the semantic content of the utterance);

(2) Utterance language and speaker and listener's proficiency in a specific language;

(3) Fundamental frequency of the speaker's voice (within the corpus of the research material in Russian and German utterances read by males was detected by listeners statistically rarely than utterances read by

NN	Word	White noise (in %)		Pink noise (in %)	
		Isolated	A part of a phrase	Isolated	A part of a phrase
1	Саша *(Eng.: «Sasha»)*	89	97	83	92
2	папа *(Eng.: «Father»)*	87	94	73	89
3	сушит *(Eng.: «dries»)*	84	91	81	86
4	сито *(Eng.: «sieve»)*	73	92	71	86
5	баба *(Eng.: «Woman»)*	79	88	62	76
6	моет *(Eng.: «washes»)*	77	84	69	75
7	била *(Eng.: «beat»)*	56	88	39	76
8	Милу *(Eng.: «Milu»)*	55	79	47	71
9	Борю *(Eng.: «Boryu»)*	52	84	41	72
10	Поле *(Eng.: «Polye»)*	38	61	27	54

Figure 13: Comparative analysis of correct recognition of words in Russian against various types of noise in various context (isolated words or as a part of a phrase), in %.

N	Signal-to-noise ratio	White noise		Pink noise	
		Isolated	A part of a phrase	Isolated	A part of a phrase
1	0 dB	1	0	1	0
2	−3 dB	1	0	3	0
3	−6 dB	3	0	5	1
4	−9 dB	4	0	9	2
5	−12 dB	8	2	14	10

Figure 14: Number of tested words (from 15), the score of recognition of which did not exceed 50% at each level of signal-to-noise ratio.

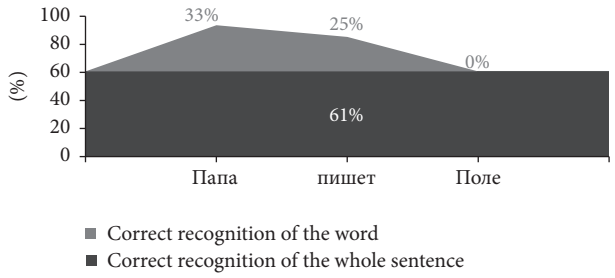


Figure 15: Recognition of isolated words in the phrase «Папа пишет Поле» *(Eng.: Father writes to Polya)* against white noise [Base: $n = 252$].

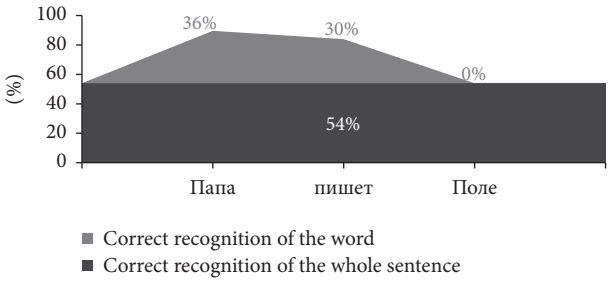


Figure 16: Recognition of isolated words in the phrase «Папа пишет Поле» *(Eng.: Father writes to Polya)* against pink noise [Base: $n = 247$].

females against both tested types of noise with all levels of signal-to-noise ratio);

(4) Context: isolated word or as a part of the phrase (within the corpus of the research material in Russian intelligibility of words within the phrase was better against both tested types of noise with all levels of signal-to-noise ratio compared to isolated words);

(5) Frequency of word occurrence in the language (according to the results of the experiment, words with higher frequency of occurrence in the Russian language showed better intelligibility);

(6) Phonetic composition of the word (within the corpus of the research material in Russian the voiceless sibilant fricative alveolar [s] and sonorant bilabial [m] among consonants and central open [a] among

Table 6: Frequency of occurrence in the Russian language of the tested words with the highest and the lowest scores of recognition (according to the results of the current experiment).

N	Word	ipm *(number of occurrences of lemma per one million word)*
1	мама *(Eng.: «mother»)*	322,6
2	Саша *(Eng.: «Sasha»)*	93,6
3	папа *(Eng.: «father»)*	143,4
4	Милу *(Eng.: «Milu»)*	10
5	Борю *(Eng.: «Boryu»)*	19,6
6	Зинин *(Eng.: «Zina's»)*	20,8
7	Поле *(Eng.: «Polye»)*	8,1

vowels showed the best intelligibility (i.e., the worst ability of masking using noise) within the tested set of sounds, while among consonants stop bilabial ones: voiced [b] and voiceless [p] and front close [i] among vowels showed the worst intelligibility, that is, the best ability of masking using noise).

Among the further possible directions of analysis the following can be mentioned:

(1) Increase of volume of bilingual research material;

(2) Expansion of the inventory of acoustic parameters for the analysis of the language sounds recognition;

(3) Increase of the number of speakers and listeners taking into account such factors as age, gender, degree of experience in listening, and proficiency in the utterance language, in relation to the studied languages;

(4) Study of the influence of linguistic and extralinguistic factors on the recognition in noisy environment for long connected texts;

(5) Organization of the database, including units of the sound composition and intonation system of various languages.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research was supported by Ministry of Education and Science of Russian Federation (Project no. 34.1254.2014K, head of the project R. K. Potapova).

References

[1] S. Anderson, E. Skoe, B. Chandrasekaran, S. Zecker, and N. Kraus, "Brainstem correlates of speech-in-noise perception in children," *Hearing Research*, vol. 270, no. 1-2, pp. 151–157, 2010.

[2] S. Anderson, A. Parbery-Clark, H.-G. Yi, and N. Kraus, "A neural basis of speech-in-noise perception in older adults," *Ear and Hearing*, vol. 32, no. 6, pp. 750–757, 2011.

[3] J. Cunningham, T. Nicol, S. G. Zecker, A. Bradlow, and N. Kraus, "Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement," *Clinical Neurophysiology*, vol. 112, no. 5, pp. 758–767, 2001.

[4] D. Poeppel, W. J. Idsardi, and V. Van Wassenhove, "Speech perception at the interface of neurobiology and linguistics," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1071–1086, 2008.

[5] J. H. Song, E. Skoe, K. Banai, and N. Kraus, "Training to improve hearing speech in noise: biological mechanisms," *Cerebral Cortex*, vol. 22, no. 5, pp. 1180–1190, 2012.

[6] K. D. R. Drager and J. E. Reichle, "Effects of discourse context on the intelligibility of synthesized speech for young adult and older adult listeners: applications for AAC," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 5, pp. 1052–1057, 2001.

[7] S. King, J. Yamagishi, and C. Valentini-Botinhao, "Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise," in *Proceedings of the SAPA-SCALE Workshop on Statistical and Perceptual Audition (SAPA-SCALE '12)*, Portland, Ore, USA, 2012.

[8] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 4, pp. 4160–4164, Orlando, Fla, USA, May 2002.

[9] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[10] N. Zoghlami and Z. Lachiri, "Application of perceptual filtering models to noisy speech signals enhancement," *Journal of Electrical and Computer Engineering*, vol. 2012, Article ID 282019, 12 pages, 2012.

[11] A. Parbery-Clark, E. Skoe, C. Lam, and N. Kraus, "Musician enhancement for speech-in-noise," *Ear and Hearing*, vol. 30, no. 6, pp. 653–661, 2009.

[12] J. Slater, E. Skoe, D. L. Strait, S. O'Connell, E. Thompson, and N. Kraus, "Music training improves speech-in-noise perception: longitudinal evidence from a community-based music program," *Behavioural Brain Research*, vol. 291, pp. 244–252, 2015.

[13] M. Pinet, P. Iverson, and B. G. Evans, "Perceptual adaptation for l1 and l2 accents in noise by monolingual British English listeners," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, August 2011.

[14] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 49, no. 1, pp. 285–310, 1999.

[15] M. J. Munro, "The effects of noise on the intelligibility of foreign-accented speech," *Studies in Second Language Acquisition*, vol. 20, no. 2, pp. 139–154, 1998.

[16] J. Song and P. Iverson, "Measuring speech-in-noise intelligibility for spontaneous speech: the effect of native and non-native accents," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, August 2015.

[17] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1600–1610, 2003.

[18] Y. S. Bykov, Theory of speech intelligibility and improving the efficiency of telephone communications, Moscow, Russia, 1959 (Russian).

[19] N. B. Pokrovskiy, "Calculation and measurement of speech intelligibility," Moscow, 1962 (Russian).

[20] M. A. Sapozhkov, "Speech signal in cybernetics and communication," Moscow, 1963 (Russian).

[21] L. V. Zlatoustova, R. K. Potapova, V. V. Potapov, and V. N. Trunin-Donskoy, *General and Applied Phonetics: Textbook*, MSU, Moscow, Russia, 2nd edition, 1997 (Russian).

[22] R. K. Potapova and O. N. Statsenko, "Preliminary results of the investigation of oral speech material considering code-switching," in *Proceedings of the 16th Session of the Russian Acoustical Society*, Moscow, Russia, November 2005.

[23] R. Potapova and V. Potapov, "Associative mechanism of foreign language perception (forensic phonetic aspect)," in *Proceedings of the Speech and Computer. SPECOM 2014*, A. Ronzhin, R. Potapova, and V. Delic, Eds., vol. 8773 of *Lecture Notes in Computer Science*, pp. 113–122, Springer International Publishing, Cham, Switzerland, 2014.

[24] R. Potapova and V. Potapov, "Cognitive mechanism of semantic content decoding of spoken discourse in noise," in *Proceedings of the Speech and Computer. SPECOM 2015*, A. Ronzhin, R. Potapova, and N. Fakotakis, Eds., vol. 9319 of *Lecture Notes in Computer Science*, pp. 153–160, Springer Internartional Publishing, Cham, Switzerland, 2015.

[25] R. K. Potapova, "Syllabic phonetics of the Germanic languages: textbook," Moscow, 1986 (Russian).

[26] R. K. Potapova, *Speech: Communication, Information, Cybernetics*, Moscow, Russia, 4th edition, 2010 (Russian).

[27] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

[28] O. V. Ivanov, "Statistics/Training course for sociologists and managers. Part 2. Confidence intervals. Hypothesis testing. Methods and their application," Moscow, 2005 (Russian).

[29] N. Sh. Kremer, *Probability Theory and Mathematical Statistics: Textbook for Universities*, Moscow State University, Moscow, Russia, 2nd edition, 2004 (Russian).

[30] Online program for creating "word clouds" Wordle™, http://www.wordle.net.

[31] O. N. Lyashevskaya and S. A. Sharov, New frequency dictionary of Russian vocabulary. The Frequency dictionary of modern Russian language (on the materials of the National corpus of the Russian language), (Russian), http://dict.ruslang.ru/freq.php.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 7437860, 11 pages
http://dx.doi.org/10.1155/2016/7437860

Research Article

A Novel DBN Feature Fusion Model for Cross-Corpus Speech Emotion Recognition

Zou Cairong,[1,2] Zhang Xinran,[2] Zha Cheng,[2] and Zhao Li[2]

[1]*Department of Information and Communication Engineering, Guangzhou Maritime University, Guangzhou 510006, China*
[2]*Key Laboratory of Underwater Acoustic signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Zhang Xinran; zxrzxr87324@126.com

Received 24 June 2016; Revised 19 October 2016; Accepted 15 November 2016

Academic Editor: Alexey Karpov

The feature fusion from separate source is the current technical difficulties of cross-corpus speech emotion recognition. The purpose of this paper is to, based on Deep Belief Nets (DBN) in Deep Learning, use the emotional information hiding in speech spectrum diagram (spectrogram) as image features and then implement feature fusion with the traditional emotion features. First, based on the spectrogram analysis by STB/Itti model, the new spectrogram features are extracted from the color, the brightness, and the orientation, respectively; then using two alternative DBN models they fuse the traditional and the spectrogram features, which increase the scale of the feature subset and the characterization ability of emotion. Through the experiment on ABC database and Chinese corpora, the new feature subset compared with traditional speech emotion features, the recognition result on cross-corpus, distinctly advances by 8.8%. The method proposed provides a new idea for feature fusion of emotion recognition.

1. Introduction

In recent years, more attention is paid to the study of emotion recognition. Speech, as one of the most important ways of communication in human daily life, contains rich emotional information. Speech emotion recognition (SER), because of its wide application significance and research value in intelligence and naturalness of human-computer interaction aspects [1], gets more and more attention from the researchers in recent years. Emotion recognition system performance determines the quality of information feedback and the efficiency of human-computer interaction, while overall performance of SER depends on the matching degree between features and classifiers [2]. Although the earlier temporal features may not be suitable for the current corpus structures [3], the emotional information contained on the time domain still has good representation ability to be reserved. In order to research SER on the broader technology level, extending the database source and searching suitable fusion model for big emotional information data have become new focuses [3, 4].

Feature layer fusion is the integration of data after preprocessing and feature extraction, so many related researches [5, 6] are applied to this area. Through specific means such as fusion, the scale of feature sources is enhanced and the data sets are expanded. Further, some effective data analysis techniques are introduced and applied, such as Neural Network and Deep Learning. Common feature fusions are often used for single source data samples. Because the emotional properties of different features are various, the cross-corpus recognition effects of current fusion methods are not satisfactory. The development of Deep Learning technology brings a new orientation to SER. Using appropriate algorithm to train the deep neural network model, more valuable features can be derived from the vast amounts of original databases which are multiple sources [7]. Accordingly, Deep Belief Nets (DBN) model, which is a commonly used model in Deep Learning area [8], is introduced in our work. Through Restricted Boltzmann Machine (RBM) [9], DBN could constantly adjust the connection weight, which can realize effective fusion of features. Previous cross-corpus studies are dependent on traditional suprasegmental acoustic global features, which are often used in emotion recognition technology [10]. Since the emotional features have great significance to SER, exploring new features to promote the development of SER

has an irreplaceable role in the cross-corpus research. Thus, this article introduces a new emotional feature category based on visual attention mechanism. The new feature space includes three kinds of image vectors: color, brightness, and orientation. Features extracted from spectrogram connect the time domain and frequency domain, so they have important significance for cross-corpus SER research. The new direction of the research which uses the spectrogram emotional features [4, 5] has advantages for its overall information. The integrated features with traditional acoustic traits could combine the global and temporal features, which supplement the original feature space.

This paper mainly studies the feature fusion method based on DBN which fuse spectrogram features and acoustic global features for SER. In Section 2, by selective attention mechanism, the features with time-frequency domain correlation traits are extracted while the emotion recognition abilities are analyzed. Then in Section 3, based on the DBN fusion method on feature level, an alternative DBN (so-called DBN21) feature fusion layer model is proposed for the extraction of spectrogram feature fusion. After that, the approximate optimal feature subset is obtained for overcoming the shortage of recognition ability differences between adjacent frames, which often appears in the traditional feature fusion method. Furthermore, as for the cross-corpus cases, a modified DBN network model (so-called DBN22) is designed for spectrogram features and acoustic features fusion. In Section 4, proven by simulation experiments on four databases, the features of proposed fusion method effectively improve the performance of SER system on cross-corpus.

2. Spectrogram Feature Extraction Based on Selective Attention

Spectrogram, namely, speech spectrum diagram, is based on the time domain signal processing, which has the horizontal axis representing time, the vertical axis representing the frequency, and the depth of the midpoint chart color representing the strength of the corresponding signal. Spectrogram is a communication between the time domain and frequency domain, which reflects the correlation of the two domains. Because spectrogram is the visual expression of the time-frequency distribution of the speech energy [11], it contains characteristic information, such as energy and formant. In our research, based on STB/Itti model [12], selective attention features on the orientation, color, and brightness of spectrogram are extracted as new characteristics for SER. Meanwhile, the dimension reduction and optimization for the features are conducted by proposed DBN model. And then an improved kernel learning K-nearest neighbor algorithm based on feature line centroid (kernel-KNNFLC) [13] classifier is carried on the experiment. The results show that the extracted features possess more powerful emotion recognition ability than their contrast. The spectrogram feature extraction process in SER with selective attention mechanism and DBN is shown in Figure 1.

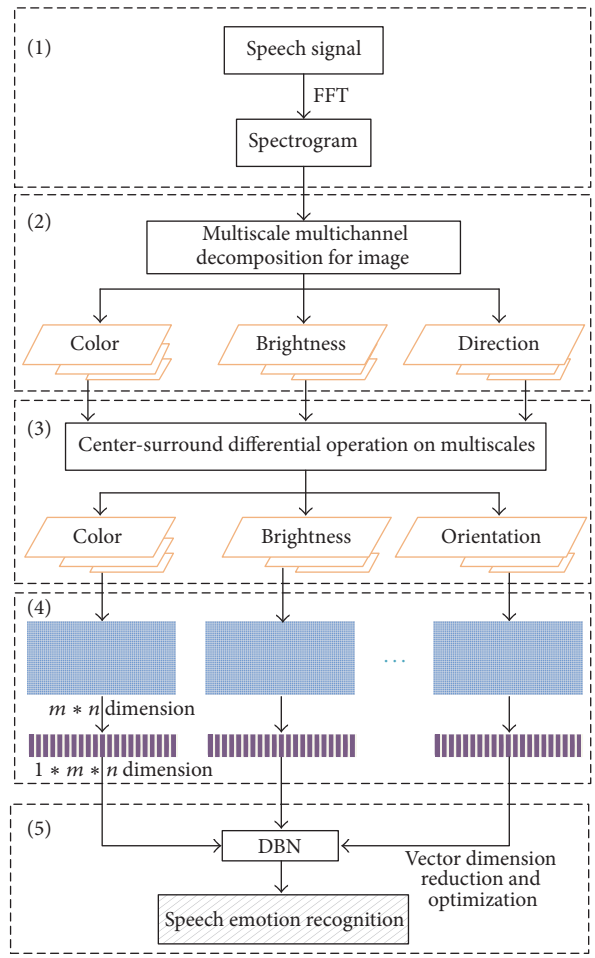


Figure 1: Spectrogram feature extraction process in SER.

2.1. Spectrogram Feature Extraction. The computation formula of spectrogram is as follows:

$$L = |Y| = \left| \sum_{n=0}^{N-1} s(n)\, \omega(n)\, e^{-j(2\pi/N)kn} \right|, \quad k \in [0, N]. \tag{1}$$

$s(n)$ represents the input signal, $\omega(n)$ represents the hamming window function, and N is the window length. Figure 2 is the spectrogram extracted from a piece of the speech labeled "aggressive" emotion in ABC corpus.

2.2. Gaussian Pyramid Decomposition. Based on the mechanism of selective attention, the area is easy to get the attention of people in a picture, which usually has strong difference compared to the surrounding area [14]. Multiscale multichannel filtering can be resolved by convolution operation with the linear Gaussian kernel. A 6×6 Gaussian kernel is used in this paper. The resulting image formula after Gaussian Pyramid Decomposition (GPD) is as follows:

$$I(\sigma + 1) = \frac{I(\sigma)}{2}, \quad \sigma = [0, 1, 2, 3, 4, 5, 6, 7, 8], \tag{2}$$

where σ represents the layer number and $I(\sigma)$ is the σ layer of the image after decomposition, in which $I(0)$ is the original

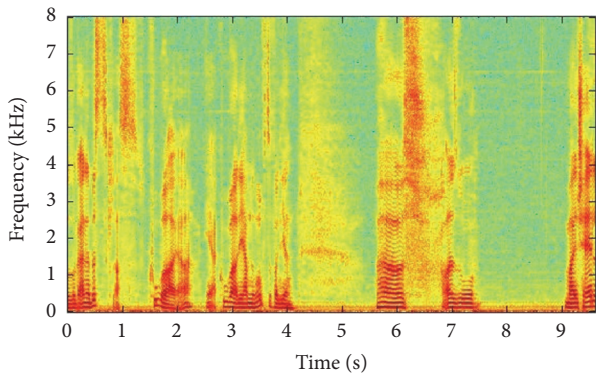


Figure 2: Spectrogram sample graph.

image. After the multiscale multichannel filtering, feature extractions are conducted of each scale image in orientation, color, and brightness, and then sequence images are formed, respectively.

In the retinal cone photoreceptors response level, the model is trichromatic mechanism. However, in the process of choosing messages for center selection in the brain, it changes into 4 primary mechanisms. As a result, 4 primary channels are defined in Itti model. Therefore, the antagonism of the *R-G* and *B-Y* colors could be used to simulate the saliency contribution which is made by the colors to images. And then the computation formula is

$$P_{R\text{-}G}(\sigma) = \frac{(r-g)}{\max(r,g,b)},$$
$$P_{B\text{-}Y}(\sigma) = \frac{(b-\min(r,g))}{\max(r,g,b)}. \tag{3}$$

In formula (3) r, g, and b, respectively, represent the three primary colors: red, green, and blue. Here are 16 GPD images based on the extracted color features of the different scale images.

The GPD images of brightness features are obtained after calculating the average of the normalized r, g, and b:

$$P_I(\sigma) = \frac{(r+g+b)}{3}. \tag{4}$$

Here are 8 GPD images of brightness.

The 2D Gabor directional filter could be used to simulate the directional selection mechanism of the retina [15]; therefore, we can use its convolution with the GPD of brightness feature to get the GPD images of local orientation feature. It has been proven that the angle $\theta \in (0^\circ, 45^\circ, 90^\circ, 135^\circ)$ can be used to represent the orientation feature:

$$P_\theta(\sigma) = \left\|P_I(\sigma) * G_0(\theta)\right\| + \left\|P_I(\sigma) * G_{\pi/2}(\theta)\right\|. \tag{5}$$

The corresponding formula is as follows:

$$G_\psi(\theta) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\delta^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right). \tag{6}$$

γ is the orientation rate which has a value of 1; δ and λ, respectively, represent the standard deviation and the wavelength which have the value of 7/3 pixels and 7 pixels; ψ is the phase and $\psi \in \{0, \pi/2\}$. 32 GPD images of orientation feature could be obtained by using a total of 8 scales and 4 directions, 2D Gabor.

2.3. Features Obtaining and Matrix Reconstruction. Relying on the color and brightness features of GPD extracted previously, they cannot attract the selective attention insufficiently, which also needs the difference contrast of image characteristics. These features compared with the traditional acoustic global features, properties, have better characterization of different speech sample sources (language, speakers, including noise, etc.) in which emotional information is implied. In our research the center-surround is applied to the computing method of calculation [16]. Experimental results show that this center-surround method brings the model more reliable robustness in cross-corpus SER. After the calculation of contrastive feature vector, the gist feature images could be obtained based on the merger strategy (local iterative normalized).

$$FM_l(\sigma_c, \sigma_s) = N\left(\left|P_l(\sigma_c) - P_l(\sigma_s)\right|\right), \tag{7}$$

where $l \in \{R\text{-}G, B\text{-}Y, I, 0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ represents the kinds of gist feature images which are a total of 7, including the *R-G* and *B-Y* 2 kinds of color features, one kind of brightness features, and four kinds of orientation features; $\sigma_c \in \{2, 3, 4\}$ is the central scale of Gaussian pyramid; and $\sigma_s = \sigma_c + d$ is the surrounding scale, among which $d = \{2, 3\}$. $N(\cdot)$ represents the merger strategy with local iterative normalized [17]. Finally we received 12 color-contrast, 6 brightness-contrast, and 24 orientation-contrast feature images. The extracted gist feature images based on the speech samples are shown in Figure 3.

A feature image is lot into m lines and n columns, forming total $m * n$ subregions. Then each subregion is replaced by its mean. Furthermore, the images are normalized to a $m * n$ feature matrix, so that a low resolution feature matrix of image is used to describe the whole spectrogram. The mathematical representation of the feature matrix is as follows:

$$FD_i(p,q) = \frac{mn}{vh}\sum_{g=pv/n}^{(p+1)v/n-1}\sum_{f=qh/m}^{(q+1)h/m-1} FM_i(g,f),$$
$$p \in [0, n-1],\ q \in [0, m-1], \tag{8}$$

among which FM_i is feature image and FD_i is corresponding feature matrix, $i \in [1, 42]$. Here, m is 4 and n is 5. h and v represent the height and width of the feature image, respectively. And then, the characteristic matrix obtained is reconstructed to a $1 \times mn$ vector, in which the feature performance on the cross-corpus SER will be validated in the subsequent experiments.

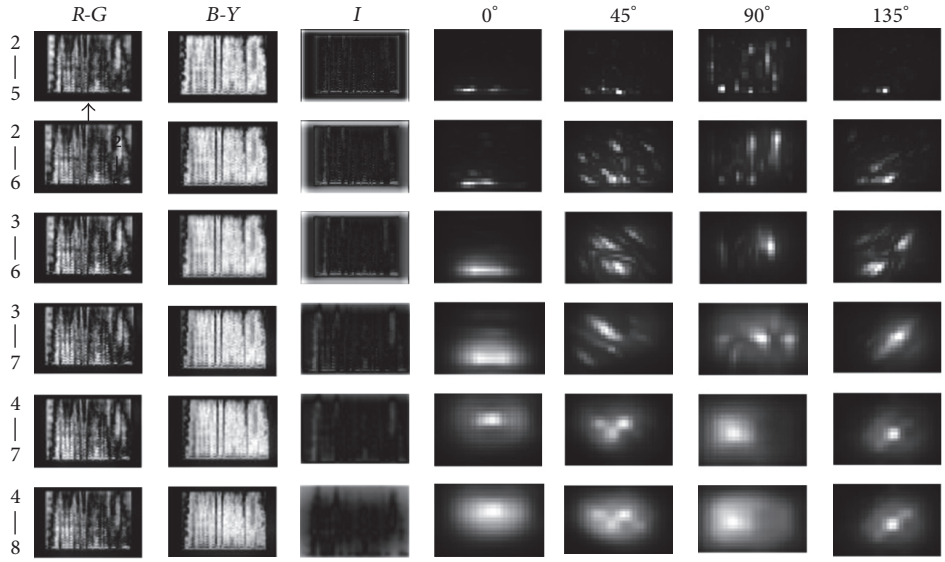


Figure 3: Gist feature images based on spectrogram.

3. DBN Feature Fusion Model for SER

The Deep Belief Nets model rooted in statistical mechanics, which is described through the energy function and probability distribution function. Energy function can reflect the stability of the system. When the system in an orderly state and the probability distribution are intense concentrated, the energy of the system is small. Conversely, if the system is in disorder and the probability distribution is uniform, the energy of system may be larger. DBN model is formed in a multilayer stack RBM, just like constructing a building. The RBM is accumulated in layers and is evaluated one by one from the bottom to the top. The training of each layer is independent while the top RBM has self-associative memory according to the information from the lower. Eventually the Error Back Propagation (BP) algorithm is applied to fine-tune weight. At the top of DBN, the kernel-KNNFLC classifier is connected for classification.

3.1. Restricted Boltzmann Machine in DBN. Boltzmann Machine (BM) is a kind of random neural network model, which is made up of two parts: visible and hidden layer. Although BM has strong unsupervised-learning ability and could learn the complex rules in the data, the training time is tremendous long. To solve this problem, Smolensky proposed the RBM, the structure of which is as shown in Figure 4.

The model figure reveals that it is inexistence of internal connection between the visible layer and hidden layer of RBM, which has the property: if the state of the hidden units is given, activated units in visible layer are conditionally independent, so that if the unit number of hidden and layers visible of RBM is m and n, respectively, which state vectors are h and v, according to a given state (v, h), the energy could be defined as follows:

$$E(v, h \mid \theta) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n} \sum_{j=1}^{m} v_i W_{ij} h_j, \quad (9)$$

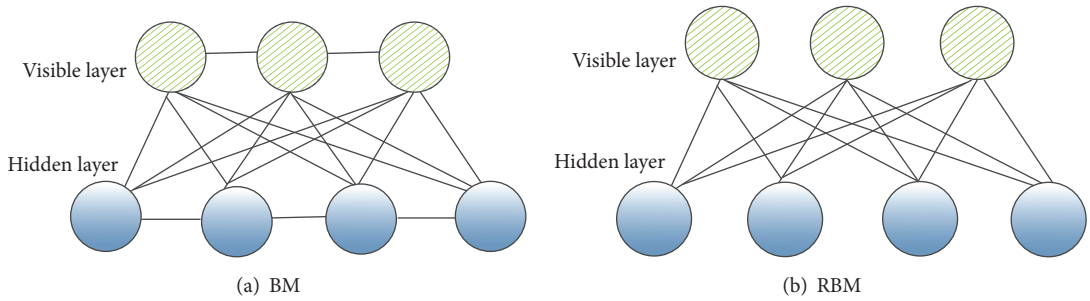
in which a_i and b_j are the values of bias of visible unit i and hidden unit j, respectively, and W_{ij} represents the connection weight of j and i. Here $\theta = \{W_{ij}, a_i, b_j\}$ is used as the whole parameter set in RBM. When the parameter set is determined, the joint probability distribution of (v, h) could be obtained according to formula (10), as shown in the following formula:

$$P(v, h \mid \theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)}, \quad Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}. \quad (10)$$

Here $Z(\theta)$ is called the partition function. Because, with the unit being given by RBM, the activated states between each hidden unit are independent, if it is in a given unit state, the activation probability of j and i could be obtained as follows:

$$P\left(h_j = 1 \mid v, \theta\right) = \sigma\left(b_j + \sum_i v_i W_{ij}\right),$$
$$P\left(v_i = 1 \mid h, \theta\right) = \sigma\left(a_i + \sum_j h_j W_{ij}\right). \quad (11)$$

3.2. The Fast Learning Algorithm Based on Contrast Divergence. Gibbs Sampling algorithm is based on Markov Chain Monte Carlo (MCMC) strategy [18]. By getting a conditional probability distribution of the weight, which can begin from any state, the algorithm implements iteration sampling in turn for each component. Gibbs Sampling method is used to obtain the probability distribution, which is often necessary to employ a lot of sampling steps. In particular within the high-dimension data, the training efficiency of model may be greatly influenced. Therefore, Hinton proposed a fast learning algorithm of RMB called contrastive divergence (CD) [19]. Unlike Gibbs Sampling, this method (CD) uses the training data to initialize and just needs k steps (usually $k = 1$) to



(a) BM (b) RBM

Figure 4: BM and RBM model.

gain a satisfactory approximation. At the beginning of the CD algorithm, the visible unit state is set to a training sample, and then formula (12) is used to calculate the unit state of the hidden layer. After that, the probability of the ist unit hidden values equaling 1 could be calculated according to formula (12). Furtherly, refactoring of visible layer is obtained.

The task of training RBM is to get parameters θ. The logarithm likelihood function is obtained through the training set that is maximized for parameter set θ, which may fit the given training data. If the number of training samples is T, there are

$$\theta^* = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \sum_{t=1}^{T} \log P\left(v^{(t)} \mid \theta\right). \tag{12}$$

Then, the Stochastic Gradient Ascent method is used to find the optimal parameters maximizing equation (12):

$$\begin{aligned} \frac{\partial}{\partial\theta} L(\theta) &= \frac{\partial}{\partial\theta} \sum_{t=1}^{T} \log \sum_{h} P\left(v^{(t)}, h \mid \theta\right) = \frac{\partial}{\partial\theta} \\ &\cdot \sum_{t=1}^{T} \log \frac{\sum_h e^{[-E(v^{(t)},h|\theta)]}}{\sum_v \sum_h e^{[-E(v^{(t)},h|\theta)]}} \\ &= \sum_{t=1}^{T} \left(\left\langle \frac{\partial}{\partial\theta} \left(-E\left(v^{(t)}, h \mid \theta\right)\right) \right\rangle_{P(h|v^{(t)},\theta)} \right. \\ &\left. - \left\langle \frac{\partial}{\partial\theta} \left(-E(v, h \mid \theta)\right) \right\rangle_{P(v,h|\theta)} \right). \end{aligned} \tag{13}$$

In formula (13) $\langle\cdot\rangle_P$ is calculating mathematical expectation of the distribution P. The first item of the formula can be determined by the training sample, while $P(v, h \mid \theta)$ in the following item need to get the joint probability distribution of visible and hidden units first. And then, for calculating the distribution function $z(\theta)$, which cannot be directly calculated, the sampling method (such as Gibbs Sampling) is introduced to approximate the related value. When using "data" as the tag of $P(h \mid v^{(t)}, \theta)$ and "model" as $P(v, h \mid \theta)$, the offset on visible and hidden units of formula (13) is a_i and b_j, respectively, and the connection weight is W_{ij}. Then a partial derivative is available:

$$\begin{aligned} \frac{\partial}{\partial a_i} \left[\log P(v \mid \theta)\right] &= \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}}, \\ \frac{\partial}{\partial b_j} \left[\log P(v \mid \theta)\right] &= \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{model}}, \\ \frac{\partial}{\partial W_{ij}} \left[\log P(v \mid \theta)\right] &= \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}. \end{aligned} \tag{14}$$

Based on the criteria of formula (14), the method of Stochastic Gradient Rise is used to maximize the value of the logarithm likelihood function on the training data. Therefore, the updating criteria of parameters are

$$\begin{aligned} \Delta a_i &= \varepsilon \left(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}} \right), \\ \Delta b_j &= \varepsilon \left(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}} \right), \\ \Delta W_{ij} &= \varepsilon \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}} \right), \end{aligned} \tag{15}$$

in which ε is the learning rate and $\langle\cdot\rangle_{\text{recon}}$ represents the distribution of model defined after one step refactoring.

From the above contents, the training procedure of RBM algorithm is divided into a few steps:

(1) Firstly, initialization of RBM is necessary. Thus, mainly the following contents are included: sample training set S; the number of neurons n_h contained in hidden layer, the number of visible layer neurons n_v; the connection weight W_{ij} of visible and hidden layer; the unit biases a_i and b_j of visible and hidden layer; the learning rate ε and the training cycle J; the number of the algorithm steps k.

(2) Rapid sampling is carried out based on the CD-k algorithm. Furtherly, according to the updates of each parameter, the value of parameter set is refreshed.

(3) The sampling process is repeated in the whole training period, until the convergence of formula (12).

3.3. DBN21 and DBN22 Models. According to the RBM, two kinds of DBN models, respectively, DBN21 and DBN22,

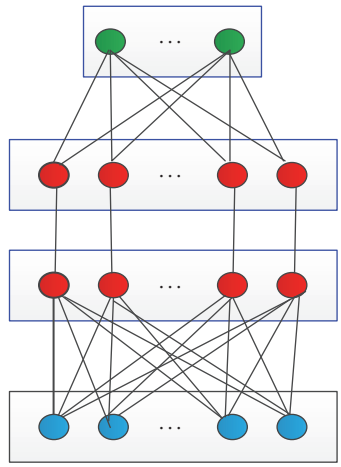


Figure 5: DBN21 model.

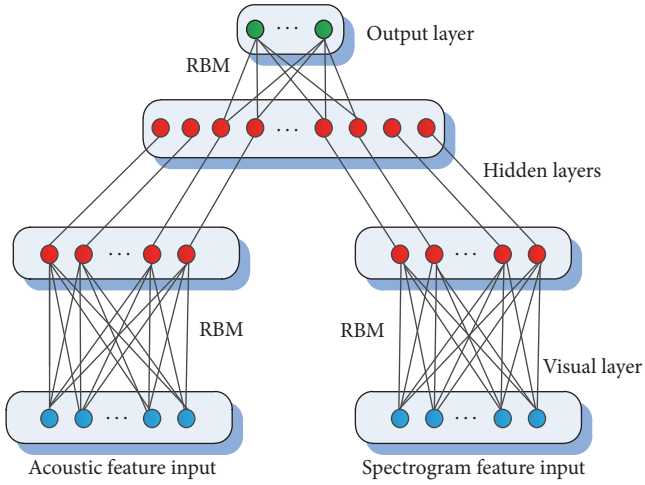


Figure 6: DBN22 model.

are structured for SER experiments on cross-database. As shown in Figures 5 and 6, (1) the DBN21 model is proposed for separate layer feature fusions with spectrogram features and traditional acoustic features (by the international general extracting method mentioned in Section 4.1.3); (2) the DBN22 model is constructed for integration of the spectrogram and traditional acoustic features in the feature layer. Because the speech emotion features extracted for SER experiments are real number data, it is not appropriate to apply the binary RBM for modeling. As a result, we chose the Gaussian-Bernoulli RBM (GRBM) [20] to build the bottom structure. The energy function of GRBM is

$$E(v, h \mid \theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H}\frac{v_i}{\sigma_i}\omega_{ij}h_j + \sum_{i=1}^{V}\frac{(v_i - b_i)^2}{2{\sigma_i}^2} - \sum_{j=1}^{H}a_j h_j. \quad (16)$$

Formula (16) represents the Gaussian noise variance of the visible neurons. Due to the change of the energy function, conditional probability is also changed, which should be amended as

$$p(h_j = 1 \mid v) = \text{logistic}\left(a_j + \sum_{i=1}^{V}\omega_{ij}\frac{v_i}{\sigma_i}\right), \quad (17)$$

$$p(v_i = 1 \mid h) = N\left(b_i + \sigma_i\sum_{j=1}^{H}\omega_{ij}h_j, {\sigma_i}^2\right). \quad (18)$$

As shown in Figure 6, the input visual, hidden, and output layers are represented as blue, red, and green colorized rounds, respectively. The restructures of models show that the DBN22 input has two representations (although in practice DBNN21 and DBN22 have the same structure once the feature vectors are combined). The training process of the DBN22 network model is conducted in accordance with the following steps.

(1) The initialization of unsupervised learning is needed at the beginning of training. The initialization process is step by step completed in each layer by multiple RBM in accordance with the order of the bottom-up.

First of all, the feature vector extracted from the traditional features is considered as the visual layer of the first left side in the RBM; the spectrogram feature vector is considered as the visual layer of the first right side in the RBM. Then, the CD-1 algorithm is applied to training for the weight of each layer, denoted by LW_1 and RW_1. According to the weights obtained and the input visible layers, the weighted summations are conducted on all the input nodes. Then, the hidden layers LH_1 and RH_1 could be obtained by mapping [21].

After that, LH_1 and RH_1 are the input of visual layer in the second RBM. Also after CD-1 training, connection weight W_2 can be obtained. Then, the hidden layer H_2 is gained according to the input visual layer and weight W_2.

(2) The Deep Belief Networks are constituted. The trained RBMs in top-bottom order are overlapped layer by layer as shown in Figure 6. The uppermost level of RBM is in the form of a two-way connection while the others are connected by top-bottom.

(3) The kernel-KNNFLC classifier is added to the top of the above for classification.

(4) The network weights are fine-tuned. Before the final network parameters being obtained, the fine-tuning is necessary by the training results and BP algorithm so that the weights may be more accurate.

The training process of DBN21 model is similar to DNB22, while the bottom layer RBM only has the left half of DBN22. The data generation process of DBN is through the top RBM with Gibbs Sampling and completed by transferring from top to bottom. The Gibbs Sampling of the top RBM is divided into multiple alternate processes, which makes the sample distribution obtained balanced. Then, the data are generated by top-bottom DBN network. This way effectively saves feature information of cross-corpus samples, so as to improve the robustness of the SER system. The operation of weights adjustment is conducted after the pretraining. Then, based on the method of Error Back Propagation, the tag data are used to fine-tune the weights. This strategy searches the

weight space locally in the process of running, which could accelerate the training speed effectively.

4. Experimental Results and Analysis

4.1. Experimental Preparation

4.1.1. Settings of Experiments. In this section, the fusion experiments are divided into three parts. First of all, DBN21 model is used to carry out layer fusion for SER across the databases, in which features are traditional acoustic (see Section 4.1.3). Then, the results of the experiments are contrasted with the traditional features without DBN fusion and this experiment group is marked as *Fusion 1*. DBN21 model is then used to extract spectrogram features of layer fusion based on selective attention mechanism. Also the results are contrasted to the features without DBN fusion, which demonstrates the cross-corpus SER ability. This group of experiments is marked as *Fusion 2*. Finally the DNB22 model is proposed to fuse the traditional and spectrogram features. The experimental results are compared with Fusion 1 and Fusion 2. To prove that DBN22 possesses the significant improvement of performance on features fusion for SER, this group of experiments is marked as *Fusion 3*.

4.1.2. Database Settings. The selection of appropriate emotional databases is also an important part of speech emotion recognition. In our research, we chose one common speech emotion database: ABC (Airplane behaviors Corpus) which is recorded in German [22]. Moreover, the Deep Learning technology is suitable for the situation with a huge number of data sets, while the international classic databases usually have less samples. Meanwhile, in order to verify the effect of the fusion method proposed on Chinese speech databases, two Chinese corpora which are widely researched in China domestic are introduced and combined. The following are the brief introductions of the 3 databases, respectively.

ABC is obtained on a holiday aircraft flight in the background of prerecorded announcement played. The flight contains 13 upcoming trip scenarios and 10 return scenarios. Eight targeted passengers are chosen to get through the set condition: false meals, aircraft navigation turbulence, sleep, and conversation with the neighbors. During this process 11.5 hours of video and 431 voices with 8.4-second length are recorded. Finally the collected segments are independent analyzed by three professional researchers, and then the selected samples are labeled in accordance with the "aggressive," "amused," "excited," "strained," "neutral," and "tired" 6 kinds of emotional categories.

Chinese corpora used in our SER experiments consist of two databases which are recorded by induced and acted speech, respectively. One of them is the Chinese Database (CNDB) created by the Key Laboratory of Underwater Acoustic Signal Processing of in Southeast University. It consists of two parts: practical speech emotion database [23] and Whisper emotion database [24]. The statement materials of practical speech emotion database are recorded by performers with histrionic or broadcast experience (8 males and 8 females, aged between 20 to 30 years, without a recent cold, standard Mandarin). The recording environment is indoors quiet. In order to guarantee the quality of the emotional corpora, the subjective listening evaluation is carried out. The statements selected with more than 85% confidence coefficient are in total 1410 from the male performers and 1429 from the female performers, including six basic emotional categories: "raged," "fear," "joy," "neutral," "sad," and "surprise." The Whisper database contains "happy," "angry," "surprise," "grieved," and "quiet" such five kinds of emotions. Then, the speech materials are divided into three types: word, phrase, and long sentence. The corpus contains 25 words, 20 phrases, and 6 long sentences for each emotion category. Each speaker repeats the whispers 3 times and with normal voice for 1 time (for later comparison), forming a total of 9600 statements. The research of whispered speech database has great significance: further improving the ability of human-computer interaction, combining with the semantic to judge the inner activities of speakers, and helping computers really understand the operators' thoughts, feelings, and attitudes. The analysis and processing of the emotional characteristics of whispered speech signals have important meaning of judgment and simulation of the emotion status from speakers in theory and application.

According to the recording criterion of corpora, the two Chinese databases are merged, ultimately forming 7839-statement CNDB Chinese corpora. The recording employs mono, 16-bit quantitative, and 11.025 kHz sampling rate. The selection of statements follows two principles: (1) the statements selected do not contain a particular emotional tendency; (2) the statements must have high emotional freedom, which could exert different emotions on the same statement.

Another Chinese corpus is the Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA) [25]. The language of the database is Chinese, made by four actors. Database contains 1200 statements and is divided into 6 categories of emotions: "angry," "fear," "happy," "neutral," "surprise," and "sad."

In order to verify the effectiveness of the method proposed in this paper, in each group of experiments two kinds of schemes (*Case I and Case II*) are adopted, respectively, for testing. According to the theory of Emotion Wheel [26], mutual or similar 4 basic emotions in the three chosen databases, "angry (aggressive, raged)," "happy (joy, amused)," "surprise (excited, amazed)," and "neutral," are chosen for experimental evaluation. Because the DBN model could show the effectiveness of the fusion under the condition of large amount of data, we merge 3 Chinese speech emotion corpora into one called *Mandarin Database*. In *Case I*, Mandarin Database is as the training data set (known label), while ABC (unknown label) is as the testing set. The cross-corpus SER of this scheme adopts rotation experiment testing method: the data set is divided into 10 portions, in which the proportion of training/testing is 9 : 1. The set of this 10-fold cross-validation is intended to optimize the parameters within the source corpus [10]. After the cross-validation, the averages are obtained as the results for the cross-corpus experiments. In *Case II*, ABC corpus (known label) is as the training set, while Mandarin Database (unknown label) is as the testing

set. Because the number of samples in German ABC corpus is less, for the sample balance of corpus in the process of SER evaluation, we join a part of the Mandarin samples (45% is the optimum by testing, known label) into ABC samples which are as the training set. Then the remaining Mandarin samples (55% corpora, unknown label) are as the testing set.

4.1.3. Settings of Feature Parameters and Classifier. With regard to the *traditional acoustic global features*, the common tool openSMILE is used for feature extraction whose tool number is set as 1 [27]. Then the feature set in the Interspeech 2010 SER competition [28], which contains a total of 1582 dimensions features, is introduced in our experiments. 38 acoustic low-level descriptors (LLDs) and their first-order differences are contained in the set. Through statistics of 21 class functions on the LLDs (16 features with 0 information are removed), we add the numbers and lengths of F0 to the feature set. In the contrast group without fusion, the feature sets extracted directly carried out the LDA dimension reduction, making its dimension match the fusion experimental group.

In this paper, the recognition experiment employs kernel-KNNFLC classifier, which may verify the SER ability of the fusion features. According to the gravity center criterion, Kernel-KNNFLC learns the sample distances and improves the K-neighbor with kernel learning method. The classifier optimizes the differentiation between kinds of the emotional feature vectors, which solves the problem about huge calculation caused by the features of prior samples. Based on the cross-corpus samples trained, the recognition model is established and then the different emotional categories are distinguished. The Gaussian Radial Basis kernel Function (RBF) is used in the classifier: $k(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$, in which $\sigma = 4$. The KNNFLC classifier based on kernel has stable SER performance on high-dimensional data. In addition our experiments use 4 kinds of speech emotions, so the dimension dropped to 3 for achieving the best recognition rate. This is due to the solving of the generalized eigenvalue principle: the optimization is achieved when the minimum number of features is solved. With the K-nearest neighbor algorithm based on feature line centroid, the kernel function of RBF is improved and the optimum value is $K = 3$ [13].

4.2. Traditional Global Acoustic Feature Fusion Experiment (Fusion 1). The purpose of *Fusion 1* is comparing the fusion feature with DBN21 to features without DBN, so that the cross-corpus recognition performance of fusional traditional features could be revealed. The extracted acoustic features are as the input of DBN21 model. Then the optimization process is carried out by DBN. After that, combined with the kernel-KNNFLC classifier introduced before, the emotion recognition missions are proceeded on cross data sets.

The settings of RBM learning rate should be moderate, because too big or too small rates will both increase the reconstruction error. GRBM learning rate of bottom layer in *Fusion 1* is set as $\varepsilon = 0.001$ and training cycle is set as $J = 200$. The upper layer RBM is set as $\varepsilon = 0.01$ and $J = 70$. Since the numbers of visible layer unit and input dimension are the same, the input units number of visible layer in the experiment is $n_v = 60$ and the number of upper hidden units is set as $n_h = 20$. The weight is set according to the Gaussian random vector $N = (0, 0.01)$. The visible and hidden unit biases are $a_i = 0$ and $b_j = 0$. Because the number of hidden units in the middle layer may influence systemic performance, therefore we enumerate the 6 units' numbers of contrast experiments: 250, 500, 750, 1000, 1250, and 1500, in order to determine the optimal number of hidden units. The experimental comparison results are shown in Figure 7.

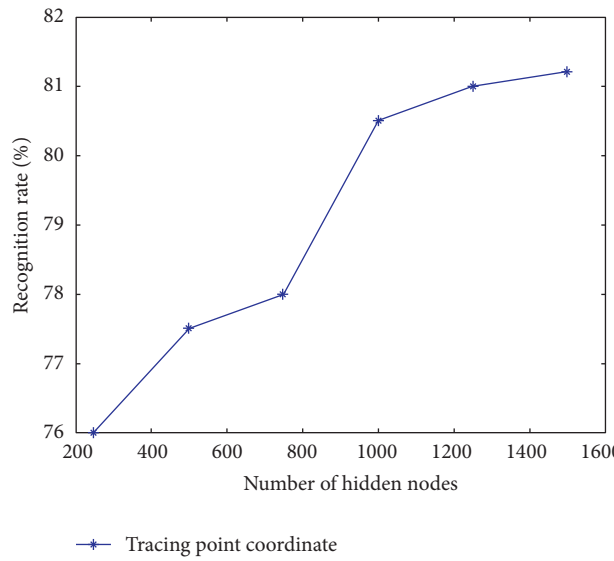


Figure 7: The influence of hidden node numbers to recognition rates *Fusion 1*.

Figure 7 reveals that, along with the increase of the hidden nodes number, the recognition efficiency of system is growing. However, the increased number of nodes may cause the extra amount of calculation. It is clear that when the nodes number rises from 750 to 1000, the recognition rate has greatly improved, and then it is steady. Hereby considering the time consumed and accuracy, the number of hidden nodes in *Fusion 1* is set as 1000.

The speech emotion recognition experiments are carried out through the DBN21 model proposed. In our testing strategy, the ABC and Mandarin databases are cross-validated, which is to verify the robustness of algorithm proposed under the cross-corpus SER task. Toward each kind of emotion in two cross-database cases, recognition rates of the traditional features which are before and after fusion are shown in Table 1.

The experimental results indicate the traditional features after optimization by DBN. The emotion recognition ability has greatly ascended, which rises by 4.6% on the average recognition rate. It reveals that the DBN model proposed in feature layer is effective for SER feature fusion research. After training on ABC and Mandarin databases in *Case I*, SER rates of ABC testing set on Mandarin training set reach 52.2%. Among them "happy" and "neutral" reach over 63% as the highest, whose recognition effect is superior to *Case II*. It related to the many similar types of samples in

Table 1: Recognition results (%) of *Fusion 1* in two cross-corpus cases.

Cross-corpus scheme	Happy	Angry	Surprise	Neutral	Average
Case I					
DBN feature fusion	63.7	38.6	41.6	64.9	52.2
Without fusion	52.8	34.3	39.7	63.7	47.6
Case II					
DBN feature fusion	57.0	37.5	40.8	62.7	49.5
Without fusion	50.1	29.7	35.3	60.4	43.9

3 training corpora. Through the great amount of emotional data training in various categories by Deep Learning, the model becomes highly mature while the matching degrees of the traditional emotional categories with high inner-class discrimination ("happy," "neutral") are high. The comparison of two experiment schemes shows that the small amount of training samples in ABC gives rise to information insufficiently. Thus, further testing in large data corpus may cause undermatching with model.

4.3. Spectrogram Feature Fusion Experiment (Fusion 2). Main aim of *Fusion 2* is to validate the feature effectiveness of spectrogram on cross-corpus. The feature sets abstracted are based on selective attention mechanism introduced in this paper. In order to reflect the promotional recognition performance on cross-databases, the experimental results after DBN21 fusion are compared with traditional features in *Fusion 1*.

The same as Fusion 1, GRBM learning rate of bottom layer in Fusion 2 is set as $\varepsilon = 0.001$ and training cycle is set as $J = 200$; the upper layer RBM is set as $\varepsilon = 0.01$ and $J = 70$. But the input units number of visible layer here is $n_v = 291$ and the number of upper hidden units is set as $n_h = 80$. The weight is set also as $N = (0, 0.01)$; meanwhile, the visible and hidden unit biases are $a_i = 0$ and $b_j = 0$. In consideration, the number of hidden nodes in layer RBM may cause the influence of system performance; this experiment still needs the discussion of the numbers of hidden nodes. The analysis of node numbers in traditional features is as in Figure 8.

As shown in the relationship in Figure 8, it is different from the traditional feature experiment; the recognition efficiency of spectrogram features has greatly promoted at 750 hidden nodes of point position. This is due to the traditional acoustic features compared to the spectrogram ones, which possess much higher input dimensions, so that, in the spectrogram feature fusion experiments, the number of hidden nodes in bottom RBM is set to 750.

According to the SER fusion model in feature layer, which is based on selective attention as shown in Figure 1, the cross-corpus experiments are carried out. In *Fusion 2*, ABC and Chinese databases are crossed training for cross-corpus testing. The SER confusion matrix in *Case I* by DBN21 fusion model is as shown in Figure 9.

It could be seen from the experimental results that the spectrogram features extracted integrally have strong ability of speech emotion recognition. When compared to traditional features, the spectrogram exhibits advantages in

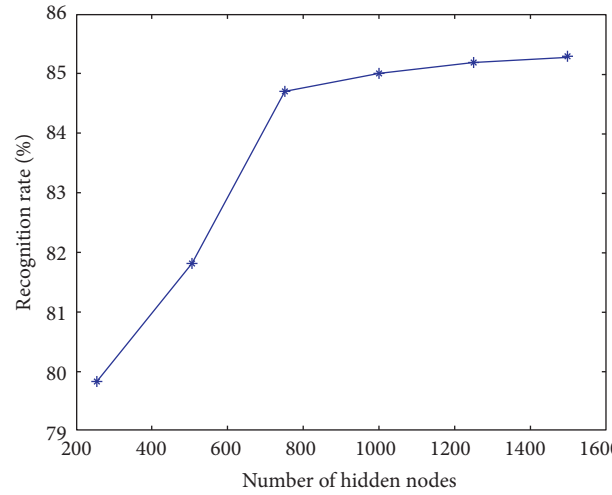


Figure 8: The influence of hidden node numbers to recognition rates in *Fusion 2*.

Figure 9: Cross-corpus SER confusion matrix in *Fusion 2* with *Case I*.

dealing with the cross data set tasks. This is because the traditional features contain only the local traits of common speech processing field, while the spectrogram, abstracted from the aspects of time and frequency domains, contains information between adjacent frames and the temporal features which can make up for a lack of global features. In spectrogram features, meanwhile, compared to traditional global features, the cascade vectors possess higher dimensions which contain more information for characterizing the emotions. Among them, the "happy," "angry," and experimental results improve significantly compared with the traditional fusion features. It reveals that spectrogram features have a relatively better distinction effect on the emotion category with high frequency domain correlation dependence.

4.4. DBN22 Feature Fusion Experiment on Cross-Corpus (Fusion 3). In experiment *Fusion 3* with DBN22 model, we conduct feature layer fusion of traditional global acoustic features and spectrogram features based on selective attention. After that the kernel-KNNFLC is combined with SER system for cross-corpus experiments. This method integrates image characteristics and acoustic characteristics, which is a novel

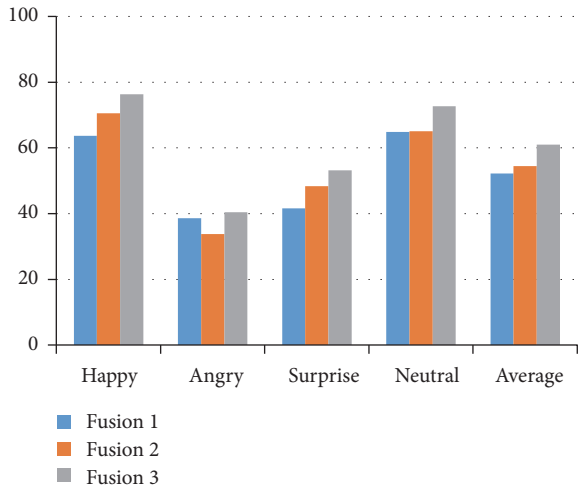


Figure 10: Recognition rates of 3 fusion models.

new attempt for data source extension in the field of speech emotion recognition. At the same time, the experiment may demonstrate that the features with thus fusion method have significant help for improving SER performance in cross-databases.

The settings of RBM in *Fusion 3* are as follow: GRBM learning rate of bottom layer is set as $\varepsilon = 0.001$ and training cycle is set as $J = 200$. The upper layer RBM is set as $\varepsilon = 0.01$ and $J = 70$. Since the numbers of visible layer unit and input dimension are the same, thus the input units numbers of visible layer in acoustic and spectrogram features are $Ln_v = 291$ and $Rn_v = 60$, respectively. The number of upper hidden units is set as $n_h = 100$. The weight is set according to the Gaussian random vector $N = (0, 0.01)$. The visible and hidden unit biases are $a_i = 0$ and $b_j = 0$. Because the number of hidden units in the middle layer may influence systemic performance, according to the two Fusion experiment advance, hidden units numbers in RBM of acoustic and spectrogram features are 1000 and 750, respectively.

After cross-database SER experiment, fusion features of traditional acoustic and spectrogram features are gained based on DBN22 network. Then the recognition results are compared with DBN21 groups in *Fusion 1* and *Fusion 2* by the bar plot (using *Case I* cross-corpus settings) (see Figure 10).

After the analysis of Figure 10, the cross-database recognition efficiency of the fused features in *Fusion 3* is the highest. Specifically "happy," "angry," "surprise," and "neutral" 4 emotional kinds compared with the traditional group rise by 12.6%, 1.8%, 11.6%, and 12.6%, respectively; the promotion of average recognition rate is 8.8%. Relative to the spectrogram features, the results increase by 5.8%, 5.8%, 6.6%, and 5.8%, respectively, and the elevation of average recognition rates up to 6.5%. The fusion by DBN22 of two kinds of features obtains excellent recognition effect in all of the emotion categories. The results benefit from the optimization of feature fusion layer in RBM stack of the DBN network, while there are also the factors of classifier and network parameters' settings. Experiments show that the DBN network model proposed successfully gains the fused features of traditional acoustic characteristics and the information of spectrogram images, which meanwhile effectively improve the cross-corpus efficiency of the SER system.

5. Conclusion

This paper mainly researches the feature layer fusion model on the strength of DBN for speech emotion recognition. First of all, based on the mechanism of selective attention, the system extracts three kinds of spectrogram features with both temporal information and global information, which are used for cross-corpus SER. The spectrogram features introduced solve the problem of information loss by the traditional feature selection methods. Further, it is a supplement to the types of emotional information under the cross-database. Then, the modified DBN models are proposed to reasonably optimize the high-dimension spectral features, to retain the useful information and to improve the robustness of cross-corpus SER system. In the subsequent simulation experiments, the DBN21 and DBN22 models designed are used in the feature layer to fuse the spectrogram and traditional acoustic traits. Furthermore, the experimental results are compared with those of the benchmark models. Through experiments in cross-databases containing three Chinese ones and a general German one, DBN networks with multilayer RBM are proved as robust feature layer fusion models for cross-corpus. Spectrogram traits, at the same time, are validated conducive to boost emotional distinguish ability after feature fusion. In this paper, on the basis of Deep Learning thought, the DBN22 model proposed effectively fuses the spectrogram and traditional acoustic emotion features. This progress realizes the features fusion of various data sources and provides a new direction for further research of SER in cross-corpus.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work has been supported by the national natural science foundation of china (NSFC) under Grants nos. 61231002 and 61375028.

References

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[2] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.

[3] E. Marchi, A. Batliner, B. Schuller et al., "Speech, emotion, age, language, task, and typicality: trying to disentangle performance and feature relevance," in *Proceedings of the International Conference on Privacy, Security, Risk and Trust (PASSAT '12) and International Conference on Social Computing (SocialCom '12)*, pp. 961–968, 2012.

[4] C. Parlak, B. Diri, and F. Gürgen, "A cross-corpus experiment in speech emotion recognition," in *Proceedings of the International Workshop on Speech, Language and Audio in Multimedia (SLAM '14)*, pp. 58–61, Penang, Malaysia, 2014.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features," in *Proceedings of the IEEE 9th International Workshop on Multimedia Signal Processing (MMSP '07)*, pp. 48–51, Crete, Greece, October 2007.

[7] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech '14)*, pp. 223–227, Singapore, September 2014.

[8] H. Lee, C. Ekanadham, and Y. Ng A, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems*, pp. 873–880, 2008.

[9] V. Nair and G. E. Hinton, "Rectified linear units improve Restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, pp. 807–814, June 2010.

[10] B. Schuller, Z. Zhang, F. Weninger et al., "Selecting training data for cross-corpus speech emotion recognition: prototypicality vs. generalization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10 '11)*, pp. 807–814, 2011.

[11] T. A. Lampert and S. E. M. O'Keefe, "On the detection of tracks in spectrogram images," *Pattern Recognition*, vol. 46, no. 5, pp. 1396–1408, 2013.

[12] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[13] X. Zhang, C. Zha, X. Xu, P. Song, and L. Zhao, "Speech emotion recognition based on LDA+kernel-KNNFLC," *Journal of Southeast University (Natural Science Edition)*, vol. 45, no. 1, pp. 5–11, 2015.

[14] O. Kalinli and R. Chen, "Speech syllable/vowel/phone boundary detection using auditory attention cues," Google Patents, 2014.

[15] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77–123, 2003.

[16] C. Stevens, B. Harn, D. J. Chard, J. Currin, D. Parisi, and H. Neville, "Examining the role of attention and instruction in at-risk kindergarteners: electrophysiological measures of selective auditory attention before and after an early literacy intervention," *Journal of Learning Disabilities*, vol. 46, no. 1, pp. 73–86, 2013.

[17] G. Evangelopoulos, A. Zlatintsi, A. Potamianos et al., "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[18] A. Smith, A. Doucet, N. de Freitas et al., *Sequential Monte Carlo Methods in Practice*, Springer Science & Business Media, 2013.

[19] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[20] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pp. 5060–5063, Prague, Czech Republic, May 2011.

[21] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, vol. 76, pp. 112–126, 2016.

[22] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. II-733–II-736, Honolulu, Hawaii, USA, April 2007.

[23] C. Huang, Y. Jin, Y. Zhao et al., "Design and establishment of practical speech emotion database," *Acoustic Technologies*, vol. 29, no. 4, pp. 396–399, 2010 (Chinese).

[24] Y. Jin, Y. Zhao, C. Huang et al., "The design and establishment of a Chinese whispered speech emotion database," *Technical Acoustics*, no. 1, pp. 63–68, 2010.

[25] Institute of Automation Chinese Academy of Sciences, The selected Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA), 2010, http://www.chineseldc.org/resource_info.php?rid=76.

[26] T. Bänziger, V. Tran, and K. R. Scherer, "The Geneva Emotion Wheel: a tool for the verbal report of emotional reactions," in *Poster Presented at ISRE*, vol. 149, pp. 149–271, Bari, Italy, 2005.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*, pp. 1459–1462, Firenze, Italy, October 2010.

[28] B. Schuller, S. Steidl, A. Batliner et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings of the International Speech and Communication Association (INTERSPEECH '10)*, pp. 2794–2797, Makuhari, Japan, 2010.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 2013658, 8 pages
http://dx.doi.org/10.1155/2016/2013658

Research Article

Experiments on Detection of Voiced Hesitations in Russian Spontaneous Speech

Vasilisa Verkhodanova and Vladimir Shapranov

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia

Correspondence should be addressed to Vasilisa Verkhodanova; interiora@gmail.com

Received 31 July 2016; Accepted 31 October 2016

Academic Editor: Alexander Petrovsky

The development and popularity of voice-user interfaces made spontaneous speech processing an important research field. One of the main focus areas in this field is automatic speech recognition (ASR) that enables the recognition and translation of spoken language into text by computers. However, ASR systems often work less efficiently for spontaneous than for read speech, since the former differs from any other type of speech in many ways. And the presence of speech disfluencies is its prominent characteristic. These phenomena are an important feature in human-human communication and at the same time they are a challenging obstacle for the speech processing tasks. In this paper we address an issue of voiced hesitations (filled pauses and sound lengthenings) detection in Russian spontaneous speech by utilizing different machine learning techniques, from grid search and gradient descent in rule-based approaches to such data-driven ones as ELM and SVM based on the automatically extracted acoustic features. Experimental results on the mixed and quality diverse corpus of spontaneous Russian speech indicate the efficiency of the techniques for the task in question, with SVM outperforming other methods.

1. Introduction

Speech technologies are often developed for different types of speech and rarely for a spontaneous one. However, almost all speech we produce and comprehend every day is spontaneous. This type of oral communication is likely to be one of the most difficult forms of speech communication among people: during very dense time interval speaker has to solve several laborious cognitive tasks. One has to form the utterance and to choose the exact linguistic form for it by selecting words, expressions, grammatical forms, and so on. This process leads to different flaws in spontaneous speech production, so-called speech disfluencies. These are self-repairs, repetitions, voiced hesitations (filled pauses and lengthening that are often referred together as FPs), slips of the tongue, and other breaks or irregularities that occur within the flow of otherwise fluent speech. These phenomena indicate the mental processes of underlying speech generation and have been viewed as a sign of word-searching problem [1] or difficulties in conceptualization at major discourse boundaries [2]. There is evidence that they can affect up to one-third of utterances [3]; for example, in conversational speech in American English, about 6 per 100 words are disfluent [3, 4].

In Russian speech filled pauses occur at a rate of about 4 times per 100 words and at approximately the same rate inside clauses and at the discourse boundaries [5]. Though evidence on filled pauses differs across languages, genres, and speakers, on average there are several filled pauses per 100 syllables [6]. They are also the most frequent speech disfluencies; filled pauses occur more often than any other speech disfluencies (repetitions, word truncations, etc.) [6], signalling not only of breaks in speech production process, but also of explication of this process [5]. According to [7] in the conversational Switchboard database [8], about 39.7% of the all disfluencies contain a filled pause. In the corpus of Portuguese lectures LECTRA filled pauses correspond to 1.8% of all the words and to 22.9% of all disfluency types being the most frequent type in the corpus [9].

The need in coping automatically with speech disfluencies appeared along with the need of spontaneous speech processing, which brought up a lot of interesting challenges to speech science and engineering. Once seen as errors, along with

other disfluencies, hesitations were acknowledged as integral part of natural conversation [5, 10]. They may play a valuable role such as helping a speaker to hold a conversational turn or expressing the speakers' thinking process of formulating the upcoming utterance fragment [10–12]. The comparison of prosodic patterns of stutterers and nonstutterers disfluencies was done in [13], where authors analysed spontaneous story-telling of 8 people: four stutterers and four nonstutterers. Results, as expected, showed that stutterers have significantly more disfluencies than nonstutterers and that disfluencies affected the adjacent tonal contexts and phrasing differently in these two groups of people, stutterers' disfluencies being accompanied by more prosodic irregularities. Details can be found in [13]. Thus, the detection of vowel lengthening and filled pauses could be an important step towards locating the disfluent regions and evaluating the spoken fluency skills of a speaker.

The problem of detecting hesitations has been addressed from various perspectives. In computational linguistics analysis of speech disfluencies is sometimes incorporated into syntactic parsing and language comprehension systems [14] and more often into automatic speech recognition systems [15]. Hesitations, as well as other speech disfluencies, were always an obstacle for automatic processing of spontaneous speech as well as its transcriptions; disfluencies are known to have an impact on ASR results; they can occur at any point of spontaneous speech; thus they can lead to misrecognition or incorrect classification of adjacent words [9, 10, 16, 17].

Hesitations exhibit universal as well as linguistic and genre specific features. Filled pauses and lengthenings are represented mainly by vocalizations with rare cases of prolonged consonants (which was shown to be a peculiarity of Armenian hesitational phenomena [18]). These vocalizations are usually phonetically different from the lexical items, since they are pronounced with minimal movements of the articulatory organs due to the articulatory economy [19]. However, it was also shown that phonological system of the language may influence the quality of FPs vocalizations [20]. Even universal characteristics of hesitations, such as lengthenings being accompanied by creaky voice, may operate differently in different languages; for example, in Finnish it was proposed that creaky voice may indicate turn-transitional locations [12], which is not the case for English [21].

Although the speech technologies and particularly ASR systems have to account for all types of disfluencies (filled pauses, lengthenings, repetitions, deletions, substitutions, fragments, editing expressions, insertions, etc.), in the present study, we focus on the detection of the most frequent disfluent category: voiced hesitations (filled pauses and sound lengthening) in Russian spontaneous speech.

2. Related Work

Various methods have been proposed for speech disfluencies detection. All of them can be roughly divided into the following: (1) those that use language modelling (LM) incorporating information on speech disfluencies into ASR systems and (2) those that take into account only acoustic parameters. The second group is more popular, since there is no need of additional large corpus of transcriptions for LM training, despite the possible way of dealing with the problem by including the filler as an ordinary word in the lexicon and ignoring it during LM-probability computation [22]. Although this inclusion may sound reasonable, it does not necessarily lead to a higher accuracy; too many filled pauses may be hypothesized due to the acoustic similarity between filled pauses and function words or single syllables of content words [23].

It has been shown that, along with duration, the prominent characteristic of voiced hesitations is a gradual fall of fundamental frequency ($F0$) [24]; they tend to be low in $F0$ and display a gradual, roughly linear $F0$ fall. In [25] it was shown that for fair detection of hesitations these two characteristics and distance to a pause are enough.

In [17], filled pauses are detected on a basis of two features (small fundamental frequency transition and small spectral envelope deformation) which are estimated by identifying the most predominant harmonic structure in the input. The method has been implemented and tested on 100 utterances extracted from a Japanese spoken language corpus. Each utterance contained at least one filled pause. The achieved results were 91.5% precision and 84.9% recall. However, the authors admit that these figures may be optimistic because in their corpus there were no low-voiced male speakers.

In [23] authors developed a detection system in order to improve the speech recognizer performance. As a classifier authors used the Multilayer Perceptron with one output. The features were segment duration, spectral stability, stable interval durations, silence before and after the hesitations, spectral centre of gravity, and simple filled pause model output (a 4-mixture GMM that was trained to model the frames belonging to a filled pause). On three Flemish parts of Spoken Dutch Corpus authors achieved precision of 85% at a recall rate of 70%.

In [9] authors focused on detection of filled pauses based on acoustic and prosodic features as well as on some lexical features. Experiments were carried on a speech corpus of university lectures in European Portuguese, LECTRA. Several machine learning methods have been applied, and the best results were achieved using Classification and Regression Trees. The performance achieved for detecting words inside of disfluent sequences was about 91% precision and 37% recall, when filled pauses and fragments were used as a feature; without it the performance decayed to 66% precision and 20% recall. Further experiments on filled pauses detection in European Portuguese were carried out using prosodic and obtained from ASR lexical features; the best results were achieved using J48, corresponding to about 61% F-measure [26].

In 2013 the INTERSPEECH Paralinguistic Challenge [27] raised interest in automatic detection of fillers providing a standardized corpus and a reference system. The winners of the Social Signals Sub-Challenge introduced a system, built upon a DNN classifier complemented with time series smoothing and masking [28]. In [29] authors presented a method for filled pauses detection using an SVM classifier, applying a Gaussian filter to infer temporal context information and performing a morphological opening to filter false

alarms. For the feature set authors used the same as was proposed for [27], extracted with the openSMILE toolkit [30]. Experiments were carried out on the LAST MINUTE corpus of naturalistic multimodal recordings of 133 German speaking subjects in a so-called Wizard-of-Oz (WoZ) experiment. The obtained results were recall of 70%, precision of 55%, and AUC of 0.94.

3. Material

Usually for studying speech disfluencies researchers use corpora with Rich Transcription [31]. An example of such corpora is English CTS Treebank with Structural Metadata corpus of English telephone conversations with metadata annotation [32], which includes, for example, filled pauses and discourse markers. Another example is the corpus Czech Broadcast Conversation MDE Transcripts [33] that consist of transcripts with metadata of the files in Czech Broadcast Conversation Speech Corpus [34]. Its annotation contains such phenomena as background noises, filled pauses, laugh, smacks, and so on [35].

For our purposes we combined different material of diverse quality and recordings situation. Thus, the material we used in this study consists of several parts.

The first part is the corpus of task-based dialogues collected at SPIIRAS in St. Petersburg in the end of 2012-beginning of 2013 [36]. Thus, the recorded speech is informal and unrehearsed, and it is also the result of direct dialogue communication, what makes it spontaneous [37]. For example, in Edinburgh and Glasgow the HCRC corpus was collected, which consists only of map-task dialogues [38], and half of the other corpus, corpus of German speech Kiel, consists of appointment tasks [39]. This corpus consists of 18 dialogues from 1.5 to 5 minutes, where students (6 women and 6 men) from 17 to 23 years fulfilled map and appointment tasks in pairs. Recordings were annotated manually into different types of disfluencies, the voiced hesitations being the majority, 492 phenomena (222 filled pauses and 270 lengthenings).

For the second part of our material we used part of Multi-Language Audio Database [40]. This database consists of approximately 30 hours of sometimes low quality, varied and noisy speech in each of three languages, English, Mandarin Chinese, and Russian. For each language there are 900 recordings taken from open source public web sites, such as http://youtube.com/. All recordings have been orthographically transcribed at the sentence/phrase level by human listeners. The Russian part of this database consists of 300 recordings of 158 speakers (approximately 35 hours). The casual conversations part consists of 91 recordings (10.3 hours) of 53 speakers [40]. From this Russian part we have taken the random 6 recordings of casual conversations (3 female speakers and 3 male speakers) that were manually annotated into hesitations. The number of annotated phenomena is 284 (188 filled pauses and 96 sound lengthenings).

The third part is the corpus of scientific reports from seminar devoted to analysis of conversational speech held at SPIIRAS in 2011. Recordings of reports of 6 people (3 female and 3 male speakers) were manually annotated into speech disfluencies. Since speakers did not base their reports on a written text, these recordings contain considerable amount of speech disfluencies. 951 hesitations were manually annotated: 741 filled pauses and 210 lengthenings.

Another part we added for making our corpus more quality and situation diverse is the records from the appendix No5 to the phonetic journal "Bulletin of the Phonetic Fund" belonging to the Department of Phonetics of Saint-Petersburg University [41]. The 12 recorded reports concerned different scientific topics (linguistics, logic, psychology, etc.). They were all recorded in 70s–80s in Moscow except one that was recorded in Prague. All speakers (6 men and 6 women) were native Russian speakers and were recorded while presenting on conferences and seminars. The number of manually annotated hesitations is 285 (225 filled pauses and 60 lengthenings).

In total, the data set we used is about 3 hours and comprises 2012 filled pauses. Distribution of hesitations duration over the corpus is shown in Figure 1.

Distribution of ten most frequent hesitations across different parts of the joint corpus is shown in Figure 2.

The duration of a single hesitation lies between 6 ms and 2.3 s; the average duration is 388 ms. Among annotated hesitation the most frequent filled pause was [əː] with total 905 utterances, and the most frequent lengthening was that of vowel /a/ - 197 utterances.

4. Experiments on Hesitations Detection in Russian with Machine Learning Techniques

To develop a good hesitations detector a proper set of prosodic and acoustic cues that are likely to mark hesitations in speech signal is needed. As it was already mentioned above, in [25], it was shown that for fair detection of hesitations these two characteristics and distance to a pause are enough. Thus, at first we started testing rule-based approaches towards hesitation detection.

4.1. Rule-Based Approaches towards Hesitations Detection in Russian. The pilot step of experiments was to try a similar simple [25] approach on Russian speech, based our method on acoustical features of voiced hesitations that are peculiar to these events in Russian. To find the most prominent ones we have checked duration, $F0$, three first formants, energy, and stableness of spectra across the corpus. Similar approaches have been applied for filled pauses detection in other languages and proved the relevancy of these acoustic properties [16, 17, 42]. As a result, we used standard deviations of $F0$, $F1$, and energy as parameters, since they showed smaller variance in hesitations (the smallest were for $F0$ and energy (Figure 3)).

We obtained the optimal values of parameters a and b for criterion

$$C = aX + bY < 1, \tag{1}$$

where X is standard deviation of logarithm of $F0$ – std(log($F0$)) and Y is standard deviation of logarithm of energy E. The optimal values are those that maximize $F1$-score for the task of selection of 150 ms windows that are part of the

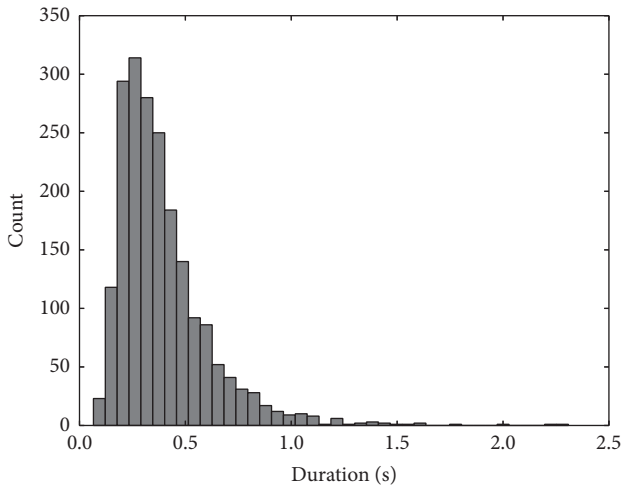


Figure 1: Distribution of hesitation duration over the joint corpus.

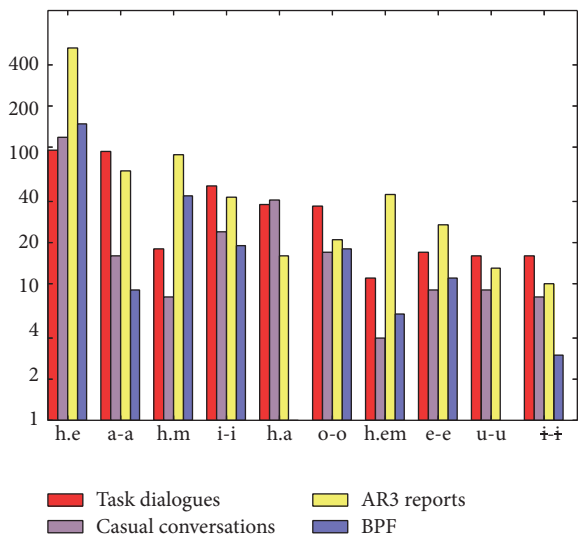


Figure 2: Distribution (in log scale) of the top ten frequent hesitations across the parts of joint corpus, where h.e, h.m, h.a and h.em are filled pauses (of [əː], [ɐː], [mː], and [əːmː] types, resp.), and others are lengthenings of certain vowels (/a/, /i/, /o/, /e/, /u/ and /i/).

hesitations (Figure 4), the standard deviation of $F1$ logarithm – std(log($F1$)) being the additional threshold.

The experiments were conducted on 85% of the corpus with 15% used as a test set. The obtained $F1$-score was 0.41.

Then we have changed the criterion and maximization process. We obtained the optimal values of parameters w_n and $E0$ for criterion

$$C = \sum_n w_n V_n < 1, \qquad E > E0, \tag{2}$$

where w_n are weights for values V_n: standard deviations of log(E) and log(F_N); and $E0$ is a minimal mean energy level. The maximization of the $F1$-score for hesitations detection was made by the gradient descent method [43]. This gave us $F1$-score of 0.46 [44].

For both these approaches the stage of comparison with annotation was the same. At first we found the intervals intersecting with the labeled ones. Then we calculated the intersection length

$$T_{\text{int}} = \text{len}\,(I \cap L) \tag{3}$$

and length of nonmatching part of the interval

$$T_{\text{ext}} = \text{len}\,(L \cap I)\,, \tag{4}$$

where I is interval and L is label. If $T_{\text{int}} > 0.2\text{len}(L)$ and $3T_{\text{ext}} < T_{\text{int}}$, the pair of label and interval is considered matching. After processing the whole signal the amount of nonmatched intervals was considered false positive count and the amount of nonmatched labels was considered false negative count.

For these two approaches misses were mainly caused by the disorder of harmonic components in hoarse voice and the laryngealized filled pauses and lengthenings. In some cases hesitation had an unstable expressive intonation contour, which was not flat or lowering, that it can be argued whether they are hesitations or interjections. Few cases of misses were the result of small duration of annotated phenomena. And noises (especially in the part from the open source multilanguage database) and overlaps (in task dialogues part) caused number of false negatives.

Thus, instead of accounting for all these phenomena in the rule-based methods, we decided to employ the data-driven approaches.

4.2. Data-Driven Approaches towards Hesitations Detection in Russian. In [45] we described experiments on hesitations detection using the Extreme Learning Machines (ELM), a particular kind of Artificial Neural Networks that solve classification and regression problems. We used the Python ELM implementation described in [46]. In our method the number of sigmoid neurons was 600. The feature set used in these experiments consisted of 21 standard deviations (for $F0$ and first three formants, energy, voicing probability and its derivative, and 14 MFCC coefficients) and of 3 mean values (for energy, voicing probability, and its derivative). The formants value was taken from Praat [47] and all other parameters from openSMILE [30]. Within each 100 ms window we calculated standard deviation for every parameter from the feature set and mean value for energy.

To create train and test sets out of the data we selected random 10% of the data for test set, and the rest was used as the train set. This operation was performed 10 times producing 10 different pairs of train and test sets. The data has been separated into two classes: FPs and Other, and since they were not balanced we downsampled the train set to avoid the bias towards the class Other [29]. This resulted in creating the subset containing randomly chosen 8% of the instances of the class Other and all the hesitation FPs data. We used this downsampled training set to train the classifier. ELM method yields a real number for every sample that was classified as a hesitation event if this number exceeded a certain threshold.

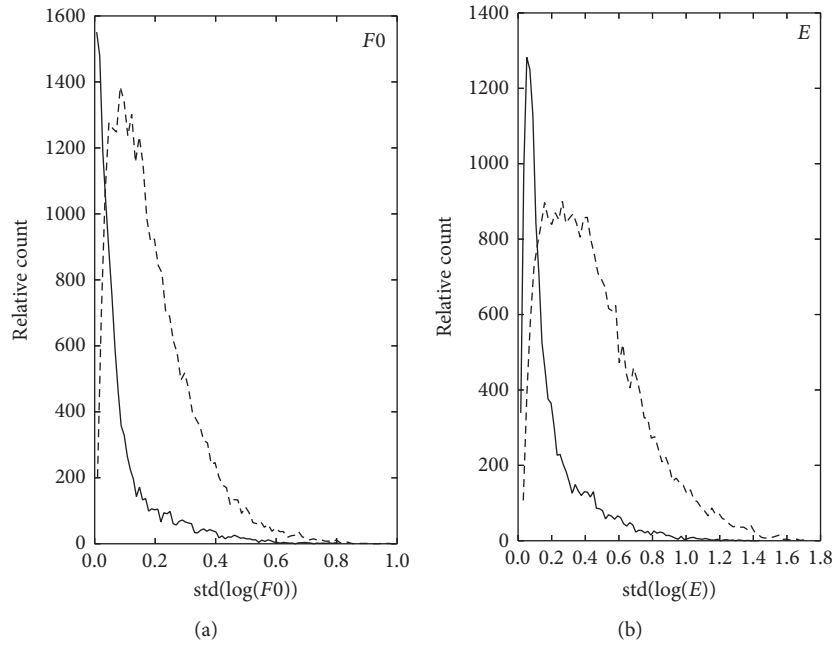


Figure 3: The standard deviation of the logarithms of $F0$ (a) and energy (b) of FPs (thick line) and of neighbouring words and phrases (dashed line).

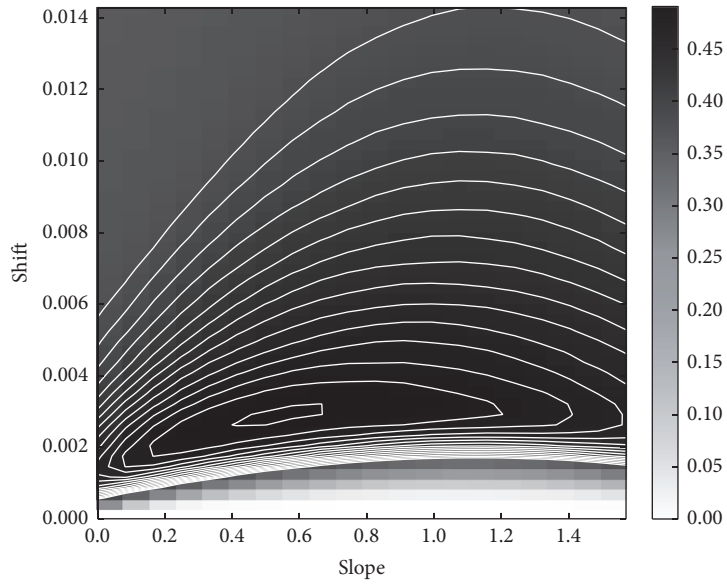


Figure 4: The $F1$-score dependence on C criterion parameters a and b, where *slope* is $\arctan(a/b)$ and *shift* is $\sqrt{a^2 + b^2}$.

This threshold was determined by a grid search in a way maximizing the $F1$-score on the training set. As the result we achieved $F1$-score of 0.42.

Our most recent experiments [48] are based on the Support Vector Machine (SVM) classifier, as we followed [29]. Compared with ELM, SVM provides better detection accuracy with better harmonized mean of precision and recall. For the experiments with SVM we used a Scikit-Learn Python library [49] implementation of SVM with polynomial kernel that enables the probability estimates by means of C-Support Vector Classification; the implementation is built upon LibSVM [50].

The feature set is based on the set that was used for the INTERSPEECH 2013 Social Signals Sub-Challenge [27]. Features were extracted with the openSMILE toolkit [30] on the frame-level basis (25 ms window, 10 ms shift). This set is derived from 54 low-level descriptors (LLDs): 14 mel-frequency cepstral coefficients (MFCCs), logarithmic energy

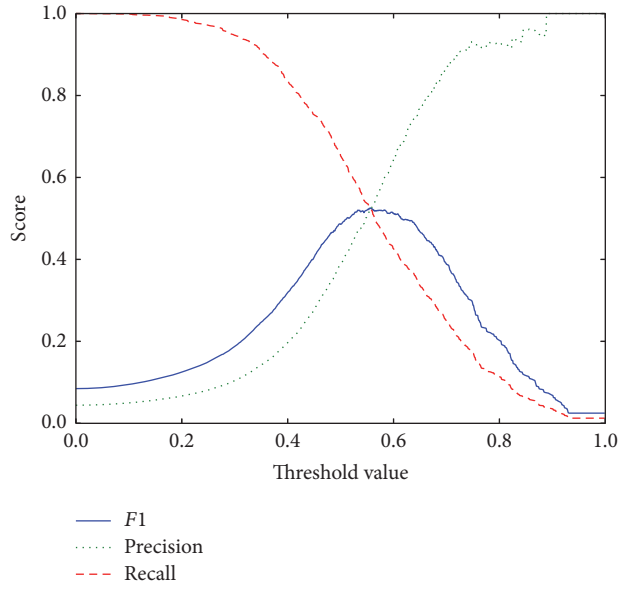


Figure 5: The dependence of results from the decision threshold.

as well as their first- and second-order delta, and acceleration coefficients; there are also voicing probability, $F0$, and zero-crossing rate, together with their deltas. For each frame-wise LLD the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four before and four after) are used as the actual features. As a result, we have 162 values per frame.

As in [45] we also separated our data into two classes: "FPs" and "Other," but changed the process of separation. Each 10th file was selected for train set, then again each 10th, for development set, and the rest was used as the test set. This operation was performed 10 times to produce 10 different triplets of train, development, and test sets.

After training our SVM classifier, as the postprocessing step we applied Gaussian filter and morphological opening [29, 51] that proved to be reasonably efficient for improving both precision and recall rates due to the usage of contextual information. Both these techniques are applied in the signal and image processing tasks for noise removal. Gaussian filter is used to smooth the spikes and remove the outliers on the probability estimates, while morphological opening is useful for making the detection of hesitations more balanced by filtering false alarms and improving $F1$-score [29]. The parameters for Gaussian and morphological opening, as well as the decision threshold, were determined using grid search on the development set.

The Gaussian filter allows us to achieve 12% improvement for $F1$-score (precision rate improving by 17% and recall rate by 5%). Morphological opening gave us only 2% improvement for $F1$-score, precision, and recall, reducing false alarm rate. The example of dependence of results from varying decision threshold on SVM output is shown in Figure 5.

As a result we achieved $F1\text{-score} = 0.54 \pm 0.027$, with precision and recall being 0.55 ± 0.05 and 0.53 ± 0.04, respectively [45]. Measures on the test set are reported in terms of mean and standard deviation over the ten evaluations using classifiers trained on ten training subsets.

The ongoing experiments are concerned with the broadening of the features set for SVM classifier by adding 4 formants with 2 derivatives for each, their standard deviations as well as means and standard deviations of their contexts, which gives us additional 36 features.

5. Conclusions and Future Work

Detection of speech disfluencies is important for many reasons from evaluating the spoken fluency skills to improving the performance of ASR systems. In this article we presented different approaches towards hesitation detection on the joint and quality diverse corpus of Russian spontaneous speech. We discussed the application of the rule-based and data-driven methods to the hesitation detection for Russian. We implemented different techniques from grid search and gradient descent in rule-based approaches to such data-driven ones as ELM and SVM based on the automatically extracted acoustic features. Experimental results on the mixed and quality diverse corpus of spontaneous Russian speech indicate the efficiency of the techniques for the task, with SVM outperforming other methods, at the moment giving us $F1\text{-score} = 0.54 \pm 0.027$, with precision and recall being 0.55 ± 0.05 and 0.53 ± 0.04, respectively. The future work will be aimed at addressing the problem of analysis of false positives and false negatives by tuning SVM, by expert analysis and by utilizing additional context levels.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This research is supported by the grant of Russian Foundation for Basic Research (Project no. 15-06-04465) and by the State Research no. 0073-2014-0005.

References

[1] F. G. Eisler, *Psycholinguistics: Experiments in Spontaneous Speech*, Academic Press, 1968.

[2] W. L. Chafe, Ed., *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, Ablex Publishing Corp, Norwood, Mass, USA, 1980.

[3] E. Shriberg, *Preliminaries to a theory of speech disfluencies [Ph.D. thesis]*, University of California at Berkeley, 1994.

[4] J. E. F. Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, no. 6, pp. 709–738, 1995.

[5] A. Kibrik and V. Podlesskaya, Eds., *Night Dream Stories: Corpus Study of Russian Discourse*, Litres, 2014.

[6] D. C. O'Connell and S. Kowal, "The history of research on the filled pause as evidence of the written language bias in linguistics (Linell, 1982)," *Journal of Psycholinguistic Research*, vol. 33, no. 6, pp. 459–474, 2004.

[7] A. Stolcke, E. Shriberg, R. A. Bates et al., "Automatic detection of sentence boundaries and disfluencies based on recognized

words," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '98)*.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switch board: telephone speech corpusfor research and development," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 1, pp. 517–520, San Francisco, Calif, USA, March 1992.

[9] H. Medeiros, H. Moniz, F. Batista, I. Trancoso, and L. Nunes, "Disfluency detection based on prosodic features for university lectures," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH '13)*, pp. 2629–2633, Lyon, France, August 2013.

[10] 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 (INTERSPEECH '05)*, pp. 1781–1784, ISCA, Lisbon, Portugal, September 2005.

[11] H. Clark, *Using Language*, Cambridge University Press, Cambridge, UK, 1996.

[12] R. Ogden, "Turn-holding, turn-yielding and laryngeal activity in finnish talkin-interaction," *Journal of the International Phonetics Association*, vol. 31, no. 1, pp. 139–152, 2001.

[13] T. Arbisi-Kelm and S. A. Jun, "A comparison of disfluency patterns in normal andstuttered speech," in *Disfluency in Spontaneous Speech*, 2005.

[14] F. Ferreira, E. F. Lau, and K. G. D. Bailey, "Disfluencies, language comprehension, and tree adjoining grammars," *Cognitive Science*, vol. 28, no. 5, pp. 721–749, 2004.

[15] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1539, 2006.

[16] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken english," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 4857–4860, April 2009.

[17] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system forspontaneous speech recognition," in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, pp. 227–230, ISCA, Budapest, Hungary, 1999.

[18] V. Khurshudian, "Hesitation in typologically different languages: an experimentalstudy," in *Proceedings of the International Conference on Computational Linguistics Dialogue*, pp. 497–501, 2005.

[19] S. Stepanova, "Some features of filled hesitation pauses in spontaneous russian," in *Proceedings of the 16th International Congress of Phonetic Sciences*, vol. 16, pp. 1325–1328, Saarbrucken, Germany, 2007.

[20] A. Giannini, "Hesitation phenomena in spontaneous italian," in *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 2653–2656, Barcelona, Spain, 2003.

[21] E. Shriberg, "To 'Errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001.

[22] J. Peters, "LM studies on filled pauses in spontaneous medical dictation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 2, pp. 82–84, Association for Computational Linguistics, Edmonton, Canada, May 2003.

[23] F. Stouten and J. P. Martens, "A feature-based filled pause detection system for dutch," in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 309–314, IEEE, 2003.

[24] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 1, pp. 521–524, IEEE, 1992.

[25] E. Shriberg, R. A. Bates, and A. Stolcke, "A prosody only decision-tree model fordisfluency detection," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 2383–2386, Rhodes, Greece, 1997.

[26] H. Medeiros, F. Batista, H. Moniz, I. Trancoso, and H. Meinedo, "Experiments onautomatic detection of filled pauses using prosodic features," *Actas de Inforum*, pp. 335–345, 2013.

[27] INTERSPEECH: Computational Paralinguistic Challenge, 2013, http://emotion-research.net/sigs/speech-sig/is13-compare.

[28] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH '13)*, pp. 173–177, Lyon, France, August 2013.

[29] D. Prylipko, O. Egorow, I. Siegert, and A. Wendemuth, "Application of image processing methods to filled pauses detection from spontaneous speech," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association: Celebrating the Diversity of Spoken Languages (INTERSPEECH '14)*, pp. 1816–1820, Singapore, September 2014.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia ACM Multimedia (MM '10)*, pp. 1459–1462, ACM, Firenze, Italy, October 2010.

[31] Y. Liu, *Structural event detection for rich transcription of speech [Ph.D. thesis]*, Purdue University, 2004.

[32] LDC: English CTS treebank with structural metadata, http://catalog.ldc.upenn.edu/LDC2009T01.

[33] LDC: Czech broadcast conversation MDE transcripts, http://catalog.ldc.upenn.edu/LDC2009T20.

[34] LDC, "Czech broadcast conversation speech," http://catalog.ldc.upenn.edu/LDC2009S02.

[35] J. Kolár, J. Svec, S. Strassel, C. Walker, D. Kozlíková, and J. Psutka, "Czech spontaneous speech corpus with structural metadata," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, Lisbon, Portugal, September 2005.

[36] V. Verkhodanova and V. Shapranov, "Automatic detection of filled pauses and lengthenings in the spontaneous Russian speech," in *Proceedings of the 7th International Conference on Speech Prosody (SP '14)*, pp. 1110–1114, Dublin, Ireland, May 2014.

[37] E. Zemskaya, Russian Spoken Speech: Linguistic Analysis and the Problems of Learning. Moscow, 1979.

[38] A. Anderson, M. Bader, E. Bard et al., "The HCRC map task corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[39] K. J. Kohler, "Labelled data bank of spoken standard German—the Kiel Corpus of read/spontaneous speech," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)*, vol. 3, pp. 1938–1941, IEEE, October 1996.

[40] S. A. Zahorian, J. Wu, M. Karnjanadecha et al., "Open-source multi-language audio database for spoken language processing applications," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH '11)*, pp. 1493–1496, Florence, Italy, August 2011.

[41] Department of Phonetics of Saint Petersburg University, http://phonetics.spbu.ru/.

[42] G. Garg and N. Ward, "Detecting filled pauses in tutorial dialogs," 2006.

[43] J. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*, vol. 97, Springer Science & Business Media, 2005.

[44] V. Verkhodanova and V. Shapranov, "Multi-factor method for detection of filledpauses and lengthenings in Russian spontaneous speech," in *Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20–24, 2015, Proceedings*, vol. 9319 of *Lecture Notes in Computer Science*, pp. 285–292, Springer, Berlin, Germany, 2015.

[45] V. Verkhodanova, V. Shapranov, and A. Karpov, "Filled pauses and lengthenings detection using machine learning techniques," in *Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics ExLing*, pp. 175–178, Saint Petersburg, Russia, July 2016.

[46] A. Akusok, K.-M. Björk, Y. Miche, and A. Lendasse, "High-performance extreme learning machines: a complete toolbox for big data applications," *IEEE Access*, vol. 3, pp. 1011–1025, 2015.

[47] P. Boersma and D. Weenink, Praat: doing phonetics by computer [computer program], version 6.0.11, http://www.praat.org/.

[48] V. Verkhodanova and V. Shapranov, "Detecting filled pauses and lengthenings in russian spontaneous speech using SVM," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23–27, 2016, Proceedings*, vol. 9811 of *Lecture Notes in Computer Science*, pp. 224–231, Springer, Berlin, Germany, 2016.

[49] Scikit-Learn: Machine learning in Python, http://scikit-learn.org.

[50] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," in *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, pp. 1–127, 2011.

[51] H. J. Heijmans, "Mathematical morphology: a modern approach in image processing based on algebra and geometry," *SIAM Review*, vol. 37, no. 1, pp. 1–36, 1995.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 8363507, 19 pages
http://dx.doi.org/10.1155/2016/8363507

Research Article

Score-Informed Source Separation for Multichannel Orchestral Recordings

Marius Miron, Julio J. Carabias-Orti, Juan J. Bosch, Emilia Gómez, and Jordi Janer

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

Correspondence should be addressed to Marius Miron; marius.miron@upf.edu

Received 24 June 2016; Accepted 3 November 2016

Academic Editor: Alexander Petrovsky

This paper proposes a system for score-informed audio source separation for multichannel orchestral recordings. The orchestral music repertoire relies on the existence of scores. Thus, a reliable separation requires a good alignment of the score with the audio of the performance. To that extent, automatic score alignment methods are reliable when allowing a tolerance window around the actual onset and offset. Moreover, several factors increase the difficulty of our task: a high reverberant image, large ensembles having rich polyphony, and a large variety of instruments recorded within a distant-microphone setup. To solve these problems, we design context-specific methods such as the refinement of score-following output in order to obtain a more precise alignment. Moreover, we extend a close-microphone separation framework to deal with the distant-microphone orchestral recordings. Then, we propose the first open evaluation dataset in this musical context, including annotations of the notes played by multiple instruments from an orchestral ensemble. The evaluation aims at analyzing the interactions of important parts of the separation framework on the quality of separation. Results show that we are able to align the original score with the audio of the performance and separate the sources corresponding to the instrument sections.

1. Introduction

Western classical music is a centuries-old heritage traditionally driven by well-established practices. For instance, large orchestral ensembles are commonly tied to a physically closed place, the concert hall. In addition, Western classical music is bounded by established customs related to the types of instruments played, the presence of a score, the aesthetic guidance of a conductor, and compositions spanning a large time frame. Our work is conducted within the PHENICX project [1], which aims at enriching the concert experience through technology. Specifically, this paper aims at adapting and extending score-informed audio source separation to the inherent complexity of orchestral music. This scenario involves challenges like changes in dynamics and tempo, a large variety of instruments, high reverberance, and simultaneous melodic lines but also opportunities as multichannel recordings.

Score-informed source separation systems depend on the accuracy of different parts which are not necessarily integrated in the same parametric model. For instance, they rely on a score alignment framework that yields a coarsely aligned score [2–5] or, in a multichannel scenario, they compute a panning matrix to assess the weight of each instrument in each channel [6, 7]. To account for that, we adapt and improve the parts of the system within the complex scenario of orchestral music. Furthermore, we are interested in establishing a methodology for this task for future research and we propose a dataset in order to objectively assess the contribution of each part of the separation framework to the quality of separation.

1.1. Relation to Previous Work. Audio source separation is a challenging task when sources corresponding to different instrument sections are strongly correlated in time and frequency [8]. Without any previous knowledge, it is difficult to separate two sections which play, for instance, consonant notes simultaneously. One way to approach this problem is to introduce into the separation framework information about the characteristics of the signals such as a well-aligned score [2–5, 9, 10]. Furthermore, previous research relates the

accuracy of the alignment to the quality of source separation [4, 11]. For Western classical music, a correctly aligned score yields the exact time where each instrument is playing. Thus, an important step in score-informed source separation is obtaining a correctly aligned score, which can be done automatically with an audio-to-score alignment system.

Audio-to-score alignment deals with the alignment of a symbolic representation such as the score with the audio of the rendition. In a live scenario, this task deals with following the musical score while listening to the live performance, and it is known as score-following. To our knowledge, with the exception of [12], audio-to-score alignment systems have not been rigorously tested in the context of orchestral music. However, with respect to classical music, limited experimental scenarios comprising Bach chorales played by a four instruments have been discussed in [4, 13]. Furthermore, in [14], a subset of RWC classical music database [15] is used for training and testing, though no details are given regarding the instrumental complexity of the pieces. Moreover, a Beethoven orchestral piece from the same database is tested in [16], obtaining lower accuracy than the other evaluated pieces. These results point out the complexity of an orchestral scenario, underlined in [12]. Particularly, a larger number of instruments, many instruments playing concurrently different melody lines [17], prove to be a more difficult problem than tracking a limited number of instruments as in pop music or, for instance, string quartets and piano pieces. Although, in this paper, we do not propose a new system for orchestral music audio-to-score alignment, the task being a complex and extensive itself, we are interested in analyzing the relation between the task and the quality of score-informed source separation in such a complex scenario.

Besides the score, state-of-the-art source separation methods take into account characteristics of the audio signal which can be integrated into the system, thus achieving better results. For instance, the system can learn timbre models for each of the instruments [7, 18]. Moreover, it can rely on the assumption that the family of featured instruments is known, and their spectral characteristics are useful to discriminate between different sections playing simultaneously, when the harmonics of the notes overlap. In a more difficult case, when neither the score or the timbre of the instrument is available, temporal continuity and frequency sparsity [19] help in distributing the energy between sources in a musically meaningful way. Furthermore, an initial pitch detection can improve the results [6, 20–22], if the method assumes a predominant source. However, our scenario assumes multichannel recordings with distant microphones, in contrast to the close-microphone approach in [6], and we cannot assume that a source is predominant. As a matter a fact, previous methods deal with a limited case: the separation between small number of harmonic instruments or piano [3, 4, 6, 18], leaving the case of orchestral music as an open issue. In this paper, we investigate a scenario characterized by large reverberation halls [23], a large number of musicians in each section, a large diversity of instruments played, abrupt tempo changes, and many concurrent melody lines, often within the same instrument section.

Regarding the techniques used for source separation, matrix decomposition has been increasingly popular for source separation during the recent years [2, 3, 6, 18–20]. Nonnegative matrix factorization (NMF) is a particular case of decomposition which restricts the values of the factor matrices to be nonnegative. The first-factor matrix can be seen as a dictionary representing spectral templates. For audio signals, a dictionary is learned for each of the sources and stored into the basis matrix as a set of spectral templates. The second-factor matrix holds the temporal activation or the weights of the templates. Then, the resulting factorized spectrogram is calculated as a linear combination of the template vectors with a set of weight vectors forming the activation matrix. This representation allows for parametric models such as the source-filter model [3, 18, 20] or the multiexcitation model [9], which can easily capture important traits of harmonic instruments and help separate between them, as it is the case with orchestral music. The multiexcitation model has been evaluated in a restricted scenario of Bach chorales played by a quartet [4] and for this particular database has been extended in the scope of close-microphone recordings [6] and score-informed source separation [11]. From a source separation point of view, in this article, we extend and evaluate the work in [6, 11, 18] for orchestral music.

In order to obtain a better separation with any NMF parametric model, the sparseness of the gains matrix is increased by initializing it with time and frequency information obtained from the score [3–5, 24]. The values between the time frames where a note template is not activated are set to zero and will remain this way during factorization, allowing for the energy from the spectrogram to be redistributed between the notes and the instruments which actually play during that interval. A better alignment leads to better gains initialization and better separation. Nonetheless, audio-to-score alignment mainly fixes global misalignments, which are due to tempo variations, and does not deal with local misalignments [21]. To account for local misalignments, score-informed source separation systems include onset and offset information into the parametric model [3, 5, 24] or use image processing in order to refine the gains matrix so that it closely matches the actual time and frequency boundaries of the played notes [11]. Conversely, local misalignments can be fixed explicitly [25–27]. To our knowledge, none of these techniques have been explored for orchestral music, although there is a scope for testing their usefulness, if we take into account several factors as the synchronization of musicians in large ensembles, concurrent melody lines, and reverberation. Furthermore, the alignment systems are monaural. However, in our case, the separation is done on multichannel recordings and the delays between the sources and microphones might yield local misalignments. Hence, towards a better separation and more precise alignment, we propose a robust method to refine the output of a score alignment system with respect to each audio channel of the multichannel audio.

In the proposed distant-microphone scenario, we normally do not have microphones close to a particular instrument or soloist and, moreover, in an underdetermined case,

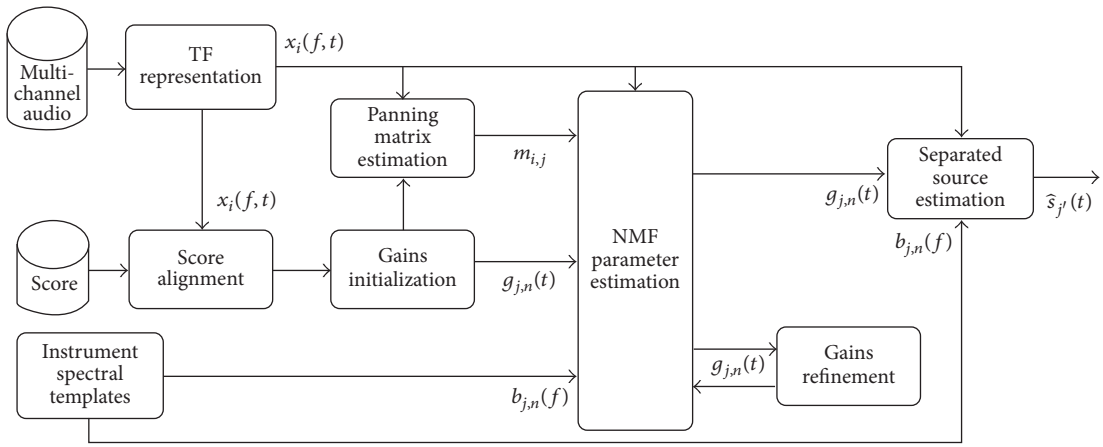


Figure 1: The diagram representing the flow of operations in the system.

the number of sources surpasses the number of microphones. Therefore, recording the sound of an entire section also captures interference from other instruments and the reverberation of the concert hall. To that extent, our task is different from interference reduction in close-microphone recordings [6, 7, 23], these approaches being evaluated for pop concerts [23] or quartets [6]. Additionally, we do not target a blind case source separation as the previous systems [6, 7, 23]. Subsequently, we adapt and improve the systems in [6, 11], by using information from all the channels, similar to parallel factor analysis (PARAFAC) in [28, 29].

With respect to the evaluation, to our knowledge, this is the first time score-informed source separation is objectively evaluated on such a complex scenario. An objective evaluation provides a more precise estimation of the contribution of each part of the framework and their influence on the separation. Additionally, it establishes a methodology for future research and eases the research reproducibility. We annotated the database proposed in [30], comprising four pieces of orchestral music recorded in an anechoic room, in order to obtain a score which is perfectly aligned with the anechoic recordings. Then, using the Roomsim software in [31], we simulate a concert hall in order to obtain realistic multitrack recordings.

1.2. Applications. The proposed framework for score-informed source separation has been used to separate recordings by various orchestras. The recordings are processed automatically and stored in the multimodal repository Repovizz [32]. The repository serves the data through its API for several applications. The first application is called instrument emphasis or orchestra focus and allows for emphasizing a particular instrument over the full orchestra. The second application relates to spatialization of the separated musical sources in the case of virtual reality scenarios and it is commonly known as Acoustic Rendering. Third, we propose an application to estimating the spatial locations of the instruments on the stage. All the three applications are detailed in Section 7.

1.3. Outline. We introduce the architecture of the framework with its main parts in Section 2. Then, in Section 3, we give an outline of the baseline source separation system. Furthermore, we present the extension of the baseline system: the initialization of the gains with score information (Section 4) and the note refinement (Section 4.1). Additionally, the proposed extension to the multichannel case is introduced in Section 5. We present the dataset and the evaluation procedures and discuss the results in Section 6. The demos and applications are described in Section 7.

2. Proposed Approach Overview

The diagram of the proposed framework is presented in Figure 1. The baseline system relies on training spectral templates for the instruments we aim to separate (Section 3.3). Then, we compute the spectrograms associated with the multichannel audio. The spectrograms along with the score of the piece are used to align the score to the audio. From the aligned score, we derive gains matrix that serves as an input for the NMF parameter estimation stage (Section 4), along with the learned spectral templates. Furthermore, the gains and the spectrogram are used to calculate a panning matrix (Section 3.1) which yields the contribution of each instrument in each channel. After the parameter estimation stage (Section 3.2), the gains are refined in order to improve the separation (Section 4.1). Then, the spectrograms of the separated sources are estimated using Wiener filtering (Section 3.5).

For the score alignment step, we use the system in [13] which aligns the scores to a chosen microphone and achieved the best results in MIREX score-following challenge (http://www.music-ir.org/mirex/wiki/2015:Real-time_Audio_to_Score_Alignment_(a.k.a._Score_Following)_Results). However, other state-of-the-art alignment systems can be used at this step, since our final goal is to refine a given score with respect to each channel, in order to minimize the errors in separation (Section 5.2). Accounting for that, we extend the model proposed by [6] and the gains refinement for monaural recordings in [11] to the case of score-informed multichannel source separation in the more complex scenario of orchestral music.

3. Baseline Method for Multichannel Source Separation

According to the baseline model in [6], the short-term complex valued Fourier transform (STFT) in time frame t and frequency f for channel $i = 1, \ldots, I$, where I is the total number of channels, is expressed as

$$\underline{x}_i(f,t) \approx \widehat{\underline{x}}_i(f,t) = \sum_{j=1}^{J} \underline{m}_{i,j} s_j(f,t), \quad (1)$$

where $s_j(f,t)$ represents the estimation of the complex valued STFT computed for the source $j = 1, \ldots, J$, with J the total number of sources. Note that, in this paper, we consider a source or instrument, one or more instruments of the same kind (e.g., a section of violins). Additionally, $\underline{m}_{i,j}$ is a mixing matrix of size $I \times J$ that accounts for the contribution of source i to channel j. In addition, we denote $x_i(f,t)$ as the magnitude spectrogram and $m_{i,j}$ as the real-valued panning matrix.

Under the NMF model described in [18], each source $s_j(f,t)$ is factored as a product of two matrices: $g_{j,n}(t)$, the matrix which holds the gains or activation of the basis function corresponding to pitch n at frame t, and $b_{j,n}(f)$, $n = 1, \ldots, N$, the matrix which holds bases, where $n = 1, \ldots, N$ is defined as the pitch range for instrument j. Hence, source j is modeled as

$$s_j(f,t) \approx \sum_{n=1}^{N} b_{j,n}(f)\, g_{j,n}(t). \quad (2)$$

The model represents a pitch for each source j as a single template stored in the basis matrix $b_{j,n}(f)$. The temporal activation of a template (e.g., onset and offset times for a note) is modeled using the gains matrix $g_{j,n}(t)$. Under harmonicity constraints [18], the NMF model for the basis matrix is defined as

$$b_{j,n}(f) = \sum_{h=1}^{H} a_{j,n}(h)\, G\left(f - h f_0(n)\right), \quad (3)$$

where $h = 1, \ldots, H$ is the number of harmonics, $a_{j,n}(h)$ is the amplitude of harmonic h for note n and instrument j, $f_0(n)$ is the fundamental frequency of note n, $G(f)$ is the magnitude spectrum of the window function, and the spectrum of a harmonic component at frequency $hf_0(n)$ is approximated by $G(f - hf_0(n))$.

Considering the model given in (3), the initial equation (2) for the computation of the magnitude spectrogram for the source j is expressed as

$$s_j(f,t) \approx \sum_{n=1}^{N} g_{j,n}(t) \sum_{h=1}^{H} a_{j,n}(h)\, G\left(f - h f_0(n)\right), \quad (4)$$

and (1) for the factorization of magnitude spectrogram for channel i is rewritten as

$$\widehat{x}_i(f,t) = \sum_{j=1}^{J} m_{i,j} \sum_{n=1}^{N} g_{j,n}(t) \sum_{h=1}^{H} a_{j,n}(h)\, G\left(f - h f_0(n)\right). \quad (5)$$

3.1. Panning Matrix Estimation. The panning matrix gives the contribution of each instrument in each channel and as seen in (1) influences directly the separation of the sources. The panning matrix is estimated by calculating an overlapping mask which discriminates the time-frequency zones for which the partials of a source are not overlapped with the partials of other sources. Then, using the overlapping mask, a panning coefficient is computed for each pair of sources at each channel. The estimation algorithm in the baseline framework is described in [6].

3.2. Augmented NMF for Parameter Estimation. According to [33], the parameters of the NMF model are estimated by minimizing a cost function which measures the reconstruction error between the observed $x_i(f,t)$ and the estimated $\widehat{x}_i(f,t)$. For flexibility reasons, we use the beta-divergence [34] cost function, which allows for modeling popular cost functions for different values of β, such as Euclidean (EUC) distance ($\beta = 2$), Kullback-Leibler (KL) divergence ($\beta = 1$), and the Itakura-Saito (IS) divergence ($\beta = 0$).

The minimization procedures assure that the distance between $x_i(f,t)$ and $\widehat{x}_i(f,t)$ does not increase with each iteration, thus accounting for the nonnegativity of the basis and the gains. By these means, the magnitude spectrogram of a source is explained solely by additive reconstruction.

3.3. Timbre-Informed Signal Model. An advantage of the harmonic model is that templates can be learned for various instruments, if the appropriate training data is available. The RWC instrument database [15] offers recordings of solo instrument playing isolated notes along all their corresponding pitch range. The method in [6] uses these recordings along with the ground truth annotation to learn instrument spectral templates for each note of each instrument. More details on the training procedure can be found in the original paper [6].

Once the basis functions $b_{j,n}(f)$ corresponding to the spectral templates are learned, they are used at the factorization stage in any orchestral setup which contains the targeted instruments. Thus, after training the basis $b_{j,n}(f)$ are kept fixed, while the gains $g_{j,n}(t)$ are estimated during the factorization procedure.

3.4. Gains Estimation. The factorization procedure to estimate the gains $g_{j,n}(t)$ considers the previously computed panning matrix $m_{i,j}$ and the learned basis $b_{j,n}(f)$ from the training stage. Consequently, we have the following update rules:

$$g_{j,n}(t) \longleftarrow g_{j,n}(t) \frac{\sum_{f,i} m_{i,j} b_{j,n}(f)\, x_i(f,t)\, \widehat{x}_i(f,t)^{\beta-2}}{\sum_{f,i} m_{i,j} b_{j,n}(f)\, \widehat{x}(f,t)^{\beta-1}}. \quad (6)$$

3.5. From the Estimated Gains to the Separated Signals. The reconstruction of the sources is done by estimating the complex amplitude for each time-frequency bin. In the case of binary separation, a cell is entirely associated with a single source. However, when having many instruments as in orchestral music, it is more advantageous to redistribute

(1) Initialize $b_{j,n}(f)$ with the values learned in Section 3.3.
(2) Initialize the gains $g_{j,n}(t)$ with score information.
(3) Initialize the mixing matrix $m_{i,j}$ with the values learned in Section 3.1.
(4) Update the gains using equation (6).
(5) Repeat Step (2) until the algorithm converges (or maximum number of iterations is reached).

Algorithm 1: Gain estimation method.

energy proportionally over all sources as in the Wiener filtering method [23].

This model allows for estimating each separated source $s_j(t)$ from mixture $x_i(t)$ using a generalized time-frequency Wiener filter over the short-time Fourier transform (STFT) domain as in [3, 34].

Let $\alpha_{j'}$ be the Wiener filter of source j', representing the relative energy contribution of the predominant source with respect to the energy of the multichannel mixed signal $x_i(t)$ at channel i:

$$\alpha_{j'}(t,f) = \frac{\left|A_{i,j'}\right|^2 \left|\underline{s}_j(f,t)\right|^2}{\sum_j \left|A_{i,j}\right|^2 \left|\underline{s}_j(f,t)\right|^2}. \quad (7)$$

Then, the corresponding spectrogram of source j' is estimated as

$$\widehat{\underline{s}}_{j'}(f,t) = \frac{\alpha_{j'}(t,f)}{\left|A_{i,j'}\right|^2} \underline{x}_i(f,t). \quad (8)$$

The estimated source $\widehat{s}_{j'}(f,t)$ is computed with the inverse overlap-add STFT of $\widehat{\underline{s}}_{j'}(f,t)$.

The estimated source magnitude spectrogram is computed using the gains $g_{j,n}(t)$ estimated in Section 3.4 and $b_{j,n}(f)$ the fixed basis functions learned in Section 3.3: $\widehat{s}_j(t,f) = g_{n,j}(t) b_{j,n}(f)$. Then, if we replace $s_j(t,f)$ with $\widehat{s}_j(t,f)$ and if we consider the mixing matrix coefficients computed in Section 3.1, we can calculate the Wiener mask from (7):

$$\alpha_{j'}(t,f) = \frac{m_{i,j'}^2 \widehat{s}_j(f,t)^2}{\sum_j m_{i,j}^2 \widehat{s}_j(f,t)^2}. \quad (9)$$

Using (8), we can apply the Wiener mask to the multichannel signal spectrogram, thus obtaining $\widehat{s}_{j'}(f,t)$, the estimated predominant source spectrogram. Finally, we use the phase information from the original mixture signal $\underline{x}_i(t)$, and, through inverse overlap-add STFT, we obtain the estimated predominant source $\widehat{s}_{j'}(t)$.

4. Gains Initialization with Score Information

In the baseline method [6], the gains are initialized following a transcription stage. In our case, the automatic alignment system yields a score which offers an analogous representation to the one obtained by the transcription. To that extent, the output of the alignment is used to initialize the gains for the NMF based methods for score-informed source separation.

Although the alignment algorithm aims at fixing global misalignments, it does not account for local misalignments. In the case of score-informed source separation, having a better aligned score leads to better separation [11], since it increases the sparseness of the gains matrix by setting to zero the activation for a time frame in which a note is not played (e.g., the corresponding spectral template of the note in the basis matrix is not activated outside this time boundary). However, in a real-case scenario, the initialization of gains derived from the MIDI score must take into account the local misalignments. This has been traditionally done by setting a tolerance window around the onsets and offsets [3, 5, 24] or by refining the gains after a number of NMF iterations and then reestimating the gains [11]. While the former integrates note refinement into the parametric model, the latter detects contours in the gains using image processing heuristics and explicitly associates them with meaningful entities as notes. In this paper, we present two methods for note refinement: in Section 4.1, we detail the method in [11] which is used as a baseline for our framework and, in Section 5.2, we adapt and improve this baseline to the multichannel case.

On these terms, if we account for errors up to d frames in the audio-to-score alignment, we need to increase the time interval around the onset and the offset for a MIDI note when we initialize the gains. Thus, the values in $g_{j,n}(t)$ for instrument j and pitch corresponding to a MIDI note n are set to 1 for the frames where the MIDI note is played, as well as the neighboring d frames. The other values in $g_{j,n}(t)$ are set to 0 and do not change during computation, while the values set to 1 evolve according to the energy distributed between the instruments.

Having initialized the gains, the classical augmented NMF factorization is applied to estimate the gains corresponding to each source j in the mixture. The process is detailed in Algorithm 1.

4.1. Note Refinement. The note refinement method in [11] aims at associating the values in the gains matrix $g_{j,n}(t)$ with notes. Therefore, it is applied after a certain number of iterations of Algorithm 1, when the gains matrix yields a meaningful distribution of the energy between instruments.

The method is applied on each note separately, with the scope of refining the gains associated with the targeted note. The gains matrix $g_{j,n}(t)$ can be understood as a greyscale image with each element in the matrix representing a pixel in the image. A set of image processing heuristics are deployed to detect shapes and contours in this image commonly known

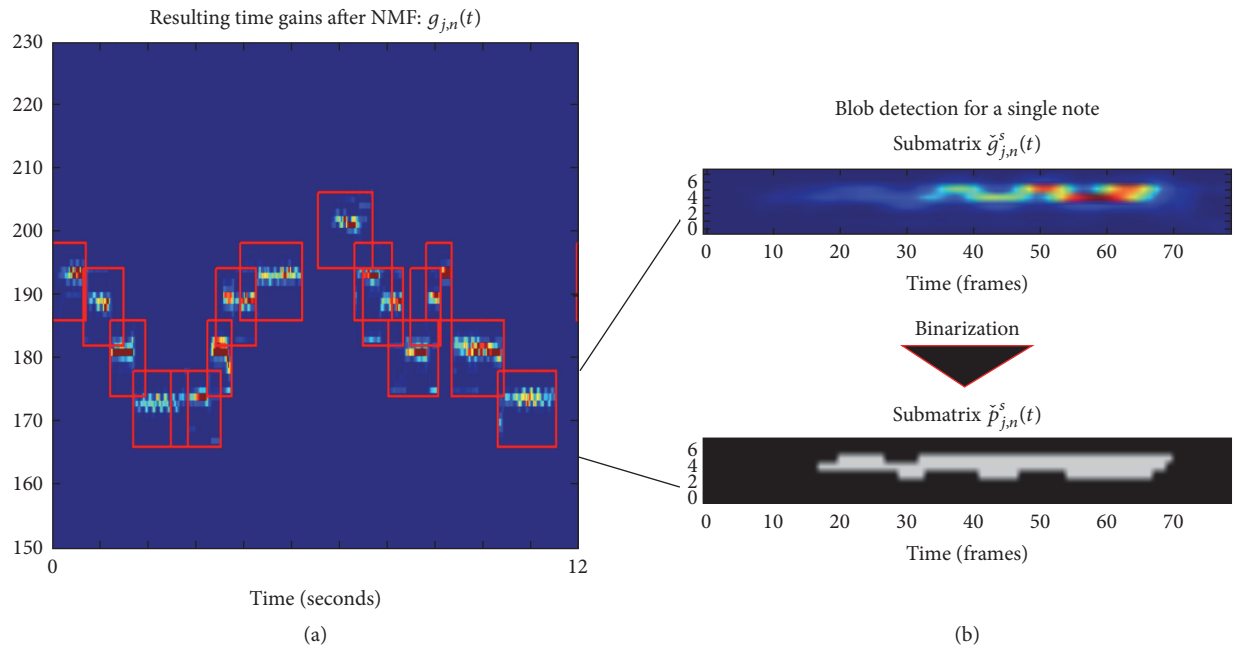


Figure 2: After NMF, the resulting gains (a) are split in submatrices (b) and used to detect blobs [11].

as blobs [35, p. 248]. As a result, each blob is associated with a single note, giving the onset and offset times for the note and its frequency contour. This representation further increases the sparsity of the gains $g_{j,n}(t)$, yielding less interference and better separation.

As seen in Figure 2, the method considers an image patch around the pixels corresponding to the pitch of the note and the onset and offset of the note given by the alignment stage, plus the additional d frames accounting for the local misalignments. In fact, the method works with the same submatrices of $g_{j,n}(t)$ which are set to 1 according to their corresponding MIDI note during the gains initialization stage, as explained in Section 3.4. Therefore, for a given note $k = 1, \ldots, K_j$, we process submatrix $\breve{g}^k_{j,n}(t)$ of the gains matrix $g_{j,n}(t)$, where K_j is the total number of notes for instrument j.

The steps of the method in [11], image preprocessing, binarization, and blob selection, are explained in Sections 4.1.2, 4.1.1, and 4.1.3.

4.1.1. Image Preprocessing. The preprocessing stage ensures through smoothing that there are no energy discontinuities within an image patch. Furthermore, it gives more weight to the pixels situated closer to the central bin in the blob in order to eliminate interference from the neighboring notes (close in time and frequency), but still preserving vibratos or transitions between notes.

First, we convolve with a smoothing Gaussian filter [35, p. 86] each row of the submatrix $\breve{g}^k_{j,n}(t)$. We choose a one-dimension Gaussian filter:

$$w(t) = \frac{1}{\sqrt{2\pi}\phi} e^{-(-t^2/2\phi^2)}, \tag{10}$$

where t is the time axis and ϕ is the standard deviation. Thus, each row vector of $\breve{g}^k_{j,n}(t)$ is convolved with $w(t)$, and the result is truncated in order to preserve the dimensions of the initial matrix by removing the mirrored frames.

Second, we penalize values in $\breve{g}^k_{j,n}(t)$ which are further away from the central bin by multiplying each column vector of this matrix with a 1-dimensional Gaussian centered in the central frequency bin, represented by vector $v(n)$:

$$v(n) = \frac{1}{\sqrt{2\pi}\nu} e^{-(n-\kappa)^2/2\nu^2}, \tag{11}$$

where n is the frequency axis, κ is the position of the central frequency bin, and ν is the standard deviation. The values of the parameters above are given in Section 6.4.2 as a part of the evaluation setup.

4.1.2. Image Binarization. Image binarization sets to zero the elements of the matrix $\breve{g}^k_{j,n}(t)$ which are lower than a threshold and to one the elements larger than the threshold. This involves deriving a submatrix $\breve{p}^k_{j,n}(t)$, associated with note k:

$$\breve{p}^k_{j,n}(t) = \begin{cases} 0, & \text{if } \breve{g}^k_{j,n}(t) < \text{mean}\left(\breve{g}^k_{j,n}(t)\right), \\ 1, & \text{if } \breve{g}^k_{j,n}(t) \geq \text{mean}\left(\breve{g}^k_{j,n}(t)\right). \end{cases} \tag{12}$$

4.1.3. Blob Selection. First, we detect blobs in each binary submatrix $\breve{p}^k_{j,n}(t)$, using the connectivity rules described in [35, p. 248] and [27]. Second, from the detected blob candidates, we determine the best blob for each note in a similar way to [11]. We assign a value to each blob, depending on its area and the overlap with the blobs corresponding to adjacent

notes, which will help us penalize the overlap between blobs of adjacent notes.

As a first step, we penalize parts of the blobs which overlap in time with other blobs from different notes $k-1, k, k+1$. This is done by weighting each element in $\check{g}^k_{j,n}(t)$ with factor γ, depending on the amount of overlapping with blobs from adjacent notes. The resulting score matrix has the following expression:

$$\check{q}^k_{j,n}(t) = \begin{cases} \gamma * \check{g}^k_{j,n}(t), & \text{if } \check{p}^k_{j,n}(t) \wedge \check{p}^{k-1}_{j,n}(t) = 1, \\ \gamma * \check{g}^k_{j,n}(t), & \text{if } \check{p}^k_{j,n}(t) \wedge \check{p}^{k+1}_{j,n}(t) = 1, \\ \check{g}^k_{j,n}(t), & \text{otherwise,} \end{cases} \tag{13}$$

where γ is a value in the interval $[0, 1]$.

Then, we compute a score for each note l and for each blob associated with the note, by summing up the elements in the score matrix $\check{q}^k_{j,n}(t)$ which are considered to be part of a blob. The best blob candidate is the one with the highest score and further on; it is associated with the note, its boundaries giving the note onset and offsets.

4.2. Gains Reinitialization and Recomputation. Having associated a blob with each note (Section 4.1.3), we discard obtrusive energy from the gains matrix $g_{j,n}(t)$, by eliminating the pixels corresponding to the blobs which were not selected, making the matrix sparser. Furthermore, the energy excluded from instrument's gains is redistributed to other instruments, contributing to better source separation. Thus, the gains are reinitialized with the information obtained from the corresponding blobs and we can repeat the factorization Algorithm 1 to recompute $g_{j,n}(t)$. Note that the energy which is excluded by note refinement is set to zero $g_{j,n}(t)$ and will remain zero during the factorization.

In order to refine the gains $g_{j,n}(t)$, we define a set of matrices $p^k_{j,n}(t)$ derived from the matrices corresponding to the best blobs $\check{p}^k_{j,n}(t)$ which contain 1 only for the elements associated with the best blob and 0 otherwise. We rebuild the gains matrix $g_{j,n}(t)$ with the set of submatrices $p^k_{j,n}(t)$. For the corresponding bins n and time frames t of note k, we initialize the values in $g_{j,n}(t)$ with the values in $p^k_{j,n}(t)$. Then, we reiterate the gains estimation with Algorithm 1. Furthermore, we obtain the spectrogram of the separated sources with the method described in Section 3.5.

5. PARAFAC Model for Multichannel Gains Estimation

Parallel factor analysis methods (PARAFAC) [28, 29] are mostly used under the nonnegative tensor factorization paradigm. By these means, the NMF model is extended to work with 3-valence tensors, where each slice of the sensor represents the spectrogram for a channel. Another approach is to stack up spectrograms for each channel in a single matrix [36] and perform a joint estimation of the spectrograms of sources in all channels. Hence, we extend the NMF model in Section 3 to jointly estimate the gains matrices in all the channels.

5.1. Multichannel Gains Estimation. The algorithm described in Section 3 estimates gains $g_{n,j}$ for source j with respect to single channel i determined as the corresponding row in column j of the panning matrix where element $m_{i,j}$ has the maximum value. However, we argue that a better estimation can benefit from the information in all channels. To this extent, we can further include update rules for other parameters such as mixing matrix $m_{i,j}$ which were otherwise kept fixed in Section 3, because the factorization algorithm estimates the parameters jointly for all the channels.

We propose to integrate information from all channels by concatenating their corresponding spectrogram matrices on the time axis, as in

$$x(f,t) = \left[x_1(f,t) \;\; x_1(f,t) \;\; \cdots \;\; x_I(f,t)\right]. \tag{14}$$

We are interested in jointly estimating the gains $g_{n,j}$ of the source j in all the channels. Consequently, we concatenate in time the gains corresponding to each channel i for $i = 1, \ldots, I$, where I is the total number of channels, as seen in (15). The new gains are initialized with identical score information obtained from the alignment stage. However, during the estimation of the gains for channel i, the new gains $g^i_{n,j}(t)$ evolve accordingly, taking into account the corresponding spectrogram $x_i(f,t)$. Moreover, during the gains refinement stage, each gain is refined separately with respect to each channel:

$$g_{n,j}(t) = \left[g^1_{n,j}(t) \;\; g^2_{n,j}(t) \;\; \cdots \;\; g^I_{n,j}(t)\right]. \tag{15}$$

In (5), we describe the factorization model for the estimated spectrogram, considering the mixing matrix, the basis, and the gains. Since we estimate a set of I gains for each source $j = 1, \ldots, J$, this will result in J estimations of the spectrograms corresponding to all the channels $i = 1, \ldots, I$, as seen in

$$\widehat{x}^j_i(f,t) = \sum_{j=1}^{J} m_{i,j} \sum_{n=1}^{N} g^i_{j,n}(t) \sum_{h=1}^{H} a_{j,n}(h)\, G\left(f - h f_0(n)\right). \tag{16}$$

Each iteration of the factorization algorithm yields additional information regarding the distribution of energy between each instrument and each channel. Therefore, we can include in the factorization update rules for mixing matrix $m_{i,j}$ as in (17). By updating the mixing parameters at each factorization step, we can obtain a better estimation for $\widehat{x}^j_i(f,t)$:

$$m_{i,j} \longleftarrow m_{i,j} \frac{\sum_{f,t} b_{j,n}(f)\, g_{n,j}(t)\, x_i(f,t)\, \widehat{x}_i(f,t)^{\beta-2}}{\sum_{f,t} b_{j,n}(f)\, g_{n,j}(t)\, \widehat{x}(f,t)^{\beta-1}}. \tag{17}$$

Considering the above, the new rules to estimate the parameters are described in Algorithm 2.

(1) Initialize $b_{j,n}(f)$ with the values learned in Section 3.3.
(2) Initialize the gains $g_{j,n}(t)$ with score information.
(3) Initialize the panning matrix $m_{i,j}$ with the values learned in Section 3.1.
(4) Update the gains using equation (6).
(5) Update the panning matrix using equation (17).
(6) Repeat Step (2) until the algorithm converges (or maximum number of iterations is reached).

Algorithm 2: Gain estimation method.

Note that the current model does not estimate the phases for each channel. In order to reconstruct source j, the model in Section 3 uses the phase of the signal corresponding to channel i where it has the maximum value in the panning matrix, as described in Section 3.5. Thus, in order to reconstruct the original signals, we can solely rely on the gains estimated in single channel i, in a similar way to the baseline method.

5.2. Multichannel Gains Refinement. As presented in Section 5.1, for a given source, we obtain an estimation of the gains corresponding to each channel. Therefore, we can apply note refinement heuristics in a similar manner to Section 4.1 for each of the gains $[g^1_{n,j}(t), \ldots, g^I_{n,j}(t)]$. Then, we can average out the estimations for all the channel, making the blob detection more robust to the variances between the channels:

$$g'_{n,j}(t) = \frac{\sum_{i=1}^{I} g^i_{n,j}(t)}{I}. \tag{18}$$

Having computed the mean over all channels as in (18), for each note $k = 1, \ldots, K_j$, we process submatrix $\overline{g}^k_{j,n}(t)$ of the new gains matrix $g'_{j,n}(t)$, where K_j is the total number of notes for an instrument j. Specifically, we apply the same steps: preprocessing (Section 4.1.1), binarization (Section 4.1.2), and blob selection (Section 4.1.3), to each matrix $\overline{g}^k_{j,n}(t)$ and we obtain a binary matrix $\overline{p}^k_{j,n}(t)$ having 1 s for the elements corresponding to the best blob and 0 s for the rest.

Our hypothesis is that averaging out the gains between all channels makes blob detection more robust. However, when performing the averaging, we do not account for the delays between the channels. In order to compute the delay for a given channel, we can compute the best blob separately with the method in Section 4.1 (matrix $\check{p}^k_{j,n}(t)$) and compare it with the one calculated with the averaged estimation ($\overline{p}^k_{j,n}(t)$). This step is equivalent to comparing the onset times of the two best blobs for the two estimations. Subtracting these onset times, we get the delay between the averaged estimation and the one obtained for a channel and we can correct this in matrix $\overline{p}^k_{j,n}(t)$. Accordingly, we zero-pad the beginning of $\overline{p}^k_{j,n}(t)$ with the amount of zeros corresponding to the delay, or we remove the trailing zeros for a negative delay.

6. Materials and Evaluation

6.1. Dataset. The audio material used for evaluation was presented by Pätynen et al. [30] and consists of four passages of symphonic music from the Classical and Romantic periods. This work presented a set of anechoic recordings for each of the instruments, which were then synchronized between them so that they could later be combined to a mix of the orchestra. Musicians played in an anechoic chamber, and, in order to be synchronous with the rest of the instruments, they followed a video featuring a conductor and a pianist playing each of the four pieces. Note that the benefits of having isolated recordings comes at the expense of ignoring the interactions between musicians which commonly affect intonation and time-synchronization [37].

The four pieces differ in terms of number of instruments per instrument class, style, dynamics, and size. The first passage is a soprano aria of Donna Elvira from the opera Don Giovanni by W. A. Mozart (1756–1791), corresponding to the Classical period, and traditionally played by a small group of musicians. The second passage is from L. van Beethoven's (1770–1827) Symphony no. 7, featuring big chords and string crescendo. The chords and pauses make the reverberation tail of a concert hall clearly audible. The third passage is from Bruckner's (1824–1896) Symphony no. 8, and represents the late Romantic period. It features large dynamics and size of the orchestra. Finally, G. Mahler's Symphony no. 1, also featuring a large orchestra, is another example of late romanticism. The piece has a more complex texture than the one by Bruckner. Furthermore, according to the musicians which recorded the dataset, the last two pieces were also more difficult to play and record [30].

In order to keep the evaluation setup consistent between the four pieces, we focus in the following instruments: violin, viola, cello, double bass, oboe, flute, clarinet, horn, trumpet, and bassoon. All tracks from a single instrument were joined into a single track for each of the pieces.

For the selected instruments, we list the differences between the four pieces in Table 1. Note that in the original dataset the violins are separated into two groups. However, for brevity of evaluation and because in our separation framework we do not consider sources sharing the same instrument templates, we decided to merge the violins into a single group. Note that the pieces by Mahler and Bruckner have a divisi in the groups of violins, which implies a larger number of instruments playing different melody lines simultaneously. This results in a scenario which is more challenging for source separation.

We created a ground truth score, by hand annotating the notes played by the instruments. In order to facilitate this process, we first gathered the scores in MIDI format and automatically computed a global audio-score alignment, using the method from [13] which has won the MIREX

Table 1: Anechoic dataset [30] characteristics.

Piece	Duration	Period	Instrument sections	Number of tracks	Max. tracks/instrument
Mozart	3 min 47 s	Classical	8	10	2
Beethoven	3 min 11 s	Classical	10	20	4
Mahler	2 min 12 s	Romantic	10	30	4
Bruckner	1 min 27 s	Romantic	10	39	12

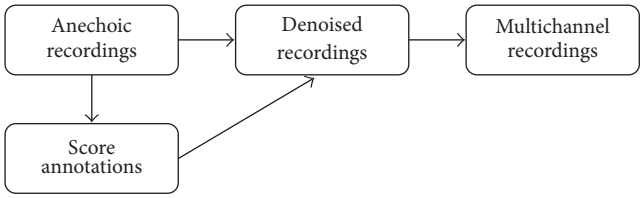


Figure 3: The steps to create the multichannel recordings dataset.

score-following challenge for the past years. Then, we locally aligned the notes of each instrument by manually correcting the onsets and offsets to fit the audio. This was performed using Sonic Visualiser, with the guidance of the spectrogram and the monophonic pitch estimation [38] computed for each of the isolated instruments. The annotation was performed by two of the authors, which cross-checked the work of their peer. Note that this dataset and the proposed annotation are useful not only for our particular task but also for the evaluation of multiple pitch estimation and automatic transcription algorithms in large orchestral settings, a context which has not been considered so far in the literature. The annotations can be found at the associated page (http://mtg.upf.edu/download/datasets/phenicx-anechoic).

During the recording process detailed in [30], the gain of the microphone amplifiers was fixed to the same value for the whole process, which reduced the dynamic range of the recordings of the quieter instruments. This led to noisier recordings for most of the instruments. In Section 6.2, we describe the score-informed denoising procedure we applied to each track. From the denoised isolated recordings, we then used Roomsim to create a multichannel image, as detailed in Section 6.3. The steps necessary to pass from the anechoic recordings to the multichannel dataset are represented in Figure 3. The original files can be obtained from the Acoustic Group at Aalto Univeristy (http://research.cs.aalto.fi/acoustics/). For the denoising algorithm, please refer to http://mtg.upf.edu/download/datasets/phenicx-anechoic.

6.2. Dataset Denoising. The noise related problems in the dataset were presented in [30]. We remove the noise in the recordings with the score-informed method in [39], which relies on learned noise spectral patters. The main difference is that we rely on a manually annotated score, while in [39] the score is assumed to be misaligned, so further regularization is included to ensure that only certain note combinations in the score occur.

The annotated score yields the time interval where an instrument is not playing. Thus, the noise pattern is learned only within that interval. In this way, the method assures that the desired noise, which is a part of the actual sound of the instrument, is preserved in the denoised recording.

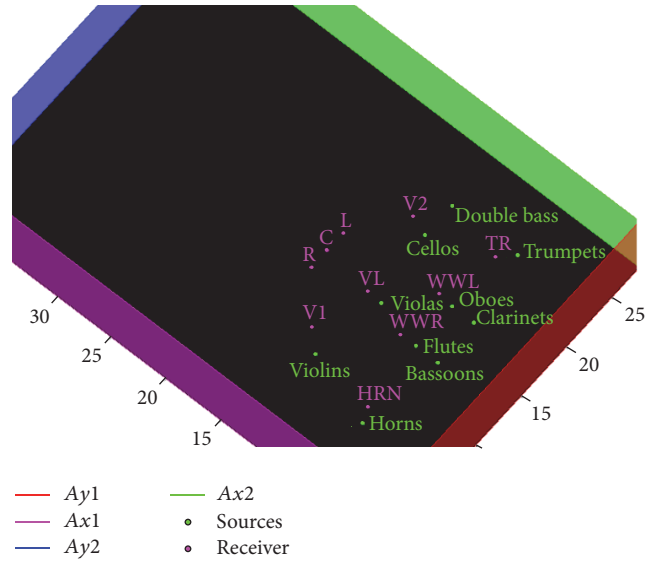


Figure 4: The sources and the receivers (microphones in the simulated room).

The algorithm takes each anechoic recording of a given instrument and removes the noise for the time interval where an instrument is playing, while setting to zero the frames where an instrument is not playing.

6.3. Dataset Spatialization. To simulate a large reverberant hall, we use the software Roomsim [31]. We define a configuration file which specifies the characteristics of the hall and, for each of the microphones, their position relative to each of the sources. The simulated room has similar dimensions to the Royal Concertgebouw in Amsterdam, one of the partners in the PHENICX project, and represents a setup in which we tested our framework. The simulated room's width, length, and height are 28 m, 40 m, and 12 m. The absorption coefficients are specified in Table 2.

The positions of the sources and microphones in the room are common for orchestral concerts (Figure 4). A configuration file is created for each microphone which contains its coordinates (e.g., (14, 17, 4) for the center microphone). Then, each source is defined through polar coordinates relative to the microphone (e.g., (11.4455, −95.1944, −15.1954) radius, azimuth, and elevation for the bassoon relative to the center microphone). We selected all the microphones to be cardioid, in order to match the realistic setup of Concertgebouw Hall.

Table 2: Room surface absorption coefficients.

Standard measurement frequencies (Hz)	125	250	500	1000	2000	4000
Absorption of wall in $x = 0$ plane	0.4	0.3	0.3	0.3	0.2	0.1
Absorption of wall in $x = Lx$ plane	0.4	0.45	0.35	0.35	0.45	0.3
Absorption of wall in $y = 0$ plane	0.4	0.45	0.35	0.35	0.45	0.3
Absorption of wall in $y = Ly$ plane	0.4	0.45	0.35	0.35	0.45	0.3
Absorption of floor, that is, $z = 0$ plane	0.5	0.6	0.7	0.8	0.8	0.9
Absorption of ceiling, that is, $z = Lz$ plane	0.4	0.45	0.35	0.35	0.45	0.3

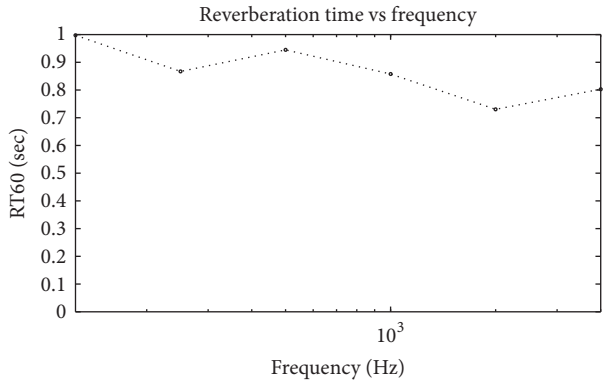


Figure 5: The reverberation time versus frequency for the simulated room.

Using the configuration file and the anechoic audio files corresponding to the isolated sources, Roomsim generates the audio files for each of the microphones along with the impulse responses for each pair of instruments and microphones. The impulse responses and the anechoic signals are used during the evaluation to obtain the ground truth spatial image of the sources in the corresponding microphone. Additionally, we plot Roomsim the reverberation time RT60 [31] across the frequencies in Figure 5.

We need to adapt ground truth annotations to the audio generated with Roomsim, as the original annotations were done on the isolated audio files. Roomsim creates an audio for a given microphone by convolving each of the sources with the corresponding impulse response and then summing up the results of the convolution. We compute a delay for each pair of microphones and instruments by taking the position of the maximum value in the associated impulse response vector. Then, we generate a score for each of the pairs by adding the corresponding delay to the note onsets. Additionally, since the offset time depends on the reverberation and the frequencies of the notes, we add 0.8 s to each note offset to account for the reverberation, besides the added delay.

6.4. Evaluation Methodology

6.4.1. Parameter Selection. In this paper, we use a low-level spectral representation of the audio data which is generated from a windowed FFT of the signal. We use a Hanning window with the size of 92 ms and a hop size of 11 ms. Here, a logarithmic frequency discretization is adopted. Furthermore, two time-frequency resolutions are used. First, to estimate the instrument models and the panning matrix, a single semitone resolution is proposed. In particular, we implement the time-frequency representation by integrating the STFT bins corresponding to the same semitone. Second, for the separation task, a higher resolution of 1/4 of semitone is used, which has proven to achieve better separation results [6]. The time-frequency representation is obtained by integrating the STFT bins corresponding to 1/4 of semitone. Note that in the separation stage, the learned basis functions $b_{j,n}(f)$ are adapted to the 1/4 of semitone resolution by replicating 4 times the basis at each semitone to the 4 samples of the 1/4 of semitone resolution that belong to this semitone. For image binarization, we pick for the first Gaussian $\phi = 3$ as the standard deviation and for the second Gaussian $\kappa = 4$ as the position of the central frequency bin and $\nu = 4$ as the standard deviation, corresponding to one semitone.

We picked 10 iterations for the NMF, and we set the beta-divergence distortion, $\beta = 1.3$, as in [6, 11].

6.4.2. Evaluation Setup. We perform three different kind of evaluations: audio-to-score alignment, panning matrix estimation, and score-informed source separation. Note that, for the alignment, we evaluate the state-of-the-art system in [13]. This method does not align notes but combinations of notes in the score (a.k.a states). Here, the alignment is performed with respect to a single audio channel, corresponding to the microphone situated in the center of the stage. On the other hand, the offsets are estimated by shifting the original duration for each note in the score [13] or by assigning the offset time as the onset for the next state. We denote these two cases as INT or NEX.

Regarding the initialization of the separation framework, we can use the raw output of the alignment system. However, as stated in Section 4 and [3, 5, 24], a better option is to extend the onsets and offsets along a tolerance window to account for the errors of the alignment system and for the delays between center channel (on which the alignment is performed) and the other channels and for the possible errors in the alignment itself. Thus, we test two hypotheses regarding the tolerance window for the possible errors. In the first case, we extend the boundaries with 0.3 s for onsets and 0.6 s for offsets (T1) and in the second with 0.6 s for onsets and 1 s for offsets (T2). Note that the value for the onset times of 0.3 s is not arbitrary but the usual threshold for onsets in the score-following in

Table 3: Score information used for the initialization of score-informed source separation.

Tolerance window size	Offset estimation
T1: onsets, 0.3 s; offsets, 0.6 s	INT: interpolation of the offset time
T2: onsets, 0.6 s; offsets, 0.9 s	NEX: the offset is the onset of the next note

MIREX evaluation of real-time score-following [40]. Two different tolerance windows were tested to account for the complexity of this novel scenario. The tolerance window is slightly larger for offsets due to the reverberation time and because the ending of the note is not as clear as its onset. A summary of the score information used to initialize the source separation framework is found in Table 3.

We label the test case corresponding to the initialization with the raw output of the alignment system as Ali. Conversely, the test case corresponding to the tolerance window initialization is labeled as Ext. Furthermore, within the tolerance window, we can refine the note onsets and offsets with the methods in Section 4.1 (Ref1) and Section 5.2 (Ref2), resulting in other two test cases. Since method Ref1 can only refine the score to a single channel, the results are solely computed with respect to that channel. For the multichannel refinement Ref2, we report the results of the alignment of each instrument with respect to each microphone. A graphic of the initialization of the framework with the four test cases listed above (Ali, Ext, Ref1, and Ref2), along with the ground truth score initialization (GT), is found in Figure 7, where we present the results for these cases in terms of source separation.

In order to evaluate the panning matrix estimation stage, we compute an ideal panning matrix based on the impulse responses generated by Roomsim during the creation of the multichannel audio (see Section 6.3). The ideal panning matrix gives the ideal contribution of each instrument in each channel and it is computed by searching the maximum in the impulse response vector corresponding to each instrument-channel pair, as in

$$m_{\text{ideal}}(i, j) = \max\left(\text{IR}(i, j)(t)\right), \tag{19}$$

where $\text{IR}(i, j)(t)$ is the impulse response of source i in channel j. By comparing the estimated matrix $m(i, j)$ with the ideal one $m_{\text{ideal}}(i, j)$, we can determine if the algorithm picked a wrong channel for separation.

6.4.3. Evaluation Metrics. For score alignment, we are interested in a measure which relates to source separation and accounts for the audio frames which are correctly detected, rather than an alignment rate computed per note onset, as found in [13]. Thus, we evaluate the alignment at the frame level rather than at a note level. A similar reasoning on the evaluation of score alignment is found in [4].

We consider 0.011 s the temporal granularity for this measure and the size of a frame. Then, a frame of a musical note is considered a true positive (tp) if it is found in the ground truth score and in the aligned score in the exact time boundaries. The same frame is labeled as a false positive (fp) if it is found only in the aligned score and a false negative (fn) if it is found only in the ground truth score. Since the gains initialization is done with score information (see Section 4), lost frames (recall), and incorrectly detected frames (precision) impact the performance of the source separation algorithm, precision is defined as $p = \text{tp}/(\text{tp} + \text{fp})$ and recall as $r = \text{tp}/(\text{tp} + \text{fn})$. Additionally, we compute the harmonic mean of precision and recall to obtain F-measure as $F = 2 \cdot ((p \cdot r)/(p + r))$.

The source separation evaluation framework and metrics employed are described in [41, 42]. Correspondingly, we use *Source to Distortion Ratio* (SDR), *Source to Interference Ratio* (SIR), and *Source to Artifacts Ratio* (SAR). While SDR measures the overall quality of the separation and ISR the spatial reconstruction of the source, SIR is related to rejection of the interferences and SAR to the absence of forbidden distortions and artifacts.

The evaluation of source separation is a computationally intensive process. Additionally, to process the long audio files in the dataset would require large memory to perform the matrix calculations. To reduce the memory requirements, the evaluation is performed for blocks of 30 s with 1 s overlap to allow for continuation.

6.5. Results

6.5.1. Score Alignment. We evaluate the output of the alignment Ali, along the estimation of the note offsets: INT and NEX in terms of F-measure (see Section 6.4.3), precision, and recall, ranging from 0 to 1. Additionally, we evaluate the optimal size for extending the note boundaries along the onsets and offsets, T1 and T2, for the refinement methods, Ref1 and Ref2, and the baseline, Ext. Since there are lots of differences between pieces, we report the results individually per song in Table 4.

Methods Ref1 and Ref2 depend on a binarization threshold which determines how much energy is set to zero. A lower threshold will result in the consolidation of larger blobs in blob detection. In [11], this threshold is set to 0.5 for a dataset of monaural recordings of Bach chorales played by four instruments. However, we are facing a multichannel scenario where capturing the reverberation is important, especially when we consider that offsets were annotated with a low energy threshold. Thus, we are interested in losing the least energy possible and we set lower values for the threshold: 0.3 and 0.1. Consequently, when analyzing the results, a lower threshold achieves better performance in terms of F-measure for Ref1 (0.67 and 0.72) and for Ref2 (0.71 and 0.75).

According to Table 4, extending note offsets (NEX), rather than interpolating them (INT), gives lower recall in all pieces, and the method leads to losing more frames which cannot be recovered even by extending the offset times in T2: NEX T2 yields always a lower recall when compared to INT T2 (e.g., $r = 0.85$ compared to $r = 0.97$ for Mozart).

The output of the alignment system Ali is not a good option to initialize the gains of source separation system. It has a high precision and a very low recall (e.g., the case INT Ali has $p = 0.95$ and $r = 0.39$ compared to case INT Ext

Table 4: Alignment evaluated in terms of F-measure, precision, and recall, ranging from 0 to 1.

			Mozart			Beethoven			Mahler			Bruckner		
			F	p	r	F	p	r	F	p	r	F	p	r
INT	T1	Ali	0.69	0.93	0.55	0.56	0.95	0.39	0.61	0.85	0.47	0.60	0.94	0.32
		Ref1	0.77	0.77	0.76	0.79	0.81	0.77	0.63	0.74	0.54	0.72	0.73	0.70
		Ref2	**0.83**	0.79	0.88	0.82	0.82	0.81	0.67	0.77	0.60	0.81	0.77	0.85
		Ext	0.82	0.73	0.94	**0.84**	0.84	0.84	**0.77**	0.69	0.87	**0.86**	0.78	0.96
	T2	Ref1	0.77	0.77	0.76	0.76	0.75	0.77	0.63	0.74	0.54	0.72	0.70	0.71
		Ref2	**0.83**	0.78	0.88	**0.82**	0.76	0.87	0.67	0.77	0.59	**0.79**	0.73	0.86
		Ext	0.72	0.57	0.97	0.79	0.70	0.92	**0.69**	0.55	0.93	0.77	0.63	0.98
NEX	T1	Ali	0.49	0.94	0.33	0.51	0.89	0.36	0.42	0.90	0.27	0.48	0.96	0.44
		Ref1	0.70	0.77	0.64	0.72	0.79	0.66	0.63	0.74	0.54	0.68	0.72	0.64
		Ref2	**0.73**	0.79	0.68	**0.71**	0.79	0.65	0.66	0.77	0.58	0.73	0.74	0.72
		Ext	**0.73**	0.77	0.68	**0.71**	0.80	0.65	**0.69**	0.75	0.64	**0.76**	0.79	0.72
	T2	Ref1	0.74	0.78	0.71	0.72	0.74	0.70	0.63	0.74	0.54	0.72	0.79	0.72
		Ref2	**0.80**	0.80	0.80	**0.75**	0.75	0.75	0.67	0.77	0.59	**0.79**	0.73	0.86
		Ext	0.73	0.65	0.85	0.73	0.69	0.77	**0.72**	0.64	0.82	**0.79**	0.69	0.91

Table 5: Instruments for which the closest microphone was incorrectly determined for different score information (GT, Ali, T1, T2, INT, and NEX) and two room setups.

	GT	INT			NEX		
		Ali	T1	T2	Ali	T1	T2
Setup 1	Clarinet	Clarinet, double bass	Clarinet, flute	Clarinet, flute, horn	Clarinet, double bass		Clarinet, flute
Setup 2	Cello	Cello, flute	Cello, flute	Cello, flute	Bassoon	Flute	Cello, flute

which has $p = 0.84$ and $r = 0.84$ for Beethoven). For the case of Beethoven, the output is particularly poor compared to other pieces. However, by extending the boundaries (Ext) and applying note refinement (Ref1 or Ref2), we are able to increase the recall and match the performance on the other pieces.

When comparing the size for the tolerance window for the onsets and offsets, we observe that the alignment is more accurate with detecting the onsets within 0.3 s and offsets within 0.6 s. In Table 4, T1 achieves better results than T2 (e.g., $F = 0.77$ for T1 compared to $F = 0.69$ for T2, Mahler). Relying on a large window retrieves more frames but also significantly damages the precision. However, when considering source separation we might want to lose as less information as possible. It is in this special case that the refinement methods Ref1 and Ref2 show their importance. When facing larger time boundaries as T2, Ref1 and especially Ref2 are able to reduce the errors by achieving better precision with the minimum amount of loss in recall.

The refinement Ref1 has a worse performance than Ref2, the multichannel refinement (e.g., $F = 0.72$ compared to $F = 0.81$ for Bruckner, INT T1). Note that, in the original version [11], Ref1 was assuming monophony within a source as it was tested in the simple case of Bach10 dataset [4]. To that extent, it was relying on a graph computation to determine the best distribution of blobs. However, due to the increased polyphony within an instrument (e.g., violins playing divisi), with simultaneous melodic lines, we disabled this feature and in this case Ref1 has lower recall, it loses more frames. On the other hand, Ref2 is more robust because it computes a blob estimation per channel. Averaging out these estimations yields better results.

The refinement works worse for more complex pieces (Mahler and Bruckner) than for simple pieces (Mozart and Beethoven). Increasing the polyphony within a source and the number of instruments, having many interleaving melodic lines, a less sparse score, also makes the task more difficult.

6.5.2. Panning Matrix. Estimating correctly the panning matrix is an important step in the proposed method, since Wiener filtering is performed on the channel where the instrument has the most energy. If the algorithm picks a different channel for this step, in the separated audio files we can find more interference between instruments.

As described in Section 3.1, the estimation of the panning matrix depends on the number of nonoverlapping partials of the notes found in the score and their alignment with the audio. To that extent, the more nonoverlapping partials we have, the more robust the estimation.

Initially, we experimented with computing the panning matrix separately for each piece. In the case of Bruckner the piece is simply too short, and there are few nonoverlapping partials to yield a good estimation, resulting in errors in the panning matrix. Since the instrument setup is the same for Bruckner, Beethoven, and Mahler pieces (10 sources in the same position on the stage), we decided to jointly estimate the matrix for the concatenated audio pieces and the associated scores. We denote Setup 1 as the Mozart piece played by 8 sources and Setup 2 the Beethoven, Mahler, and Bruckner pieces played by 10 sources.

Since the panning matrix is computed using the score, different score information can yield very different estimation of the panning matrix. To that extent, we evaluate the influence of audio-to-score alignment, namely, the cases INT, NEX, Ali, T1, and T2, and the initialization with the ground truth score information, GT.

In Table 5, we list the instruments for which the algorithm picked the wrong channel. Note that, in the room setup generated with Roomsim, most of the instruments in Table 5 are placed close to other sources from the same family of instruments: for example, cello and double bass, flute with clarinet, bassoon, and oboe. In this case, the algorithm makes more mistakes when selecting the correct channel to perform source separation.

In the column GT of Table 5, we can see that having a perfectly aligned score yields less errors when estimating the panning matrix. Conversely, in a real-life scenario, we cannot rely on hand annotated score. In this case, for all columns of the Table 5 excluding GT, the best estimation is obtained by the combination of NEX and T1: taking the offset time as the onsets of the next note and then extending the score with a smaller window.

Furthermore, we compute the SDR values for the instruments in Table 5, column GT (clarinet and cello), if the separation was to be done in the correct channel or in the estimated channel. For Setup 1, the channel for clarinet is wrongly mistaken to WWL (woodwinds left), the correct one being WWR (woodwinds right), when we have a perfectly aligned score (GT). However, the microphones WWL and WWR are very close (see Figure 4), and they do not capture significant energy from other instrument sections and the SDR difference is less than 0.01 dB. However, in Setup 2, the cello is wrongly separated in the WWL channel, and the SDR difference between this audio and the audio separated in the correct channel is around −11 dB for each of the three pieces.

6.5.3. Source Separation. We use the evaluation metrics described in Section 6.4.3. Since there is a lot of variability

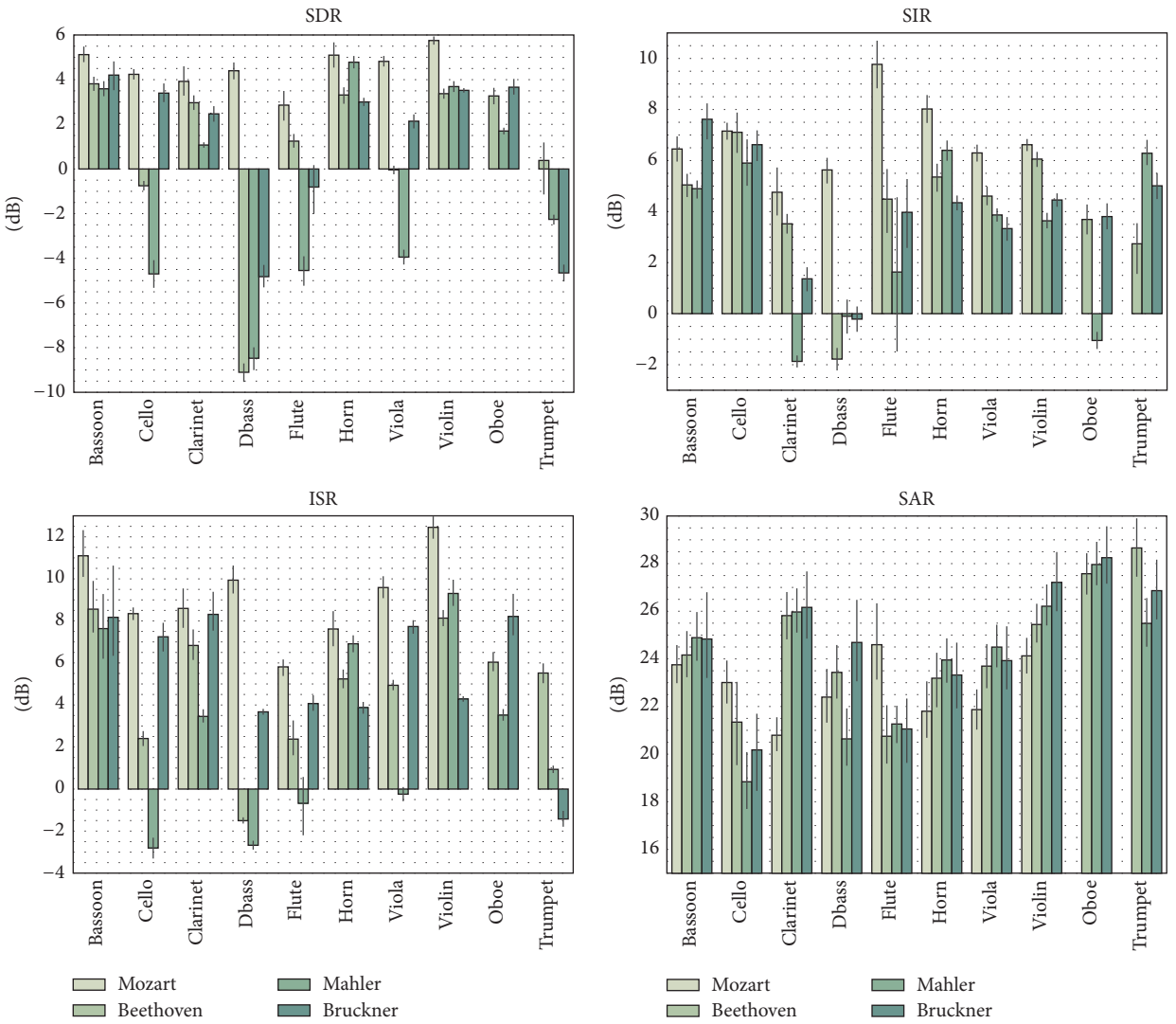


Figure 6: Results in terms of SDR, SIR, SAR, and ISR for the instruments and the songs in the dataset.

between the four pieces, it is more informative to present the results per piece rather than aggregating them.

First, we analyze the separation results per instruments in an ideal case. We assume that the best results for score-informed source separation are obtained in the case of a perfectly aligned score (GT). Furthermore, for this case, we calculate the separation in the correct channel for all the instruments, since, in Section 6.5.2, we could see that picking a wrong channel could be detrimental. We present the results as a bar plot in Figure 6.

As described in Section 6.1 and Table 1, the four pieces had different levels of complexity. In Figure 6, we can see that the more complex the piece is, the more difficult it is to achieve a good separation. For instance, note that cello, clarinet, flute, and double bass achieve good results in terms of SDR on Mozart piece but significantly worse results on other three pieces (e.g., 4.5 dB for cello in Mozart, compared to −5 dB in Mahler). Cello and double bass are close by in both of the setups, similarly for clarinet and flute, and we expect interference between them. Furthermore, these instruments usually share the same frequency range which can result in additional interference. This is seen in lower SIR values for double bass (5.5 dB SIR for Mozart, but −1.8, −0.1, and −0.2 dB SIR for the others) and flute.

An issue for the separation is the spatial reconstruction, measured by the ISR metric. As seen in (9), when applying the Wiener mask, the multichannel spectrogram is multiplied with the panning matrix. Thus, wrong values in this matrix can yield wrong amplitude values of the resulting signals.

This is the case for trumpet, which is allocated a close microphone in the current setup, and for which we expect a good separation. However, trumpet achieves a poor ISR (5.5, 1, and −1 dB) but has a good separation in terms of SIR and SAR. Similarly, other instruments as cello, double bass, flute, and viola face the same problem, particularly for the piece by Mahler. Therefore, a good estimation of the panning matrix is crucial for a good ISR.

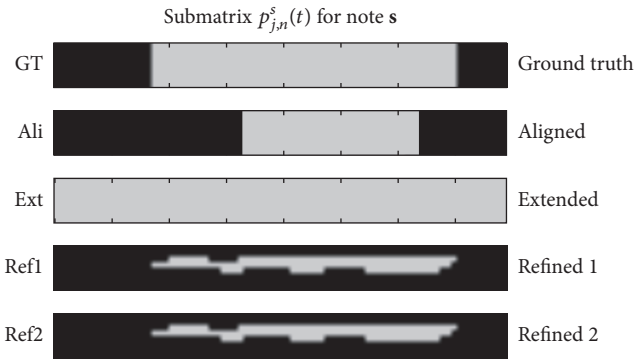


FIGURE 7: The test cases for initialization of score-informed source separation, for the submatrix $p^s_{j,n}(t)$.

A low SDR is obtained in the case of Mahler that is related to the poor results in alignment obtained for this piece. As seen in Table 4 for INT case, F-measure is almost 8% lower in Mahler than in other pieces, mainly because of the bad precision.

The results are considerably worse for double bass for the more complex pieces of Beethoven (−9.1 dB SDR), Mahler (−8.5 dB SDR), and Bruckner (−4.7 dB SDR), and, for further analysis, we consider it as an outlier, and we exclude it from the analysis.

Second, we want to evaluate the usefulness of note refinement in source separation. As seen in Section 4, the gains for NMF separation are initialized with score information or with a refined score as described in Section 4.2. A summary of the different initialization options and seen in Figure 7. Correspondingly, we evaluate five different initializations of the gains: the perfect initialization with the ground truth annotations (Figure 7 (GT)), the direct output of the score alignment system (Figure 7 (Ali)), the common practice of NMF gains initialization in state-of-the-art score-informed source separation [3, 5, 24] (Figure 7 (Ext)), and the refinement approaches (Figure 7 (Ref1 and Ref2)). Note that Ref1 is the refinement done with the method in Section 4.1 and Ref2 with the multichannel method described in Section 5.2.

We test the difference between the binarization thresholds 0.5 and 0.1, used in the refinement methods Ref1 and Ref2. One-way ANOVA on SDR results gives F-value = 0.0796 and p value = 0.7778, which shows no significant difference between both binarization thresholds.

The results for the five initializations, GT, Ref1, Ref2, Ext, and Ali, are presented in Figure 8, for each of the four pieces. Note that, for Ref1, Ref2, and Ext, we aggregate information across all possible outputs of the alignment: INT, NEX, T1, and T2. Analyzing the results, we note that the more complex the piece, the more difficult to separate between the instruments, the piece by Mahler having the worse results, and the piece by Bruckner a large variance, as seen in the error bars. For these two pieces, other factors as the increased polyphony within a source, the number of instruments (e.g., 12 violins versus 4 violins in a group), and the synchronization issues we described in Section 6.1 can increase the difficulty of separation up to the point that Ref1, Ref2, and Ext have a minimal improvement. To that extent, for the piece by Bruckner, extending the boundaries of the notes (Ext) does not achieve significantly better results than the raw output of the alignment (Ali).

As seen in Figure 8, having a ground truth alignment (GT) helps improving the separation, increasing the SDR with 1–1.5 dB or more for all the test cases. Moreover, the refinement methods Ref1 and Ref2 increase SDR for most of the pieces with the exception of the piece by Mahler. This is due to an increase of SIR and decrease of interferences in the signal. For instance, in the piece by Mozart, Ref1 and Ref2 increase the SDR with 1 dB when compared to Ext. For this piece, the difference in SIR is around 2 dB. Then, for Beethoven, Ref1 and Ref2 increase 0.5 dB in terms of SDR when compared to Ext and 1.5 dB in SIR. For Bruckner, solely Ref2 has a higher SDR; however SIR increases with 1.5 dB in Ref1 and Ref2. Note that not only do Ref1 and Ref2 refine the time boundaries of the notes, but also the refinement happens in frequency, because the initialization is done with the contours of the blobs, as seen in Figure 7. This can also contribute to a higher SIR.

Third, we look at the influence of the estimation of note offsets: INT and NEX, and the tolerance window sizes, T1 and T2, which accounts for errors in the alignment. Note that for this case we do not include the refinement in the results and we evaluate only the case Ext, as we leave out the refinement in order to isolate the influence of T1 and T2. Results are presented in Figure 9 and show that the best results are obtained for the interpolation of the offsets INT. This relates to the results presented in Section 6.5.1. Similarly to the analysis regarding the refinement, the results are worse for the pieces by Mahler and Bruckner, and we are not able to draw a conclusion on which strategy is better for the initialization, as the error bars for the ground truth overlap with the ones of the tested cases.

Fourth, we analyze the difference between the PARAFAC model for multichannel gains estimation as proposed in Section 5.1, compared with the single channel estimation of the gains in Section 3.4. We performed a one-way ANOVA on SDR and we obtain a F-value = 1.712 and a p-value = 0.1908. Hence, there is no significant difference between single channel and multichannel gain estimation, when we are not performing postprocessing of the gains using grain refinement. However, despite the new updates rule do not help, in the multichannel case we are able to better refine the gains. In this case, we aggregate information all over the channels, and blob detection is more robust, even to variations of the binarization threshold. To account for that, for the piece by Bruckner, Ref2 outperforms Ref1 in terms of SDR and SIR. Furthermore, as seen in Table 4 the alignment is always better for Ref2 than Ref1.

The audio excerpts from the dataset used for evaluation, as well as tracks separated with the ground truth annotated score are made available (http://repovizz.upf.edu/phenicx/anechoic_multi/).

7. Applications

7.1. Instrument Emphasis. The first application of our approach for multichannel score-informed source separation

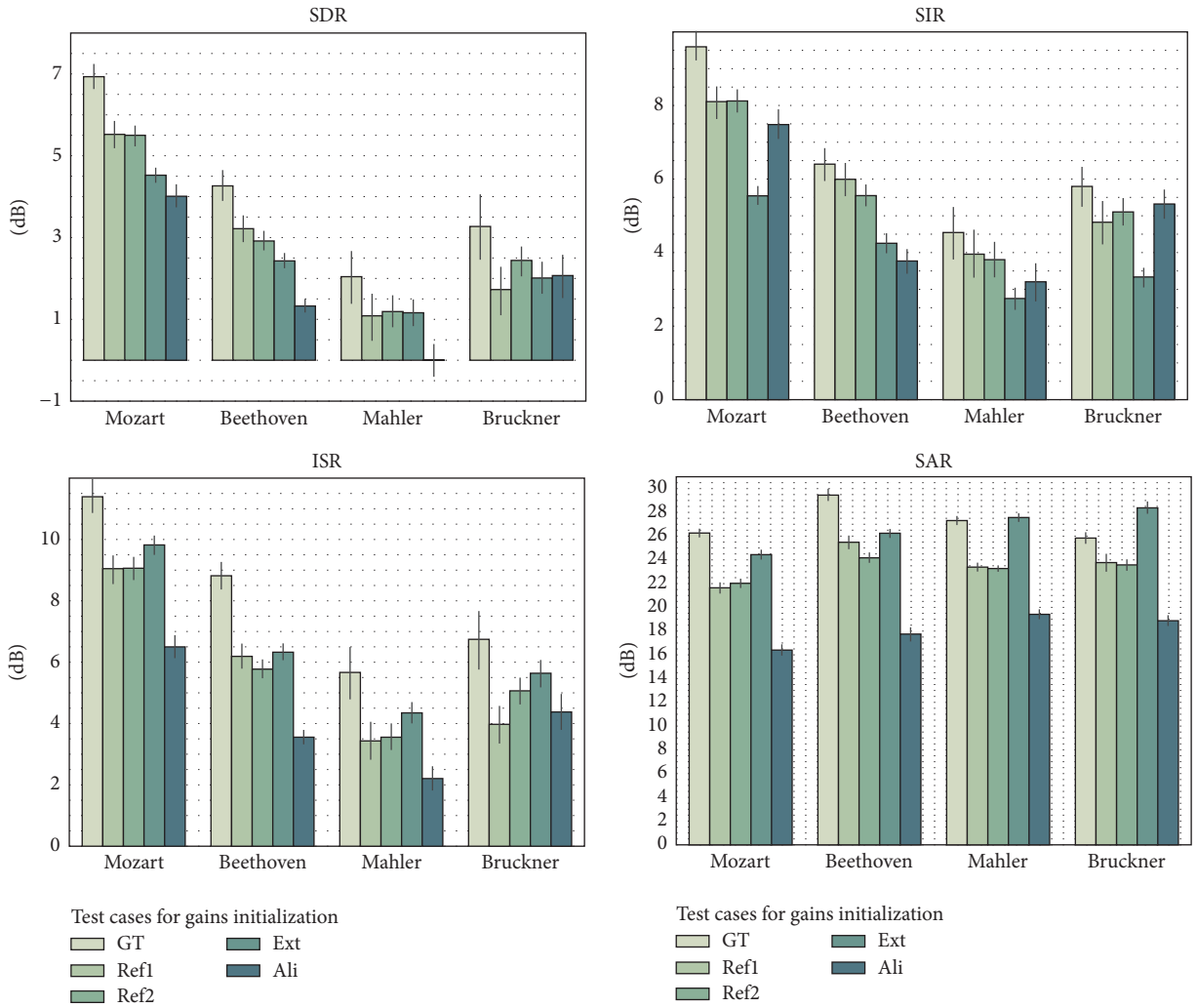


Figure 8: Results in terms of SDR, SIR, AR, and ISR for the NMF gains initialization in different test cases.

is *Instrument Emphasis*, which aims at processing multitrack orchestral recordings. Once we have the separated tracks, it allows emphasizing a particular instrument over the full orchestra downmix recording. Our workflow consists in processing multichannel orchestral recording from which the musical score is available. The multitrack recordings are obtained from a typical on-stage setup in a concert hall, where multiple microphones are placed on stage at certain distance of the sources. The goal is to reduce leakage of other sections, obtaining enhanced signal for the selected instrument.

In terms of system integration, this application has two parts. The front-end is responsible for interacting with the user in the uploading of media content and present the results to the user. The back-end is responsible for managing the audio data workflow between the different signal processing components. We process the audio files in batch estimating the signal decomposition for the full length. For long audio files, as in the case of symphonic recordings, the memory requirements can be too demanding even for a server infrastructure. Therefore, to overcome this limitation, audio files are split into blocks. After the separation has been performed, the blocks associated with each instrument are concatenated, resulting in the separated tracks. The separation quality is not degraded if the blocks have sufficient duration. In our case, we set the block duration to 1 minute. Examples of this application are found online (http://repovizz.upf.edu/phenicx/) and are integrated into the PHENICX prototype (http://phenicx.com/).

7.2. Acoustic Rendering. The second application for the separated material is augmented or virtual reality scenarios, where we apply a spatialization of the separated musical sources. Acoustic Rendering aims at recreating acoustically the recorded performance from a specific listening location and orientation, with a controllable disposition of instruments on the stage and the listener.

We have considered binaural synthesis as the most suitable spatial audio technique for this application. Humans locate the direction of incoming sound based on a number of cues: depending on the angle and distance between listener and source, the sound will arrive with a different intensity

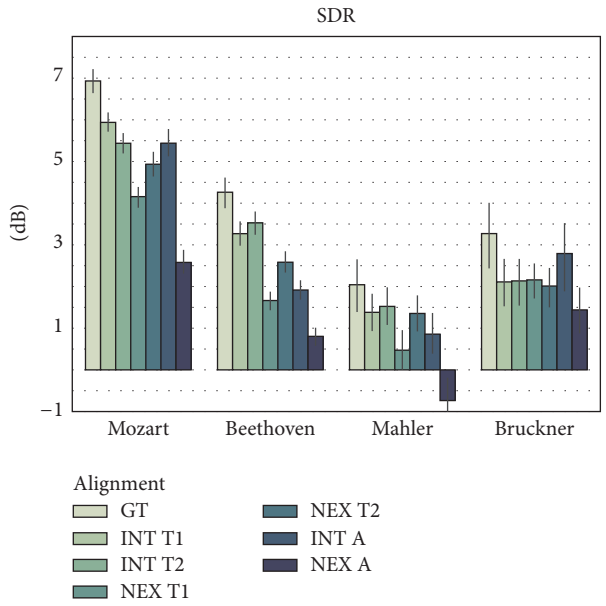


Figure 9: Results in terms of SDR, SIR, SAR, and ISR for the combination between different offset estimation methods (INT and Ext) and different sizes for the tolerance window for note onsets and offsets (T1 and T2). GT is the ground truth alignment and Ali is the output of the score alignment.

and at different time instances at both ears. The idea behind binaural synthesis is to artificially generate these cues to be able to create an illusion of directivity of a sound source when reproduced over headphones [43, 44]. For a rapid integration into the virtual reality prototype, we have used the noncommercial plugin for Unity3D of binaural synthesis provided by 3DCeption (https://twobigears.com/index.php).

Specifically, this application opens new possibilities in the area of virtual reality (VR), where video companies are already producing orchestral performances specifically recorded for VR experiences (e.g., the company WeMakeVR has collaborated with the London Symphony Orchestra and the Berliner Philharmoniker https://www.youtube.com/watch?v=ts4oXFmpacA). Using a VR headset with headphones, the Acoustic Rendering application is able to perform an acoustic zoom effect when pointing at a given instrument or section.

7.3. Source Localization. The third application aims to estimate the spatial location of musical instruments on stage. This application is useful for recordings where the orchestra layout is unknown (e.g., small ensemble performances) for the instrument visualization and Acoustic Rendering use-cases introduced above.

As for inputs for the source localization method we need the multichannel recordings and the approximate position of the microphones on stage. In concert halls, the recording setup consists typically of a grid structure of hanging overhead mics. The position of the overhead mics is therefore kept as metadata of the performance recording.

Automatic Sound Source Localization (SSL) methods make use of microphone arrays and complex signal processing techniques; however, undesired effects such as acoustic reflections and noise make this process difficult, being currently a hot-topic task in acoustic signal processing [45].

Our approach is a novel time difference of arrival (TDOA) method based on note-onset delay estimation. It takes the refined score alignment obtained prior to the signal separation (see Section 4.1). It follows two steps: first, for each instrument source, the relative time delays for the various microphone pairs are evaluated, and, then, the source location is found as the intersection of a pair of a set of half-hyperboloids centered around the different microphone pairs. Each half-hyperboloid determines the possible location of a sound source based on the measure of the time difference of arrival between the two microphones for a specific instrument.

To determine the time delay for each instrument and microphone pair, we evaluate a list of time delay values corresponding to all note onsets in the score and take the maximum of the histogram. In our experiment, we have a time resolution of 2.8 ms, corresponding to the Fourier transform hop size. Note that this method does not require time intervals in which the sources to play isolated as SRP-PHAT [45] and can be used in complex scenarios.

8. Outlook

In this paper, we proposed a framework for score-informed separation of multichannel orchestral recordings in distant-microphone scenarios. Furthermore, we presented a dataset which allows for an objective evaluation of alignment and separation (Section 6.1) and proposed a methodology for future research to understand the contribution of the different steps of the framework (Section 6.5). Then, we introduced several applications of our framework (Section 7). To our knowledge, this is the first time the complex scenario of orchestral multichannel recordings is objectively evaluated for the task of score-informed source separation.

Our framework relies on the accuracy of an audio-to-score alignment system. Thus, we assessed the influence of the alignment on the quality of separation. Moreover, we proposed and evaluated approaches to refining the alignment which improved the separation for three of the four pieces in the dataset, when compared to other two initialization options for our framework: the raw output of the score alignment and the tolerance window which the alignment relies on.

The evaluation shows that the estimation of panning matrix is an important step. Errors in the panning matrix can result into more interference in separated audio, or to problems in recovery of the amplitude of a signal. Since the method relies on finding nonoverlapping partials, an estimation done on a larger time frame is more robust. Further improvements on determining the correct channel for an instrument can take advantage of our approach for source localization in Section 7, provided that the method is reliable enough in localizing a large number of instruments. To that extent, the best microphone to separate a source is the closest one determined by the localization method.

When looking at separation across the instruments, viola, cello, and double bass were more problematic in the more complex pieces. In fact, the quality of the separation in our experiment varies within the pieces and the instruments, and future research could provide more insight on this problem. Note that increasing degree of consonance was related to a more difficult case for source separation [17]. Hence, we could expect a worse separation for instruments which are harmonizing or accompanying other sections, as the case of viola, cello, and double bass in some pieces. Future research could find more insight on the relation between the musical characteristics of the pieces (e.g., tonality and texture) and source separation quality.

The evaluation was conducted on the dataset presented in Section 6.1. The creation of the dataset was a very laborious task, which involved annotating around 12000 pairs of onsets and offsets, denoising the original recordings and testing different room configurations in order to create the multichannel recordings. To that extent, annotations helped us to denoise the audio files, which could then be used in score-informed source separation experiments. Furthermore, the annotations allow for other tasks to be tested within this challenging scenario, such as instrument detection, or transcription.

Finally, we presented several applications of the proposed framework related to Instrument Emphasis or Acoustic Rendering, some of which are already at the stage of functional products.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank all the partners of the PHENICX consortium for a very fruitful collaboration. Particularly, they would like to thank Agustin Martorell for his insights on the musicology side and correcting the scores and Oscar Mayor for his help regarding Repovizz and his contribution on the applications. This work is partially supported by the Spanish Ministry of Economy and Competitiveness under CASAS Project (TIN2015-70816-R).

References

[1] E. Gómez, M. Grachten, A. Hanjalic et al., "PHENICX: performances as highly enriched aNd Interactive concert experiences," in *Proceedings of the SMAC Stockholm Music Acoustics Conference 2013 and SMC Sound and Music Computing Conference*, 2013.

[2] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 129–132, Kyoto, Japan, March 2012.

[3] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 888–891, Vancouver, Canada, May 2013.

[4] Z. Duan and B. Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[5] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, pp. 45–48, May 2011.

[6] J. J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. J. Rodriguez-Serrano, "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, article 184, 2013.

[7] T. Pratzlich, R. M. Bittner, A. Liutkus, and M. Muller, "Kernel Additive Modeling for interference reduction in multi-channel music recordings," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '15)*, pp. 584–588, Brisbane, Australia, April 2014.

[8] S. Ewert, B. Pardo, M. Mueller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: an overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[9] F. J. Rodriguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, and J. J. Carabias-Orti, "Online score-informed source separation with adaptive instrument models," *Journal of New Music Research*, vol. 44, no. 2, pp. 83–96, 2015.

[10] F. J. Rodriguez-Serrano, S. Ewert, P. Vera-Candeas, and M. Sandler, "A score-informed shift-invariant extension of complex matrix factorization for improving the separation of overlapped partials in music recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '16)*, pp. 61–65, Shanghai, China, 2016.

[11] M. Miron, J. J. Carabias, and J. Janer, "Improving score-informed source separation for classical music through note refinement," in *Proceedings of the 16th International Society for Music Information Retrieval (ISMIR '15)*, Málaga, Spain, October 2015.

[12] A. Arzt, H. Frostel, T. Gadermaier, M. Gasser, M. Grachten, and G. Widmer, "Artificial intelligence in the concertgebouw," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pp. 165–176, Buenos Aires, Argentina, 2015.

[13] J. J. Carabias-Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada, "An audio to score alignment framework using spectral factorization and dynamic time warping," in *Proceedings of the 16th International Society for Music Information Retrieval (ISMIR '15)*, Malaga, Spain, 2015.

[14] Ö. Izmirli and R. Dannenberg, "Understanding features and distance functions for music sequence alignment," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR '10)*, pp. 411–416, Utrecht, The Netherlands, 2010.

[15] M. Goto, "Development of the RWC music database," in *Proceedings of the 18th International Congress on Acoustics (ICA '04)*, pp. 553–556, Kyoto, Japan, April 2004.

[16] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1869–1872, IEEE, Taipei, Taiwan, April 2009.

[17] J. J. Burred, *From sparse models to timbre learning: new methods for musical source separation [Ph.D. thesis]*, 2008.

[18] F. J. Rodriguez-Serrano, J. J. Carabias-Orti, P. Vera-Candeas, T. Virtanen, and N. Ruiz-Reyes, "Multiple instrument mixtures source separation evaluation using instrument-dependent NMF models," in *Proceedings of the 10th international conference on Latent Variable Analysis and Signal Separation (LVA/ICA '12)*, pp. 380–387, Tel Aviv, Israel, March 2012.

[19] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and EM algorithm," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 5510–5513, March 2010.

[20] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO '09)*, pp. 15–19, Glasgow, UK, August 2009.

[21] J. J. Bosch, K. Kondo, R. Marxer, and J. Janer, "Score-informed and timbre independent lead instrument separation in real-world scenarios," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO '12)*, pp. 2417–2421, August 2012.

[22] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR '14)*, Taipei, Taiwan, October 2014.

[23] E. K. Kokkinis and J. Mourjopoulos, "Unmixing acoustic sources in real reverberant environments for close-microphone applications," *Journal of the Audio Engineering Society*, vol. 58, no. 11, pp. 907–922, 2010.

[24] S. Ewert and M. Müller, "Score-informed voice separation for piano recordings," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR '11)*, pp. 245–250, Miami, Fla, USA, October 2011.

[25] B. Niedermayer, *Accurate audio-to-score alignment-data acquisition in the context of computational musicology [Ph.D. thesis]*, Johannes Kepler University Linz, Linz, Austria, 2012.

[26] S. Wang, S. Ewert, and S. Dixon, "Compensating for asynchronies between musical voices in score-performance alignment," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '15)*, pp. 589–593, April 2014.

[27] M. Miron, J. Carabias, and J. Janer, "Audio-to-score alignment at the note level for orchestral recordings," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR '14)*, Taipei, Taiwan, 2014.

[28] D. Fitzgerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation. acoustics, speech and signal processing," in *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal (ICASSP '06)*, Toulouse, France, 2006.

[29] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," in *Exploring Music Contents: 7th International Symposium, CMMR 2010, Málaga, Spain, June 21–24, 2010. Revised Papers*, vol. 6684 of *Lecture Notes in Computer Science*, pp. 102–115, Springer, Berlin, Germany, 2011.

[30] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica United with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[31] D. Campbell, K. Palomäki, and G. Brown, "A Matlab simulation of shoebox room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.

[32] O. Mayor, Q. Llimona, M. Marchini, P. Papiotis, and E. Maestre, "RepoVizz: a framework for remote storage, browsing, annotation, and exchange of multi-modal data," in *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*, pp. 415–416, Barcelona, Spain, October 2013.

[33] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[34] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[35] M. Nixon, *Feature Extraction and Image Processing*, Elsevier Science, 2002.

[36] R. Parry and I. Essa, "Estimating the spatial position of spectral components in audio," in *Independent Component Analysis and Blind Signal Separation: 6th International Conference, ICA 2006, Charleston, SC, USA, March 5–8, 2006. Proceedings*, J. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, Eds., vol. 3889 of *Lecture Notes in Computer Science*, pp. 666–673, Springer, Berlin, Germany, 2006.

[37] P. Papiotis, M. Marchini, A. Perez-Carrillo, and E. Maestre, "Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data," *Frontiers in Psychology*, vol. 5, article 963, 2014.

[38] M. Mauch and S. Dixon, "PYIN: a fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '14)*, pp. 659–663, Florence, Italy, May 2014.

[39] F. J. Cañadas-Quesada, P. Vera-Candeas, D. Martínez-Muñoz, N. Ruiz-Reyes, J. J. Carabias-Orti, and P. Cabanas-Molero, "Constrained non-negative matrix factorization for score-informed piano music restoration," *Digital Signal Processing*, vol. 50, pp. 240–257, 2016.

[40] A. Cont, D. Schwarz, N. Schnell, and C. Raphael, "Evaluation of real-time audio-to-score alignment," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR '07)*, pp. 315–316, Vienna, Austria, September 2007.

[41] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[42] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[43] D. R. Begault and E. M. Wenzel, "Headphone localization of speech," *Human Factors*, vol. 35, no. 2, pp. 361–376, 1993.

[44] G. S. Kendall, "A 3-D sound primer: directional hearing and stereo reproduction," *Computer Music Journal*, vol. 19, no. 4, pp. 23–46, 1995.

[45] A. Martí Guerola, *Multichannel audio processing for speaker localization, separation and enhancement*, Universitat Politècnica de València, 2013.

Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 4062786, 9 pages
http://dx.doi.org/10.1155/2016/4062786

Research Article

A Russian Keyword Spotting System Based on Large Vocabulary Continuous Speech Recognition and Linguistic Knowledge

Valentin Smirnov,[1] Dmitry Ignatov,[1] Michael Gusev,[1] Mais Farkhadov,[2] Natalia Rumyantseva,[3] and Mukhabbat Farkhadova[3]

[1]*Speech Drive LLC, Saint Petersburg, Russia*
[2]*V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia*
[3]*RUDN University, Moscow, Russia*

Correspondence should be addressed to Mais Farkhadov; mais.farhadov@gmail.com

Received 11 July 2016; Revised 27 October 2016; Accepted 14 November 2016

Academic Editor: Alexey Karpov

The paper describes the key concepts of a word spotting system for Russian based on large vocabulary continuous speech recognition. Key algorithms and system settings are described, including the pronunciation variation algorithm, and the experimental results on the real-life telecom data are provided. The description of system architecture and the user interface is provided. The system is based on CMU Sphinx open-source speech recognition platform and on the linguistic models and algorithms developed by Speech Drive LLC. The effective combination of baseline statistic methods, real-world training data, and the intensive use of linguistic knowledge led to a quality result applicable to industrial use.

1. Introduction

The need to understand business trends, ensure public security, and improve the quality of customer service has caused a sustainable development of speech analytics systems which transform speech data into a measurable and searchable index of words, phrases, and paralinguistic markers. Keyword spotting technology makes a substantial part of such systems. Modern keyword spotting engines usually rely on either of three approaches, namely, phonetic lattice search [1, 2], word-based models [3, 4], and large vocabulary speech recognition [5]. While each of the approaches has got its pros and cons [6] the latter starts to be prominent due to public availability of baseline algorithms, cheaper hardware to run intensive calculations required in LVCSR and, most importantly, high-quality results.

Most recently a number of innovative approaches to spoken term detection were offered such as various recognition system combination and score normalization, reporting 20% increase in spoken term detection quality (measured as ATWV) [7, 8]. Deep neural networks application in LVCSR is starting to achieve wide adoption [9]. Thanks to the IARPA Babel program aimed at building systems that can be rapidly applied to any human language in order to provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech [10] in recent years wide research has been held to develop technologies for spoken term detection systems for low-resource languages. For example, [11] describes an approach for keyword spotting in Cantonese based on large vocabulary speech recognition and shows positive results of applying neural networks to recognition lattice rescoring. Reference [12] provides an extensive description of modern methods used to build a keyword spotting system for 10 low-resource languages with primary focus on Assamese, Bengali, Haitian Creole, Lao, and Zulu. Deep neural network acoustic models are used both as feature extractor for a GMM-based HMM system and to compute state posteriors and convert them into scaled likelihoods by normalizing by the state priors. Data augmentation via using multilingual bottleneck features is offered (the topic is also covered in [13]). Finally language independent and unsupervised acoustic models are trained

for languages with no training data. An average MTWV reported for these languages ranges from 0.22 for Zulu to 0.67 for Haitian Creole. In [14] the use of recurrent neural networks for example-based word spotting in real time for English is described. Compared to more widespread text-based systems, this approach makes use of spoken examples of a keyword to build up a word-based model and then do the search within speech data. As an alternative to hybrid ANN-HMM approaches authors in [15] offer a pure NN based keyword search system for conversational telephone speech in Vietnamese and Lao. For Vietnamese the "pure" NN system provides ATWV comparable with that reported for a baseline hybrid system while working significantly faster (real-time factor 3.4 opposed to 5.3 for a hybrid system).

As high-quality language modeling is an indispensable part of any modern keyword spotting system, a lot of effort is now aimed at improving LMs. One of the most recent trends is to use web data in training. The advent of the Internet has provided rich amount of data to be easily available for speech recognition community [16]. This is of particular interest for low-resource languages and among most recent improvements [17] suggests an approach to effectively deal with the challenge of normalizing and filtering the web data for keyword spotting. Two methods are offered, one using perplexity ranking and the other using out-of-vocabulary words detection. This resulted in more than 2% absolute improvement in ATWV across 5 low-resourced languages. Reference [18] covers the aspect of augmenting baseline LMs with carefully chosen web data, showing that blogs and movie subtitles are more relevant for language modeling of conversational telephone speech and help to obtain large reductions in out-of-vocabulary keywords.

Russian research in the domain of speech recognition falls in line with global scientific trends. It is noteworthy however that most frequently the research is conducted to meet a more general target of creating LVCSR systems per se with no specific focus on spoken term detection. The most well-known systems include Yandex SpeechKit [19] used to recognize spoken search queries via web and mobile applications, real-time speech recognition system by Speech Technology Center [20] used for transcribing speech in the broadcasting news, LVCSR system developed by SPIIRAS [21, 22] used for recognizing speech in multimodal environments, and speech recognition system by scientific institute Specvuzavtomatika [23] based on deep neural networks.

Current paper presents the results of the ongoing research underlying the commercial software for speech analytics. The software design follows the concept of a minimum viable product, which motivates incremental complication of the technology while the product evolves. Such approach motivated us to rely on generally available open-source toolkits and a number of readily available knowledge-based methods developed under our previous studies.

Sections 2 and 3 outline the overall setup of applying LVCSR technology to keyword spotting for Russian telephone conversational speech, including the key system parameters and the description of experiments run to assess the quality and performance of the system. Special focus is given to linguistic components used at the training and spotting stage. Section 4 describes the off-the-shelf speech analytics system developed using the ideas and results discussed in this paper.

2. Key System Parameters

The system described in the paper is intended to be used to perform keyword search in telephone conversational speech. The system is provided both as SDK to be integrated with speech recording systems and as a stand-alone MS Windows application. The system is created on top of CMU Sphinx [24]; this framework has been chosen due to its simplicity and licensing model which allows for freely using the code in commercial applications. Following the idea of minimum viable product we mostly use the standard settings across all system modules. 13 MFCCs with their derivatives and acceleration are used in the acoustic front-end; triphone continuous density acoustic models are trained on around 200 hours of telephone-quality Russian speech (8 kHz, 8 bit, Mono) recorded by 200 native speakers. 5-state HMMs are used with diagonal covariation matrix, and CART (classification and regression trees) algorithm is used to cluster acoustic models into 9000 senones, each senone being described by 10 to 30 Gaussians. Texts in the training database for language models are transcribed automatically with a morphological and contextual linguistic processor [25]. A set of transcription variation rules are applied. Unigram and bigram language models are trained on hundreds of thousands of modern Russian e-books generally available on the Internet. Decoder makes use of a standard CMU Sphinx token-passing algorithm with pruning methods widely employed in the system setup including maximum beam width, word insertion penalty, and acoustic likelihood penalty.

The core novelty of the system is granted by extensive use of linguistic knowledge on both the training and spoken term detection steps. The system uses a linguistic processor with built-in information on Russian morphology which helps to generate high-quality transcriptions for any word form and thus train more viable acoustic models. The same processor is used to generate various forms of words which ensures better spoken term detection on the spotting step. A rule-based transcription variation algorithm is applied to generate alternative phoneme sequences. Ultimately on the language modeling step the texts are automatically prefiltered by the type of text to let only dialogues stay in the training corpus.

3. Algorithms, Associated Problems, and Solution

3.1. Acoustic Front-End. While throughout the system standard MFCCs are used, an additional effort was required to make the front-end work for keyword spotting in a real-world application. First, audio files to be analyzed are chunked into 10-second long chunks in order to split the decoding process over multiple CPUs. An overlap of 1 second is used to guarantee that a keyword is not truncated between two subsequent speech segments. Further on, a parsing algorithm

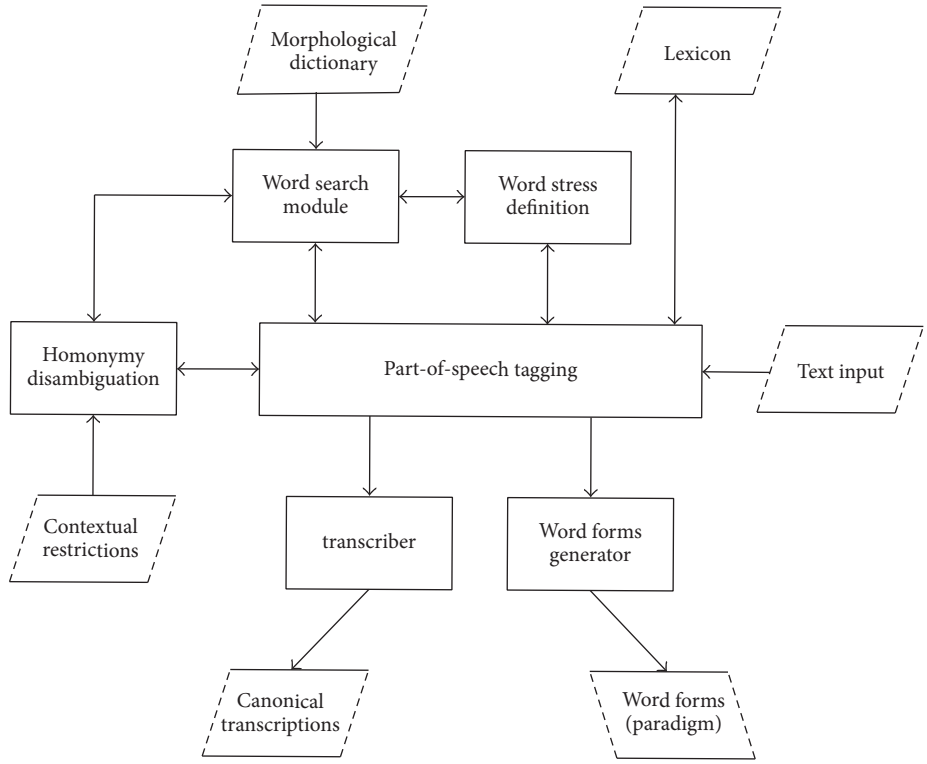


Figure 1: Linguistic processor.

is applied to combine partial decoding results into a single file in order to avoid redundant false alarms. The future plan is to use VAD to divide the audio stream into phrases which would better suit the LVCSR-based approach used in this paper; however, our current VAD implementation has shown worse results, hence the use of chunks of equal length.

3.2. Acoustic Modeling, Grapheme-to-Phoneme Conversion, and a Transcription Variation Algorithm. The system discussed in the paper is intended to be used in real-world telephone environment under low sound quality conditions. To cover this requirement the acoustic model is trained on real-world data encountering the telephone channel quality speech in Russian telephone networks. Continuous density HMMs are used, resulting in a representative set of 9000 senones each described with a GMM with 10–30 components.

Under our previous research [25] a linguistic processor has been developed which makes use of information on morphological characteristics of around 600 000 Russian words (see the structure on Figure 1) to transcribe words and generate forms of words. Processor parses the text and defines the part of speech for every word in the sentence; then the word stress is defined, and a set of preprogrammed contextual grapheme-to-phoneme rules is applied to derive a canonical ("ideal") word transcription.

The current state of the art for transcribing words in speech recognition systems is to use statistical grapheme-to-phoneme converters [26, 27]. The research has been held on combining various techniques, for example, in [28] Conditional Random Fields and Joint-Multigram Model are used to bring an additional improvement in quality. Studies have been done [29, 30] to introduce weighted finite state transducers to grasp the probabilities of in-word phonetic sequences. Altogether these studies outline the key advantages of probabilistic approach compared to knowledge-based methods, namely, language independency (easily ported to a new language), ability to generalize and provide transcriptions to new (out-of-vocabulary) words, and the need of a smaller footprint of linguistic data (and hence effort) to train a grapheme-to-phoneme converter.

On the other hand, the majority of the results shared in cited studies relate to languages characterized with low number of word forms (e.g., English and French). Meanwhile Russian is a highly inflectional language with a word stress depending on the exact word form in the paradigm and a high frequency of homonymy also affecting word stress and thus being a source for potential transcription errors [31]. This means that one needs a much bigger hand-crafted lexicon to train a high-quality probabilistic grapheme-to-phoneme converter for Russian. This obstacle together with the concept of minimum viable product described above motivated us to set probabilistic grapheme-to-phoneme conversion as a target for our future research and to use a readily available high-quality knowledge-based linguistic processor instead. Another important factor which guided this choice is the ability to disambiguate homonymy and to generate word forms (to be discussed later on).

The key element of the acoustic model training process is transcription variation. Every phrase used to train the models receives several alternative transcriptions by applying

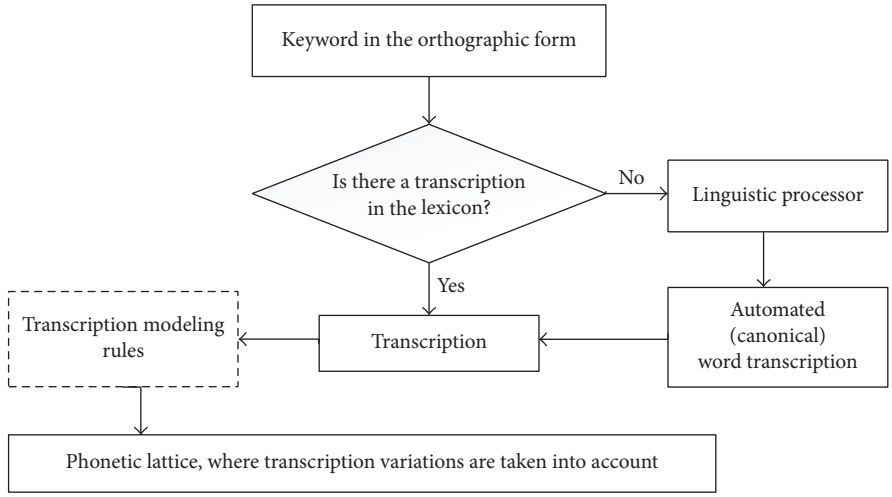


Figure 2: Transcription variation algorithm.

a set of predefined linguistic rules. Then on the training step CMU Sphinx train module chooses the best alternative which maximizes expectation. The experiments showed a 4% absolute increase in keyword detection rate achieved thanks to such implementation (please refer to Section 4 for more details on experiments). At the moment the rules are derived manually based on linguistic knowledge. The list of rules is then embedded in the recognizer which is run on a training dataset to define which rules provide for quality improvement and should be kept in the production system. As the next step of our research we plan to develop a sufficient corpus to train these rules automatically.

The ultimate list of transcription variation rules chosen on the training set contains 30 contextual phoneme-to-phoneme and grapheme-to-phoneme mappings based on both from the modern research on Russian spontaneous speech [32] and from the authors' proper experiments with real-life data audio analysis. The main steps of the transcription variation algorithm are outlined below (please also refer to Figure 2):

(1) A textual annotation of the trained database is loaded.

(2) If the word is not in the lexicon (used mainly for foreign words and named entities), automatic transcriber is launched which makes use of the digitized dictionary of the Russian language, containing 600 thousand words with morphological information and a part-of-speech (POS) tagger. As a result of this stage the word stress is assigned to the right syllable of every word.

(3) Automated, or canonical, transcription is generated by applying context-dependent letter-to-phone rules.

(4) Pronunciation variation is executed by iteratively applying a separate set of phoneme-to-phoneme and grapheme-to-phoneme transcription modeling rules to the canonical transcriptions.

It is well known that knowledge-based rules, being "laboratory" in origin, may happen to be inadequate when confronted with real-world data. However this was our intent to check this critical assumption on our test material. Moreover, during the past decades, Russian phonetics has undergone a general shift from laboratory speech to fully spontaneous [32, 33], and the rules we use are based on vast research on spontaneous speech traits.

The rules are divided into two main groups. The first contains substitution, deletion, and insertion rules, which apply to initial phonetic transcriptions. Here are some examples of such rules:

(i) [@] ("schwa"), followed by a consonant, is deleted in the unstressed position after stressed syllable.

(ii) [f] is deleted from consonant sequence [fs] + (any) unvoiced consonant.

(iii) Affricates [c] and [t$\int$'] are substituted by fricatives [s] and [$\int^{j}$], respectively (sign j denotes that a consonant is palatalized).

(iv) Sonorant [j] is deleted before unstressed vowel at the beginning of words.

(v) Noise stops (e.g., [p], [t], [p^{j}], and [t^{j}]) are deleted in the final position after vowels due to implosive pronunciation (i.e., without burst following articulators closure).

The second group of rules makes use of both morphological and orthographical level of linguistic representation. Hence, this is not correction to initial transcriptions (phoneme-to-phoneme rules) but a separate set of grapheme-to-phoneme rules. Here are some examples:

(i) [@j@] and [uju] in unstressed inflections of adjectives "–ая" and "–ую" are changed to [@e] and [u], respectively.

(ii) [@v@], [ɨv@], and [iv@] in unstressed noun inflections "–ого" and "–его" are changed to [@@], [ɨ@], and [i@].

(iii) [@t] in verb inflections "–ат" is changed to [it].

For frequent words we also added another set of rules, which generate simplified pronunciation, which is common to informal spontaneous speech. These include [d^j] and [v] deletion in intervocalic position, [s^jt^j] changing to [s^j], and so forth.

3.3. Language Models and the Choice of Relevant Content to Train Them. Initially language models have been trained with a few gBs of user-generated content to be found on the Internet, including public forums, social networks, and chats. The idea behind this was that such content would better represent spontaneous speech and thus ensuring more sustainable keyword spotting results. However the experiments have shown that such linguistic material occurred to bear an intrinsic drawback, because it contains enormous number of spelling errors which led to a statistic bias and wrong lemmas to appear in the lexicon. Hence a decision was taken to rely on standard and error-free texts derived from a wide range of books of different genres available on the Internet. Only books by modern authors (1990s and later) were chosen to reflect current traits of Russian speech. However only the dialogues have been extracted from such books to guarantee the "live" style of communication, which is characteristic of real-world telephone speech. 2 gB of raw text data was used as a result to train a unigram and bigram language models containing 600 000 Russian lemmas. The LMs were trained using SRILM toolkit [34] with Good-Turing discounting algorithm applied.

Current research in the domain of language modeling is focused on applying deep neural networks and high-level LM rescoring [35]. In our case there is insufficient data to train such models, which motivated us to shift to much simpler models. As outlined in Section 3.4 we do not rely on the most probable word sequence in the recognition result to detect keywords; rather we want to generate as diverse and "rough" lattice on the indexing step to guarantee high probability for the spoken term detection. Simple bigram/unigram language modeling fits this aim quite well.

3.4. Decoding, Word Spotting, and Automated Word Form Generation. The main idea behind using LVCSR to find keywords is to transform speech signal into a plain text and then search for the word in this text. However due to diverse types of communication context in the telephone conversational speech it is not viable to use the top decoding result per se. Rather, it makes sense to parse the resulting recognition lattice to find every possible node with the keyword. Hence speech is first indexed into recognition lattices; the keyword search is performed on-demand at a later stage.

To improve spotting results we make intensive use of the linguistic processor described above. When a word is entered as a search query its forms are automatically generated by addressing the morphological dictionary (see Figure 1) and a set of variants are derived for the word which are then searched in the lattice and appear in the recognition results list. For example, when the word "кусок" is to be searched (Russian word for "a piece") all the words containing this sequence will be searched within a recognition lattice; hence the user will be able to spot the words "куска" and "куском" and so forth. Since Russian is an inflectional language numerous forms are available for one word. Consequently low-order (unigram and bigram) language models used in our system cause the recognizer to make errors in the word endings. The simple idea described above helps to avoid errors and achieve much better results.

Table 1: Experimental results.

Parameter	Value
MTWV	0.37
RTF	2.0

4. Experimental Results

The system described hereby is intended to be used in real-world applications to analyze telephone-quality speech. To test it a 10-hour database including the recordings of dialogues of around 50 speakers has been recorded using the hardware and software of SpRecord LLC (http://www.sprecord.ru/). 1183 different keywords are searched within the database. The signal-to-noise ratio falls between 5 and 15 dB, reflecting an adverse real telephone channel environment.

Maximum Term-Weighted Value (MTWV) is a predicted metric corresponding to the best TWV for all values of the decision threshold; θ (see formula (1)) and real-time factor (RTF) metrics (formula (2)) are used to evaluate system performance; the former metric reflects the quality of word spotting, and the latter reflects its speed. RTF parameter is calculated on 1 CPU unit of 3 gHz. The results are shown in Table 1.

$$\mathrm{TWV}(\theta) = 1 - \left[P_{\mathrm{Miss}}(\theta) + \beta \cdot P_{\mathrm{FA}}(\theta)\right]. \quad (1)$$

θ is the threshold used to determine a hit or a miss, and β is the weight that accounts for the presumed prior probability of a term and the relative costs of misses and false alarms are equal to 0.999 in our study.

$$\mathrm{RTF} = \frac{T_{\mathrm{proc}}}{T_{\mathrm{set}}}. \quad (2)$$

T_{proc} is the time spent on processing the file, and T_{set} is the duration of the test set.

In order to understand whether these results correspond to the current state of the art we compared them to the result of another scientific group for spoken term detection in telephone conversational of another underresourced language (Cantonese) [11]. What we saw is that our results in terms of keyword search quality fall in between those reported for Cantonese when GMMs are used in the acoustic model and are slightly worse when deep neural networks are used (MTWV 0.335 and 0.441, resp.). As for the real-time factor our results outperform those reported in [14], which may be attributed to a relatively small number of Gaussians we use per senone.

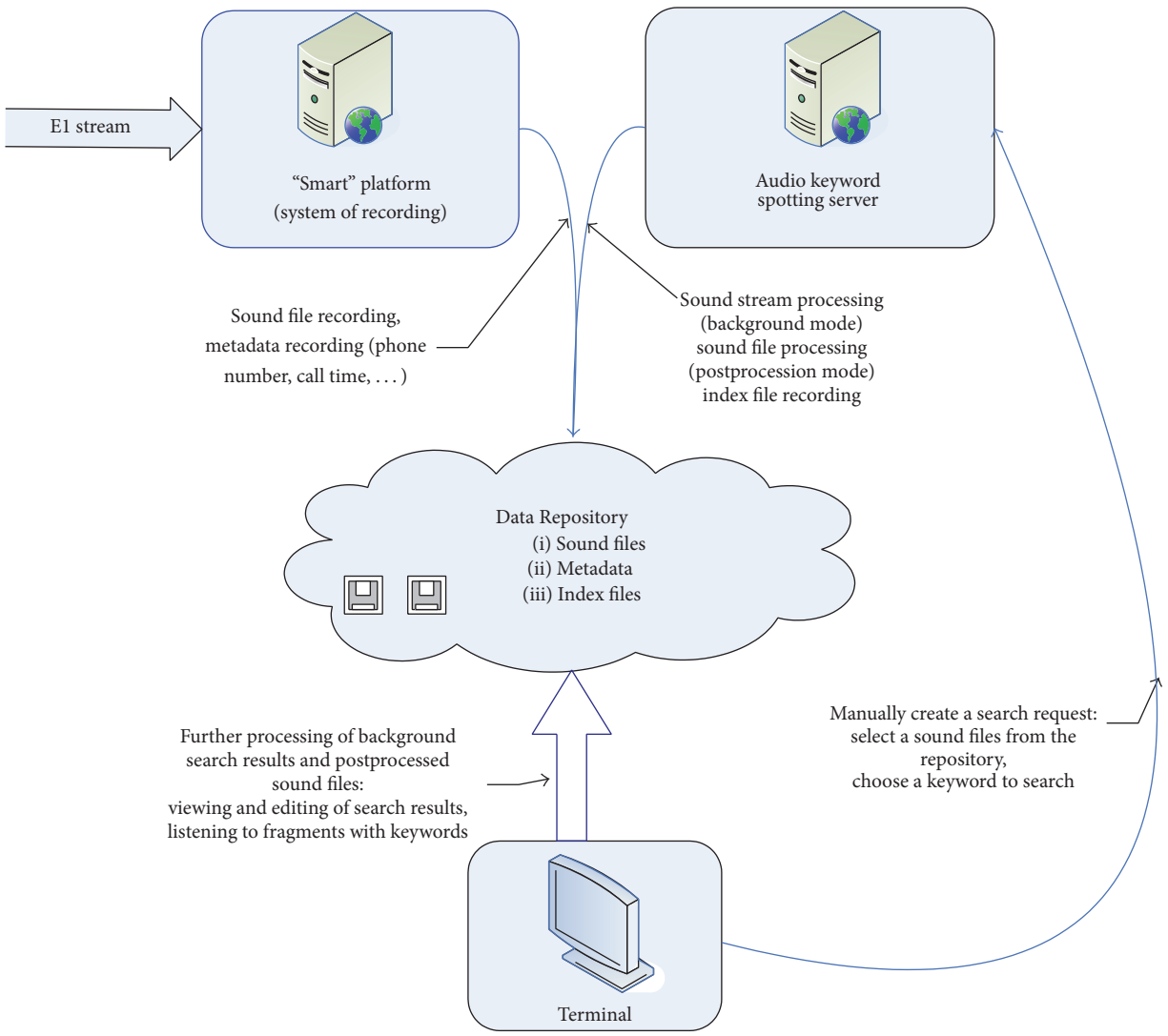


Figure 3: System architecture.

5. System Architecture and User Interface

5.1. Principal Components. The algorithms described in Section 2 were used in creating "ANALYZE" software—an industrial speech analytics system. Figure 3 outlines the key system components: word spotting server, terminal, and data repository. Word spotting server processes speech data and saves index with positions of searched keywords into the database. The terminal is used to schedule or launch immediate search queries and to view the search results. The search is performed in two steps: first, the lattice with speech recognition results is generated for each wave file; second, the keyword is found via a substring search within this lattice. The data repository contains both speech files and corresponding indices.

5.2. User Interface. The key problems of human-machine interaction within speech analytics systems, including accurate treatment of the keyword spotting results, and the role of in the optimization of workflows in modern organizations are reflected in [36–39]. Figure 4 outlines the user interface of the ANALYZE software which has been developed based on use-cases validated with the end-users. Usability and use-case integrity were tested in the real-world environment. All settings are available in 1-2 clicks; real-time reporting is shown on the screen; navigation panel provides access to all needed functions. Table 1 with search results provides easy filtering and listening modes. Figure 3 presents the main board of the system's user interface.

An essential benefit of the software is the ability to work in real-time mode on the workstations with limited resources, which makes it worthy for small organizations with a fraction of telephone lines in use.

6. Conclusion and Further Plans

A keyword spotting system for Russian based on LVCSR has been described in this paper. General availability of

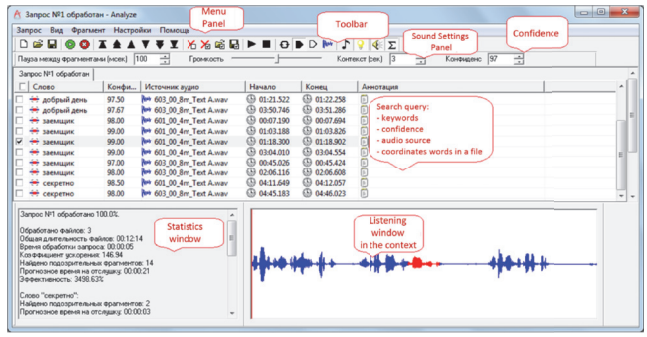


Figure 4: The user interface of the "ANALYZE" software.

open-source software made it easy to be implemented and linguistic modules helped to improve the system quality, while representative training and test data ensured the applicability of the system to real-world problems.

The ongoing research is aimed at further tuning the acoustic and language models, trying probabilistic frameworks for grapheme-to-phoneme conversion, data-driven transcription variation, introducing noise compensation and pause detection into the front-end and at creating specific confidence measures to minimize false alarms which are caused by frequent words in the language model.

In building our automated keyword spotting system based on large vocabulary continuous speech recognition we relied on the results of the scientific community, namely, the open-source software CMU Sphinx for acoustic modeling and decoding and SRILM for language modeling. At the same time the system has several technological advantages: the use of linguistic knowledge in training and decoding, namely, a morphological parser of texts and transcription variation to generate word transcriptions, transcription variation rules, and automated generation of word forms on the spotting step; real-world industrial data used to train acoustic models; accurate language modeling achieved via cautious choice of training data; real-time operation mode on limited computer resources.

We believe that high-quality automated keyword spotting system based on large vocabulary continuous speech recognition for online speech data analysis can be used both as a technological platform to create effective integrated systems for monitoring and as a ready-to-use solution to monitor global information space.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank SpRecord LLC authorities for providing real-world telephone-quality data used in training and testing of the keyword spotting system described in this paper.

References

[1] T. J. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, pp. 659–664, Kyoto, Japan, December 2007.

[2] D. A. James and S. J. Young, "A fast lattice-based approach to vocabulary independent wordspotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 377–380, Adelaide, Australia, 1994.

[3] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát, and J. Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," in *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12–15, 2005. Proceedings*, vol. 3658 of *Lecture Notes in Computer Science*, pp. 302–309, Springer, Berlin, Germany, 2005.

[4] M. Yamada, M. Naito, T. Kato, and H. Kawai, "Improvement of rejection performance of keyword spotting using anti-keywords derived from large vocabulary considering acoustical similarity to keywords," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[5] M. Matsushita, H. Nishizaki, H. Nishizaki, S. Nakagawa et al., "Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven web retrieval task," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 1205–1208, Geneva, Switzerland, September 2003.

[6] I. Szoke et al., "Comparison of keyword spotting approaches for informal continuous speech," in *Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (INTERSPEECH '05)*, pp. 633–636, Edinburgh, UK, 2005.

[7] D. Karakos, R. Schwartz, S. Tsakalidis et al., "Score normalization and system combination for improved keyword spotting," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '13)*, pp. 210–215, Olomouc, Czech Republic, December 2013.

[8] J. Mamou, J. Cui, X. Cui et al., "System combination and score normalization for spoken term detection," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 8272–8276, Vancouver, Canada, May 2013.

[9] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] M. Harper, "IARPA Babel Program," https://www.iarpa.gov/index.php/research-programs/babel?highlight=WyJiYWJlbCJd.

[11] J. Cui, X. Cui, B. Ramabhadran et al., "Developing speech recognition systems for corpus indexing under the IARPA Babel program," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13)*, pp. 6753–6757, IEEE, Vancouver, Canada, May 2013.

[12] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low resource languages: babel project research at CUED," in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 16–23, Petersburg, Russia, 2014.

[13] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and Bottle-Neck features in multilingual environment," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '11)*, pp. 359–364, December 2011.

[14] P. Baljekar, J. F. Lehman, and R. Singh, "Online word-spotting in continuous speech with recurrent neural networks," in *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT '14)*, pp. 536–541, South Lake Tahoe, Nev, USA, December 2014.

[15] K. Kilgour and A. Waibel, "A neural network keyword search system for telephone speech," in *Speech and Computer: 16th International Conference, SPECOM 2014, Novi Sad, Serbia, October 5–9, 2014. Proceedings*, A. Ronzhin, R. Potapova, and V. Delic, Eds., pp. 58–65, Springer, Berlin, Germany, 2014.

[16] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and Ö. Çetin, "Web resources for language modeling in conversational speech recognition," *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, article 1, 2007.

[17] A. Gandhe, L. Qin, F. Metze, A. Rudnicky, I. Lane, and M. Eck, "Using web text to improve keyword spotting in speech," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '13)*, pp. 428–433, Olomouc, Czech Republic, December 2013.

[18] G. Mendels, E. Cooper, V. Soto et al., "Improving speech recognition and keyword search for low resource languages using web data," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech '15)*, pp. 829–833, Dresden, Germany, September 2015.

[19] SpeechKit API, http://api.yandex.ru/speechkit/.

[20] K. Levin, I. Ponomareva, A. Bulusheva et al., "Automated closed captioning for Russian live broadcasting," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH '14)*, pp. 1438–1442, Singapore, September 2014.

[21] A. Karpov, I. Kipyatkova, and A. Ronzhin, "Speech recognition for east slavic languages: the case of russian," in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU '12)*, pp. 84–89, Cape Town, South Africa, 2012.

[22] A. Karpov, K. Markov, I. Kipyatkova, D. Vazhenina, and A. Ronzhin, "Large vocabulary Russian speech recognition using syntactico-statistical language modeling," *Speech Communication*, vol. 56, no. 1, pp. 213–228, 2014.

[23] M. Zulkarneev, R. Grigoryan, and N. Shamraev, "Acoustic modeling with deep belief networks for Russian Speech," in *Speech and Computer: 15th International Conference, SPECOM 2013, Pilsen, Czech Republic, September 1–5, 2013. Proceedings*, vol. 8113 of *Lecture Notes in Computer Science*, pp. 17–23, Springer, Berlin, Germany, 2013.

[24] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990.

[25] V. A. Smirnov, M. N. Gusev, and M. P. Farkhadov, "The function of linguistic processor in the system for automated analysis of unstructured speech data," *Automation and Modern Technologies*, no. 8, pp. 22–28, 2013.

[26] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[27] T. Hain, "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech Communication*, vol. 46, no. 2, pp. 171–188, 2005.

[28] D. Jouvet, D. Fohr, and I. Illina, "Evaluating grapheme-to-phoneme converters in automatic speech recognition context," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, pp. 4821–4824, IEEE, Kyoto, Japan, March 2012.

[29] L. Lu, A. Ghoshal, and St. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '13)*, pp. 374–379, Olomouc, Czech Republic, December 2013.

[30] S. G. Paulo and L. C. Oliveira, "Generation of word alternative pronunciations using weighted finite state transducers," in *Proceedings of the Interspeech 2005*, pp. 1157–1160, Lisbon, Portugal, September 2005.

[31] I. Kipyatkova, A. Karpov, V. Verkhodanova, and M. Železný, "Modeling of pronunciation, language and nonverbal units at conversational Russian speech recognition," *International Journal of Computer Science and Applications*, vol. 10, no. 1, pp. 11–30, 2013.

[32] L. V. Bondarko, A. Iivonen, L. C. W. Pols, and V. de Silva, "Common and language dependent phonetic differences between read and spontaneous speech in russian, finnish and dutch," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS '03)*, pp. 2977–2980, Barcelona, Spain, 2003.

[33] L. V. Bondarko, N. B. Volskaya, S. O. Tananaiko, and L. A. Vasilieva, "Phonetic properties of russian spontaneous speech," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS '03)*, p. 2973, Barcelona, Spain, 2003.

[34] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colo, USA, September 2002.

[35] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," *Interspeech*, vol. 2, no. 3, 2010.

[36] R. V. Bilik, V. A. Zhozhikashvili, N. V. Petukhova, and M. P. Farkhadov, "Analysis of the oral interface in the interactive servicing systems. II," *Automation and Remote Control*, vol. 70, no. 3, pp. 434–448, 2009.

[37] V. A. Zhozhikashvili, N. V. Petukhova, and M. P. Farkhadov, "Computerized queuing systems and speech technologies," *Control Sciences*, no. 2, pp. 3–7, 2006.

[38] V. A. Zhozhikashvili, R. V. Bilik, V. A. Vertlib, A. V. Zhozhikashvili, N. V. Petukhova, and M. P. Farkhadov, "Open queuing system with speech recognition," *Control Sciences*, no. 4, pp. 55–62, 2003.

[39] N. V. Petukhova, S. V. Vas'kovskii, M. P. Farkhadov, and V. A. Smirnov, "Architecture and features of speech recognition systems," *Neurocomputers: Development, Applications*, no. 12, pp. 22–30, 2013.