Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 250160, 16 pages
http://dx.doi.org/10.1155/2014/250160

Research Article

Language Recognition Using Latent Dynamic Conditional Random Field Model with Phonological Features

Sirinoot Boonsuk,[1] Atiwong Suchato,[1] Proadpran Punyabukkana,[1] Chai Wutiwiwatchai,[2] and Nattanun Thatphithakkul[2]

[1] *Department of Computer Engineering, Chulalongkorn University, Bangkok 10330, Thailand*
[2] *HLT, National Electronics and Computer Technology Center (NECTEC), Bangkok 10400, Thailand*

Correspondence should be addressed to Atiwong Suchato; atiwong.s@chula.ac.th

Received 27 September 2013; Revised 23 December 2013; Accepted 23 December 2013; Published 20 February 2014

Academic Editor: Yue Wu

Spoken language recognition (SLR) has been of increasing interest in multilingual speech recognition for identifying the languages of speech utterances. Most existing SLR approaches apply statistical modeling techniques with acoustic and phonotactic features. Among the popular approaches, the acoustic approach has become of greater interest than others because it does not require any prior language-specific knowledge. Previous research on the acoustic approach has shown less interest in applying linguistic knowledge; it was only used as supplementary features, while the current state-of-the-art system assumes independency among features. This paper proposes an SLR system based on the latent-dynamic conditional random field (LDCRF) model using phonological features (PFs). We use PFs to represent acoustic characteristics and linguistic knowledge. The LDCRF model was employed to capture the dynamics of the PFs sequences for language classification. Baseline systems were conducted to evaluate the features and methods including Gaussian mixture model (GMM) based systems using PFs, GMM using cepstral features, and the CRF model using PFs. Evaluated on the NIST LRE 2007 corpus, the proposed method showed an improvement over the baseline systems. Additionally, it showed comparable result with the acoustic system based on *i*-vector. This research demonstrates that utilizing PFs can enhance the performance.

1. Introduction

Spoken language recognition (SLR) is the task of determining the language of a spoken utterance. SLR has become an important component in many speech processing applications such as being the preprocessor of multilingual speech recognition systems and of automatic selection of the appropriate language for information service applications. Recent research works on SLR can be divided into two approaches: (1) the *acoustic approach* [1, 2] which directly models the distributions of acoustic features from speech signals; and (2) the *phonotactic approach* [3, 4] which utilizes phone-sequences tokenized from speech utterances to construct language modeling of *n*-grams of these phones. An obvious shortcoming of the phonotactic approach is that manual phonetic transcription of speech data is required for constructing language modeling. The acoustic approach has become a popular alternative to overcome this issue due to the fact that it does not require prior knowledge of a specific language and transcription of phonetic data. Furthermore, the acoustic approach captures the differences in spectral features between languages and directly models the distribution of the spectral features given in the speech utterance in each language. The acoustic system based on *i*-vector approach [5] that provided superior performance has become state-of-the-art in the language recognition field.

The performance of the overall language recognition system depends on preprocessing techniques, feature extraction, and classification techniques. Some research studies focused on feature extraction to improve the performance of SLR system. A typical acoustic-based SLR system uses the Gaussian mixture model (GMM) [6, 7] to model conventional

speech features by applying Mel-frequency cepstral coefficients (MFCC) or perceptual linear prediction (PLP). In [1], a shifted-delta cepstral coefficient (SDC) which is employed to capture longer temporal information across multiple frames has improved the performance of the acoustic-based and it is the most commonly used feature in SLR. Recent work by [8] enhanced acoustic-based SLR by using MLP for feature extraction and achieved a better result than using conventional features. However, those features modeled the behavior of the human auditory system. By relying only on the conventional speech features, the SLR system may be limited in its ability to incorporate linguistic knowledge to improve the performance. Linguistic knowledge components such as articulatory features have been proposed to improve accuracy in speech applications [9] and to increase the tolerance in noisy environment [10]. Articulatory features have played an important role in language since the properties of sound segments in each language can be described by the articulatory configuration; however, there have been few research studies using these features in SLR.

Another factor that impacts the SLR performance is the modeling technique. Traditional acoustic-based SLR systems model the spectral features by using a generative model such as GMM [6]. In the past few decades, many approaches to SLR have been developed by exploiting discriminative models in acoustic system. Whereas recent approaches have used discriminative models, such as the support vector machine (SVM) with a polynomial kernel [11], SVM with a generalized linear discriminant sequence kernel (GLDS) [12, 13], GMM-MMI (which is GMM trained with the discriminative maximum mutual information) [11, 14], and a hybrid SVM/GMM (which used a GMM supervector as a feature vector of SVM) [15–18]. The results showed significant improvement over the previously reported results. Over the last few years, the Joint Factor Analysis (JFA) [19] and *i*-vector [20] that have achieved a success in speaker verification have shown excellent performance when applied to language recognition task. Recent modeling techniques used to improve the acoustic systems based on the *i*-vector [5, 21] have provided superior improvement in language recognition. Nevertheless, discriminative modeling in previous acoustic-based SLR studies assumed independency among speech features. It needed feature space transformation for extracting a new feature vector before applying to the discriminative model. Moreover, the recent research failed to consider the sequence of feature vectors; that is, the model was estimated based on the assumption of independence between the input features.

This study utilizes phonological features (PFs) to capture the acoustic characteristics of speech signals and to integrate linguistic information into SLR to represent the language characteristics. PF attributes can represent relation between articulatory configurations and phone units in spoken languages. They can meaningfully describe the cooccurrence of articulatory gesture patterns and can describe articulatory transition better than using conventional cepstral features. We used PFs as an alternative frame-based speech features to represent language information. Typically, a speech utterance is mapped to a bundle of PF attributes. Each PF correlates with a particular articulatory attribute and it corresponds with other articulatory attributes. As each PF attribute correlates with others, we require a modeling technique that considers a sequence of these feature bundles and captures the relationship between the bundles of PF attributes. As a language classifier, this paper focuses on utilizing the latent-dynamic conditional random field (LDCRF) model which is a successful discriminative model applied to sequential data.

The purpose of this study was to employ the LDCRF model in an SLR system to capture the behaviors of PF attributes that reflect language characteristics. Baseline systems were conducted to evaluate the proposed method. The evaluation of the effectiveness of using PF attributes as speech representation was undertaken by comparing the SLR system using PF attributes with a system using spectral features such as MFCC, PLP coefficients, and SDC features. To measure the performance of classifier, we compared the SLR using the LDCRF model with that using the GMM model. Moreover, the performance of the proposed system is compared with state-of-the-art acoustic SLR system based on *i*-vector space [5].

This paper is organized as follows. Section 2 presents the literature review while Section 3 describes the background of PF. Section 4 presents the LDCRF model. Section 5 outlines the proposed SLR system. In Section 6 the experimental setup including the speech corpus, configuration of the feature extraction, and classification is described and the experimental results and discussion are presented in Section 7. Finally, the conclusion and suggestions for further work are presented in Section 8.

2. Related Works

PF research has focused on two areas: (1) PF extraction which estimates the underlying phonology in speech signals and (2) integrating PFs to improve their performance in SLR.

In general, extracting PFs can be achieved by manually mapping the phone transcriptions to PF attributes. The disadvantage is that it lacks the flexibility to integrate a new language. To avoid manual mapping, many different techniques have been developed, such as using multilayer perceptron artificial neural networks (MLP ANNs) [22], time-delay recurrent neural networks (RNN) [23], the hidden Markov model (HMM) [24], and the GMM [25] and SVM [26–28] techniques. Due to the popularity and success of MLP used in speech recognition, we make use of an MLP model to detect PF attributes in this paper.

The literature on the integration of PFs into an SLR system shows that few studies have employed PFs in an acoustic-based SLR system. In [26], the combination of SDC features and distinctive features (which are similar to PF attributes) is used as an input for acoustic GMM system and it showed better results than only using SDC features.

Another line of research in modeling techniques suggested utilizing a statistical model that considered the dependence of the assumptions between the input features. Conditional random field (CRF), a discriminative model, is proposed to solve sequential labeling (i.e., a kind of sequential data problem). The label sequence is calculated over the entire

sequences from a log-linear combination of input features. It showed a significant improvement in speech recognition [29] such as phone classification [30] and phone recognition [31]. In [32], the integration of PFs into the phone recognition task by using CRF achieved superior performance compared with conventional features. Furthermore, the deep structure CRF [33] also yielded a better result than other discriminative models in SLR. Although CRF can capture the extrinsic dynamics between the behaviors of features, it cannot capture the intrinsic dynamics of feature sequences.

LDCRF, a variant of CRF, has been successfully applied to continuous gesture recognition tasks [34–36] and it outperformed CRF. It was designed to learn the substructure sequential label. The advantage of using LDCRF is that it captures the intrinsic dynamics of the sequence of the features and it also explicitly learns the substructure of the features as well as the extrinsic dynamics between the class labels. In this paper, LDCRF is used to capture the dynamic characteristic of articulatory configurations within each phone and across phone sequences to represent a model of language.

This paper proposes an acoustic-based SLR system using LDCRF with PF attributes. The contribution of this paper is to incorporate linguistic information and acoustic information to improve the performance of the SLR system. In addition, employing LDCRF to learn the dynamic sequences of PF attributes for modeling the language can resolve the problem of the independence of observations. Furthermore, this work provides analysis on the language discriminative ability of PFs compared with conventional cepstral features.

3. Phonological Features

In linguistics, phonological features (PFs) represent speech sounds as bundles of positive-valued (+) or negative-valued (−) features where positive value shows a presence of the feature while otherwise, the value is negative. The phonological component, mapping speech production characteristics to phones, is considered to be a linear sequence of these feature bundles. Many studies proposed different PF concepts [23] such as the Sound Pattern of English (SPE), the Government Phonology (GP), the multi-valued (MV) features, and the hybrid features (HF). These concepts are defined from different articulatory, acoustical or phonological aspects and the different concepts of relationships between those aspects. In this paper, we use the Sound Pattern of English (SPE) definition, a widely used definition for describing phone inventory, since it has no redundant mapping rules. SPE, defined by Chomsky and Halle [37], illustrates speech production as binary values. According to the SPE definition, each phone can be broken down into 14 PF attributes: vocalic, consonantal, high, low, back, round, tense, continuant, anterior, coronal, voice, nasal, strident, and silence as shown in Table 1. They are classified as (1) the major class features which describe the obstruction in the way of airstream: vocalic, consonantal, and nasal; (2) the manner of articulation features which describe the primary constriction of air flow: continuant; (3) the place of articulation features which describe the body of tongue: coronal, anterior, and round; (4) the source features: voice, strident, and tense; (5) vowel features which describe the position of tongue: back, high, and low; and (6) other features: silence.

Table 1: Example of phonological features in the Thai word "khun."

IPA	[k^h]	[u]	[n]
Vocalic	−	+	−
Consonantal	+	−	+
High	+	+	−
Back	+	+	−
Low	−	−	−
Anterior	−	−	+
Coronal	−	−	−
Round	−	+	−
Tense	−	+	−
Voice	−	+	+
Continuant	−	+	−
Nasal	−	−	+
Strident	−	−	−
Silence	−	−	−

Table 1 shows an example of the word "khun" in Thai (defined as [k^hun] by IPA) which means "you." Each column represented a sequence of phone segments, and each segment is characterized by a set of PF attributes.

Moreover, the patterns of PF sequences occur differently for each language. For instance, some diphthongs occur in some languages, but they are not allowed in others. The diphthong /aw/, representing an articulatory configuration running from the vowel [a] to the glide [w], occurs in Thai but not in Russian. In Japanese, the onset of a velar nasal /ŋ/ is allowed but this does not occur in English. This causes the absence of a movement pattern of articulatory configuration from the velar nasal to vowel in English. With the benefit of different patterns of PF sequences in different languages, we can utilize the sequential PF attributes to discriminate between languages.

4. LDCRF for Language Classification

The latent-dynamic conditional random field (LDCRF) model is discriminative and relaxes the conditional independence assumptions between input features. It was designed to identify the substructure sequence label; thus, it captures the intrinsic characteristics within a class and interclass of patterns by associating a set of hidden states with each class label. To apply this in the SLR task, we used these hidden states to model the internal substructures of different language patterns and provide the overall likelihood for classification. Each hidden state can be treated similarly to a CRF. The overall likelihood is the sum of individual likelihoods from the hidden states. In a language recognition problem, we assume a training set of N speech utterances given as $X = \{x^1, x^2, \ldots, x^N\}$ which contains speech utterances from the class label of languages. The corresponding label can be denoted as $Y = \{y^1, y^2, \ldots, y^N\}$ where each y^n is a member

of a set of possible speech labels. Given the above definitions, a latent conditional model is defined as

$$P(Y \mid X, \theta) = \sum_{h} p(Y \mid h, X, \theta)\, p(h \mid X, \theta), \quad (1)$$

where x is the concatenation of all feature vectors x_i for the entire sequence of the utterance and $\theta\{x_i\}$ are the parameters of the training model.

4.1. Problem Formulation. SLR problems can be represented in mathematical formulation as $L = \{l_1, l_2, \ldots, l_k\}$, where L denotes a set of n different languages. Given that X denotes the input speech signal, the most likely recognized language, L^*, can be represented as

$$L^* = \arg\max_{L} P(L \mid X). \quad (2)$$

Since the problem of SLR is a multiclass classification problem, this study broke the problem down into multiple binary classifications by using a one-versus-one schema. In language classification, we focus on the presence or absence of languages; thus, the class label from each LDCRF model is limited to the binary value $y^n \in \{0, 1\}$. For example, the class label $y^n \in \{Chinese, Arabic\}$ represents the pair language classification between *Chinese* and *Arabic*.

The task of the LDCRF model is to learn the mapping between the sequence of observation inputs $X = \{x_1, x_2, ..., x_T\}$ and the sequence of class labels $Y = \{y_1, y_2, \ldots, y_T\}$. Each y_i is a class label for the ith frame of speech sequence and is a member of a set L of possible class labels. We assign $y_1 = y_2 = y_T$ because one training speech utterance has one language. Each frame of observation is represented by a feature vector x_i. The LDCRF model incorporates a vector of hidden state variables $h = \{h_1, h_2, \ldots, h_n\}$ to model the substructure of the speech sequence. Each h_i is a member of the set of all possible hidden states H and the hidden variables are not observable in the training examples.

The model is limited to having disjointed sets of hidden states corresponding to each class label. Given each $h_i \in H_{y_i}$ where H_{y_i} is the possible hidden states for the class label y_i and H is the set of all possible hidden states which is the union of all sets H_{y_i}, then a sequence has $P(Y \mid h, X, \theta) = 0$ for any $h_i \notin H_{y_i}$, or otherwise 1. The model is shown as

$$P(Y \mid X, \theta) = \sum_{h:\forall h_i \in H_{y_i}} p(h \mid X, \theta), \quad (3)$$

where $p(h \mid X, \theta)$ is the conditional random field. It is defined as

$$P(h \mid X, \theta) = \frac{1}{Z(X, \theta)} \exp\left(\sum_{k} \theta_k \cdot F_k(h, X)\right), \quad (4)$$

where the partition function Z is a normalization value with respect to all candidate paths for the input sequence. It can be written as

$$Z(X, \theta) = \sum_{h} \exp\left(\sum_{k} \theta_k \cdot F_k(h, X)\right), \quad (5)$$

and the feature vector F_k is defined as

$$F_k(h, X) = \sum_{i=1}^{T} f_k(h_{i-1}, h_i, X, i). \quad (6)$$

The $F_k(h, X)$ vector is the sum over all feature functions. In (6), each feature function $f_k(h_{i-1}, h_i, X, i)$ can be represented by two kinds of feature functions: a state function $s_k(h_i, X, i)$ or a transition function $t_k(h_{i-1}, h_i, X, i)$. The state function s_k depends on a single hidden variable in the model while the transition function t_k can depend on pairs of hidden variables.

4.2. Training LDCRF. Given a training set consisting of n labeled sequences (X_i, Y_i) for $i = 1, \ldots, n$, the objective function for training is defined similarly to the reported work on CRF [38, 39] as

$$L(\theta) = \sum_{i=1}^{n} \log p(Y_i \mid X_i, \theta) - \frac{1}{2\sigma^2}\|\theta\|^2, \quad (7)$$

where the log-likelihood of the first term is given by the conditional log-likelihood of each training sequence and the second term is the Gaussian prior likelihood with the variance σ^2 (i.e., $P(\theta) \sim \exp(1/2\sigma^2\|\theta\|^2)$). The gradient ascent technique was used to optimize the parameter values $\theta^* = \arg\max_\theta L(\theta)$.

4.3. Feature Functions. The state functions are considered by a subset of observations preceding the current variable because the dependency of the entire observation sequence is ignored. It can be said that $s_k(h_i, X, i) = s_k(h_i, \widetilde{X}_i, i)$ where $\widetilde{X}_i = [X_{i-m}, \ldots, X_i]$ is a window of size $m + 1$ preceding the variable at the ith frame. The state function $s_k(h_i, X, i)$ represents the value of hidden states with a neighborhood relationship with the entry in $\widetilde{X}_i$. The details of state function are described in Section 5.2.

4.4. Inference on LDCRF. To classify a test sequence X, we find the most likely label sequence Y^* that maximizes the conditional model. In the inference phase, we use the optimal model parameter θ^* obtained from training data to estimate the model output as

$$Y^* = \arg\max_{Y} p(Y_i \mid X_i, \theta^*). \quad (8)$$

Because the LDCRF model is incorporating hidden states, it cannot be produced directly by a searching path like CRF. Thus, the most probable sequence can be estimated from summing the probabilities of the hidden paths, where each class label is associated with the sets of hidden states. It can be written as

$$Y^* = \arg\max_{Y} \sum_{h:\forall h_i \in H_{y_i}} p(h \mid X, \theta^*). \quad (9)$$

Applying the LDCRF model in SLR, the likelihood for estimating a specific language l is equal to the marginal

probability $P(Y_i = l \mid X, \theta^*)$. The belief propagation algorithm is employed to estimate the marginal probabilities. This probability is equal to the summation of the marginal probabilities of the hidden states of the subset H_l:

$$p(Y_i = l \mid X, \theta^*) = \sum_{h:\forall h_i \in H_l} p(h \mid X, \theta^*), \quad (10)$$

where X is the concatenation of the feature vector x_i for the entire sequence of speech. The observation X can be represented by speech features (see Section 5.1 for details). The model parameter θ^* is learned during the training phase.

5. Proposed SLR System

The block diagram of the proposed SLR system is shown in Figure 1. It consists of two main components: feature extraction and language classification. In feature extraction, the input speech is first converted into PF attributes by MLP attribute detectors. Then, the feature vectors are used as the input to language classifiers. Additionally, the feature vectors can be PF attributes or PF attributes applied with the shift delta operation. The language classifiers, which are trained for each target language, are used to determine the most likely language.

5.1. Feature Extraction. The main goal of the feature extraction is to extract discriminative speech representations that highlight the relevant information of language. A continuous speech is converted to a sequence of feature vectors containing information of language characteristics. Then, the feature vectors are used as the input to the language classification component.

5.1.1. Phonological Features (PFs). The PFs are the linguistic knowledge that conveys language information. They contain articulatory characteristics and different patterns of articulatory configuration movements. For PF attributes extraction, we employed MLP to detect the PF attributes from speech utterance. The MLP attribute detectors were applied to input speech and classify the values of the attribute for each frame.

The MLP attribute detectors were trained on the TIMIT speech corpus. The TIMIT database is a corpus recorded readings of a large set of English sentences. It contains 6,300 sentences, 10 sentences spoken by each of 630 speakers, recorded from male and female speakers of eight dialects of American English. The TIMIT corpus has been divided into a training set (4,620 utterances from 462 speakers) and a test set (1,680 utterances from 168 speakers). The training set was used for training each MLP attribute detector and each MLP attribute detector was evaluated on the test set.

For training MLP attribute detectors, we generated labeling of PF attribute from TIMIT phonetic transcription. We used the set of 13 PF attributes (excluding silence) that followed the SPE definition shown in Section 3. The mapping between phone transcription and PF values is based on [23]. The MLPs used PF transcription to separately train the attribute detector. Speech parameters were 39-coefficient MFCC vector (12 Mel-frequency cepstral coefficients and energy plus their delta and acceleration coefficients) with 25 ms window length and 10 ms frame shift. Thirteen MLP attribute detectors (one for each PF attribute) were trained by using the NICO toolkit [40]. Each MLP consisted of three layers and the input layers had 39 nodes. The optimal number of hidden units was determined through MLP that provided the best performance in tuning experiments. Thus, each MLP was trained with the various numbers of hidden units. From the tuning results shown in Table 2, the MLP with 250 hidden units provided the best performance. The output layer of MLP was an estimation of the posterior probability of the PF attributes. Each MLP has two output nodes representing binary values of PF attribute. The results of thirteen MLP attribute detectors were composed to a 26-dimensional feature vector. Then, the feature vectors were used as the input of language classifiers.

Table 2: The tuning result of MLP attribute detectors with the number of hidden units.

Number of hidden units	Accuracy
100	82.2
150	85.6
200	87.6
250	87.9
300	87.9

Table 3 shows the performance of the MLP attribute detectors evaluated on the test set. The 39-MFCCs feature vectors were used to test classification performance of MLP attribute detectors. The results show that the MLP attribute detector using the nasal attribute outperformed other attributes (silence was not considered). The average accuracy of overall MLP attribute detectors was 87.95% and the range of classification accuracy at the frame level for each attribute detector had value between 81% and 98%.

Figure 2 represents the spectrogram and PF attributes of Thai utterance "พึ่งโทรมาครั้งสอง" which can be transcribed as /pvng^ toz^ maz^ kraang^ thiiz^ s@@ng^/, along with the canonical values of PFs. The top subplot shows the spectrogram of the speech utterance. The values of PF attributes, as derived from MLP attribute detectors, are presented on the 2nd to 8th subplots. The outputs are the continuous values between 0 and 1, where 1 represents the most likely classifying PF attribute. Although each PF attribute was derived from separately-trained attribute detectors, the result shows that the values of each PF correlate with others. That is, the PFs did not perform frame synchronous as the manual mapping did while they change in similar pattern. For example, the vowel features that describe the position of tongue: *Back* and *High* attributes (on the 7th and 8th subplot) and the *Round* attribute, the feature describing the body of tongue and involving vowel (on the 6th subplot), simultaneously change in the same manner. For another example, the relationship between three features *Vocalic*, *Consonantal*, and *Continuant* attributes (on the 2nd, 3rd, and 5th subplot, resp.) is correlated. The rising of *Vocalic* and

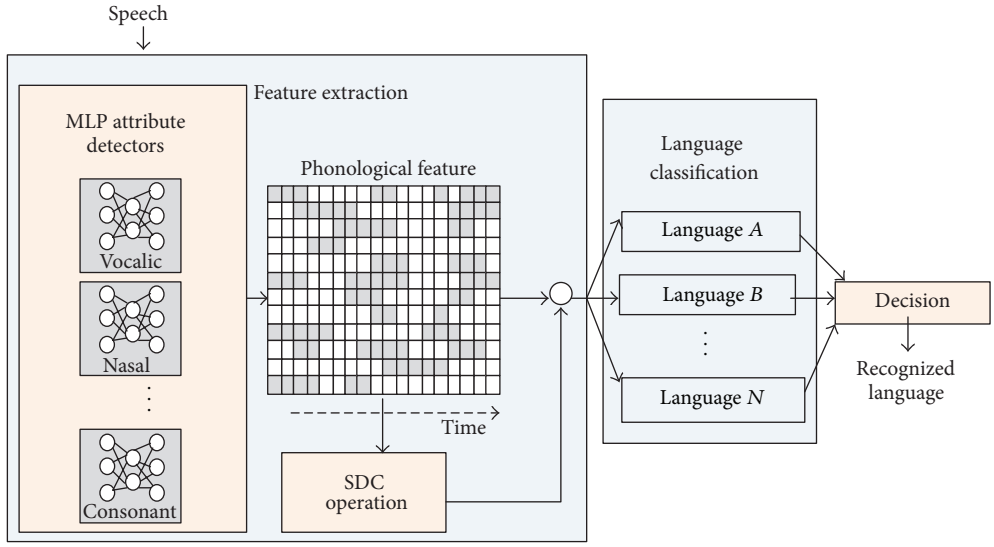

Figure 1: Proposed SLR system.

Table 3: Percentage accuracy of MLP attribute detectors tested on TIMIT corpus.

Attribute	Accuracy
High	81.9
Back	82.6
Coronal	82.8
Vocalic	82.8
Anterior	83.3
Tense	84.3
Consonantal	84.8
Continuant	88.7
Low	88.8
Voice	90.3
Round	91.7
Strident	95.0
Nasal	**95.8**
Silence	98.5
Average	87.95

The high accuracy result is displayed in bold.

Continuant feature simultaneously occurred with the falling of *Consonantal* feature.

In addition, the feature values do not switch instantaneously and all features do not always change simultaneously at the phone boundaries. It means that each feature makes the smooth transition between the side of target-like and non-target-like values. With the manner of using detector, it helps relaxing the limitation of value changing at phone boundaries. It has low possibility for the insertion error to occur. We can make use of the loosening and the properly asynchronous manner of the features to use these outputs as features rather than the output derived from the manually mapping procedure.

5.1.2. Shifted Delta Coefficient Operation. In this paper, the shifted delta operation was applied to compute delta PF attributes across successive frames, denoted as SDPF, since shifted delta coefficients (SDCs) have been successfully used in SLR [41, 42] to capture the cepstral dynamics of a long temporal window. We used the shifted delta operation to estimate the changes of PF attributes in multiple frames and to capture the language characteristic resulting from the pattern of the changes. The SDC features are obtained by stacking delta cepstral coefficients across multiple speech frames. The computation of the shifted delta operation is shown in Figure 3. There are four parameters, N-d-P-k, which are used in computing a shifted delta operation. The parameter N is the total dimension of the coefficients in each time frame. The parameter d is the time advance and delay for the delta computation. The parameter k is the number of blocks whose delta coefficients are concatenated to form the final feature vector. The parameter P is the time shift between consecutive blocks. In this study, the four parameters in a shifted delta operation were set to 7-1-3-7 following the configuration which has been successfully used in SLR [21]. The SDC feature vector at frame time t is given by the concatenation of all the blocks of delta vectors, $\Delta c(t + iP)$, where

$$\Delta c\,(t + iP) = c\,(t + iP + d) - c\,(t + iP - d)\,. \tag{11}$$

5.2. Language Classification. In the language classification, the classifiers employed the feature vectors as input and provided a decision score of the hypothesis language as output. We used a binary classifier scheme to construct language classifiers. Each language classifier is performed using speech features extracted from Section 5.1. For recognition, the language of speech utterance is determined from the result of sequence classification of language classifiers. This result is obtained from the summation of sequence of classification

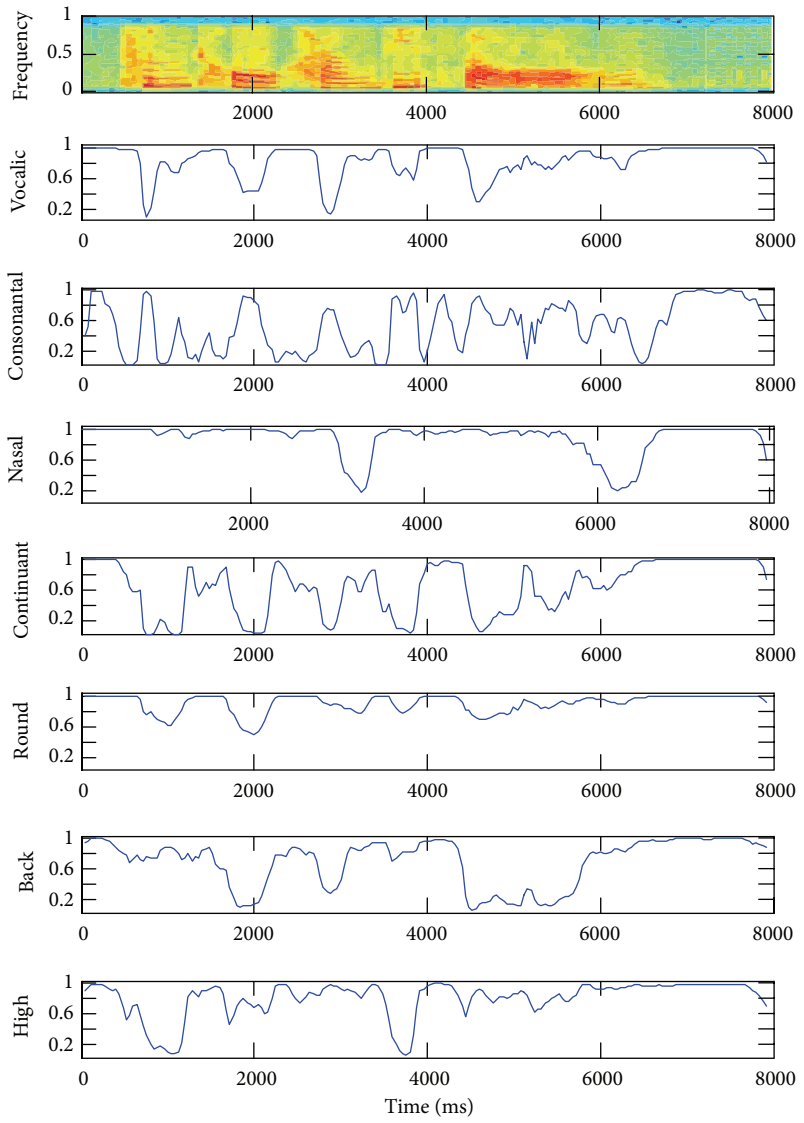

Figure 2: Spectrogram and PF attributes of the sentence "พึ่งโทรมาครั้งสอง", resulting from MLP attribute detectors.

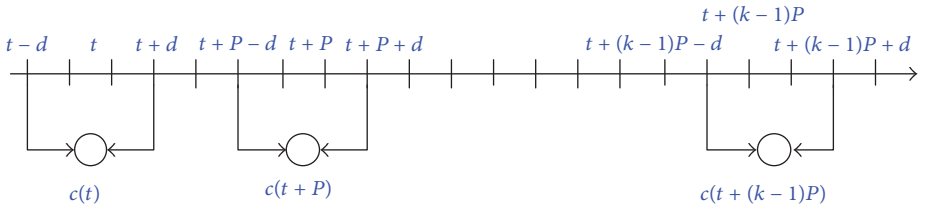

Figure 3: Computation of SDC feature vector at frame t for the parameters N-d-P-K.

scores. The performance of the classifiers is evaluated by considering the correctness of the sequence classification.

In this paper, we investigated the use of the outputs of the MLP attribute detector to construct feature functions for the LDCRF model. We used the posterior probability values obtained by the MLP attribute detectors to determine their frame-level result as state feature functions of the LDCRF model.

The feature functions for PF attributes are defined using the language label/PF attribute pairs. Let state feature function $s_{\phi,\text{Att}}(h_i, X, i)$ depend on a single hidden variable to link the output label of language ϕ. The feature function that ties together the output label of language ϕ is based on the output of the MLP attribute detector for feature Att. This allows for the association between a language ϕ and the Att attribute even if the Att attribute is not traditionally associated with

language ϕ. Thus, the state feature function is defined as follows:

$$s_{\phi,\text{Att}}(h_i, X, i) = \begin{cases} \text{MLP}_{\text{Att}}(x_i), & \text{if } h_i = \phi \\ 0; & \text{otherwise,} \end{cases} \quad (12)$$

where the output of the MLP detector for feature Att at time i is $\text{MLP}_{\text{Att}}(x_i)$. For example, the state feature function is defined for the output label language ϕ where the MLP output of feature *vocalic* can be defined as

$$s_{\phi,\text{vocalic}}(h_i, X, i) = \begin{cases} \text{MLP}_{\text{vocalic}}(x_i), & \text{if } h_i = \phi \\ 0; & \text{otherwise,} \end{cases} \quad (13)$$

where $\text{MLP}_{\text{vocalic}}(x_i)$ is the output of the MLP detector for the vocalic attributes at frame i of speech sequence X. As for the language classifier model above, state feature functions are defined for all possible language label/phonological attribute pairings, not just for the canonical attributes for the label. The $|H| \times |H|$ transitions functions t_k are defined as one for each hidden state pair (h', h''). Each transition function is expressed as

$$t_k(h_{i-1}, h_i, X, i) = \begin{cases} 1, & \text{if } h_{i-1} = h',\ h_i = h'' \\ 0; & \text{otherwise.} \end{cases} \quad (14)$$

Weights associated with a transition function for hidden states that are in the subset H_y will model the substructure patterns.

Furthermore, we also used the PF attributes applied with shifted delta operation and other speech attributes from the feature extraction as features in the LDCRF model. The definition of feature functions or the state function of the LDCRF model is similar to the above described function; the only difference is we used the dimension of the feature vector of the speech attribute instead of the output from the MLP attribute detector to define the feature function.

For LDCRF model training, the speech utterance was labeled on the frame level either as a target language or as a nontarget language. The configuration parameters of the LDCRF model in the training and validation phase were set as follows: the hidden states were set to 3 and the window size was set to 3, while the regularization term used in BFGS was set to 300. It was observed that 3 hidden states per label gave better results. The LDCRF model was trained using the objective function described in Section 4. For performance evaluation, the LDCRF model was compared with two classification models: the CRF and the GMM models.

6. Experimental Method and Data

6.1. Speech Corpus. In the language classification and recognition tasks, the speech corpus from a portion of the NIST 2007 language recognition evaluation (LRE) [43] development data set was used as a speech corpus for training and testing. The corpus contains 7,530 utterances in eight languages: Arabic (Arb), Bengali (Ben), Thai (Tha), Urdu (Urd), Russian (Rus), Chinese-Cantonese or Yue Chinese (CH.can), Chinese-Min (CH.min), and Chinese-Wu (CH.wuu). The language recognition task experiment focused on closed-set recognition where target languages must be included in the test set.

6.2. Language Classifiers. The speech features were extracted from speech utterances into a sequence of feature vectors and they were used to train the language classifiers. In this experiment, the classification problem of eight target languages was broken down into a pairwise language classification. Thus, the total number of classifiers was an elementary combination $C(n, k) = C(8, 2) = 28$.

6.3. Performance Evaluation. Performance evaluation of the proposed system by k-fold cross validation ($k = 5$) was used to reduce the bias of the trained model and to generalize classification ability of the model. Under fivefold cross-validation, the speech dataset was randomly partitioned to form five disjoint subsets. Four sets were used for training and the remaining set was test set. The average accuracy and average error rate were calculated by repeating training and testing five times on different combinations of data sets.

Receiver operating characteristic (ROC) curve is used as another evaluation method of the performance of a language pair classifier. It used the maximal marginal probabilities of (10), which was computed from the 5-fold models. ROC shows the relation of the true positive rate, which is computed as the ratio of the number of recognized frames and the total number of ground truth frames, and the false positive rate, which is computed as the ratio of the number of falsely recognized frames and the total number frames of the nontarget class.

Additionally, the performance of language classifiers can be statistically measured by applying a paired t-test at the significance level 0.01 to consider whether or not the classifiers using different speech features are significantly different.

6.4. Preliminary Studies

6.4.1. Preliminary Studies on the Discriminative Ability of Speech Features Using MANOVA. Based on the hypothesis that the distributions of PF attributes of each language are different and they have the capability to discriminate between languages, this experiment was conducted to study the variance of the distribution of speech features occurring in different languages.

MANOVA (multivariate analysis of variance) was used to analyze the difference of the means of features among all the languages. It is an extension of the F-test which analyzes the distribution of features with more than one dependent variable. We used MANOVA for observing the means of the values of the speech attributes and for analyzing the distribution of the speech features across different languages.

The discriminating powers of features were evaluated using two different criteria, the F-ratio and the P value. The feature is more significant if the F-value is very large and the P-value is small. A level of significance of 0.01 was used as the criterion for checking the statistical significance

Table 4: MANOVA of language classification using MFCC, PLP, and PF attributes.

Attribute	P value	F-ratio
PF	0	1425.43
MFCC	0	22.63
PLP	1	0.11

of the P value. In this study, we compared three speech features: PF, MFCC. and PLP, which were obtained from eight languages.

In Table 4, the P value and the F-ratio of three speech features are listed in descending order of F-value. The P value of PF attributes which is less than 0.01 indicates that the distribution of PF attributes for at least one of the eight languages was different. The PF attributes showed the largest F-ratio; thus, they contributed the most to discriminating between languages. The MFCC with a smaller F-ratio had less discriminative ability than the PFs. Based on the discrimination factors of PLP where the P value was 1, the mean values of PLP were not significantly different among eight languages.

6.4.2. Preliminary Studies on the Variance of the Distribution of PF Attributes Using ANOVA. The purpose of this experiment was to study the variance of the distribution of each PF attribute occurring in different languages. To analyze the distributions of each PF occurring in each language, the PF attributes were extracted by using MLP attribute detectors corresponding to the phonetic features.

The variance of the means of the PF attribute distributions were compared among the eight languages. ANOVA (analysis of variance) was used to analyze the difference of the means of the PF attributes among all the languages and in pairwise analysis. F-ratios and P values were used to evaluate the differences. A level of significance of 0.01 was used as the criterion for statistical significance of the F-ratios for the individual ANOVA. The number of language pairwise combinations from the eight languages was 28 language pairs. Thus, there were 364 cases used to analyze the distribution of 13 PFs.

Table 5 shows the results of the overall ANOVA analysis of PF attributes in terms of the F-ratio and P value. The P values of all PF attributes were less than 0.01 indicating that the distribution of the PF attributes for at least one of the eight languages is different. It showed that there were significant differences among the eight languages. The results of the F-ratio analysis show that (i) the low attribute, a member of the vowel features, has the best discriminative ability; (ii) the features involving obstruction in the vocal tract (*Vocalic* and *Continuant* attributes) show better discriminative ability than the *Consonantal* attribute; (iii) among the vowel features of the *High*, *Back*, and *Low* attributes, the *Low* attribute is more discriminative than the others; and (iv) three attributes (including *Consonantal*, *Nasal*, and *Round*) showed poorer discriminative ability.

For the language pairwise analysis, there were 296 cases out of 364 (about 81%) where the pairwise P values were less than 0.01. The results show that the mean values of the PF attributes were statistically different between two languages. From the results of the pairwise analysis, the discriminating capability of each PF attribute was different. Thus, incorporating all PF attributes can provide supplementary information that reflects the language characteristics and can compensate for the discriminating ability of a weak PF attribute.

Table 5: F-ratio and P value statistics for each PF attribute from pairwise analysis.

PF classes	F-ratio	P
Vocalic	3706.52	0
Consonantal	830.29	0
Nasal	970.02	0
Continuant	4984.65	0
Coronal	3414.93	0
Anterior	4919.50	0
Round	748.32	0
Voice	2683.94	0
Strident	2695.36	0
Tense	5237.42	0
Back	3041.22	0
High	3214.64	0
Low	6597.28	0

From analyzing *Consonantal* attribute, there were 6 out of 28 pairs where the pairwise P values were greater than 0.01 which means that there are no differences between the two languages. They were CH.min versus Arb, CH.min versus Ben, CH.min versus Tha, CH.min versus CH.can, Rus versus Arb, and Rus versus CH.wuu. It can be noticed that the *Consonantal* attribute of CH.min had less ability to discriminate between languages. From observing the *Nasal* attribute on language pairwise analysis, about 15% of all pairwise cases were having P values that were greater than 0.01 (such as Arb versus Rus, Arb versus Tha, Arb versus Urd, and Ben versus CH.can). The reason that the mean values of the PF attributes were not statistically different could be due to the occurrence of nasal vowel and nasal consonant. The nasal vowel, which is vowel that is adjacent to nasal consonants, is allowed in some languages such as Bengali, Urdu, CH.min, and CH.wuu. Thus, nasal vowel can be used to discriminate between languages. Some nasal consonants such as velar nasal [ŋ] (which represents the sound /ng/) is allowed in some languages (e.g., CH.min, CH.can, CH.wuu, Urdu, Thai, and Russian) but it is not allowed in others. From observing *Round* attribute, there were three cases, including CH.min versus CH.can, CH.min versus CH.wuu, and CH.min versus Thai, where pairwise P values were larger than 0.01. The mean values of *Round* attribute among Chinese languages are not significantly different. It is interesting to describe why the *Round* attribute is not good to distinguish among these languages. Since *Round* attribute is one of vowel characteristics, we used the vowel chart to illustrate the manner of occurrence of this attribute. Vowels occurred in each language are summarized in the vowel chart and used to analyze the possibility of vowel feature in each language. From vowel chart analyzing as shown in Figure 4, we found

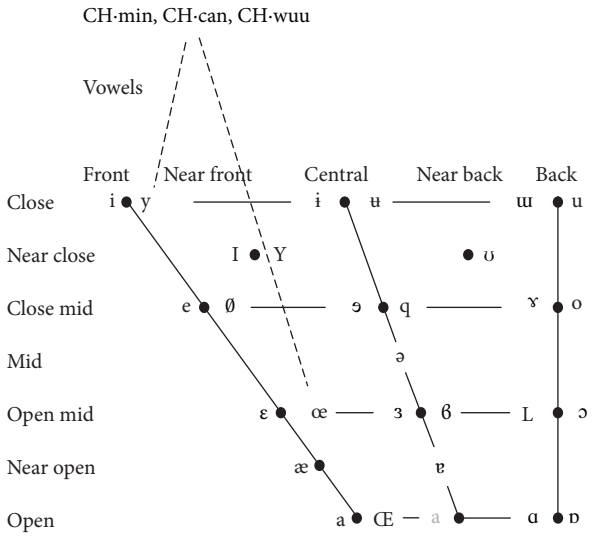

Figure 4: IPA vowel chart.

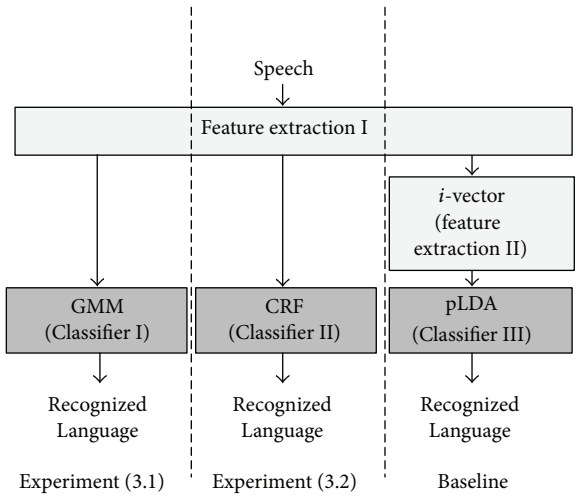

Figure 5: The baseline systems.

that that some rounded vowels are commonly present in Chinese languages such as the close front rounded vowel (represented as [y]) and the open-mid front rounded vowel (represented as [œ]). These examples of rounded vowels occur in Chinese languages (CH.min, CH.wuu, and CH.can) and are not present in others. The benefit of *Round* attribute can be used to distinguish Chinese languages from others but it cannot separate among Chinese dialects. Additionally, the mean value of *Round* attribute was not statistically different between CH.min and Thai because the rounded vowels can be often found in Thai and CH.min.

6.5. Baseline Systems and Experimental Evaluation. The architecture of baseline system and experimental evaluation is shown in Figure 5. It consists of two feature extraction methods and three classifiers. There are one baseline system

Table 6: Summary of speech features used in the experiment.

Feature	Description	Dimension
PF	Phonological features	26
SDPF	SDC over PF	182
PLP	C0 + 7 PLPs	8
MFCC	(12 MFCCs + E) + Δ + ΔΔ	39
SDCPLP	C0 + 7 PLPs + shifted delta cepstral	49
SDC + PLP	Concatenation of SDCPLP and 7 PLPs	56

and two experiments (as described in Section 6.6) which utilize different feature extractions and classifiers. The speech utterance is decoded by different feature extractions; then the feature vectors are fed into classifiers to determine the language of speech utterance. We evaluated the discriminative LDCRF model against the GMM to compare ability with generative model and we also evaluated it against the CRF model to examine the ability of discriminative model. Additionally, the performance of proposed system was also compared to the acoustic system based on *i*-vector space.

For comparing performance of proposed feature, the LDCRF language classifiers using different speech features were compared with conventional speech features in the acoustic-based SLR system. Summary of speech features used and compared in the experiment is provided in Table 6. There are six types of speech features: PF, SDPF, PLP, MFCC, SDC, and SDC + PLP. The details of the speech features are described in the following section.

6.5.1. Cepstral Feature. Typically, speech utterance is represented by cepstral feature vectors (including MFCC and PLP) which have been demonstrated to perform well in speech applications. This paper used these features to compare the discriminative ability for language classification. For MFCC and PLP feature extraction, the frames of speech utterance were analyzed using a 25 ms window with a 10 ms overlap. The MFCC feature vectors with 13 dimensions, including 12 MFCCs and energy, along with their first and second temporal derivatives, were calculated for each frame. For the PLP feature vector, we used the 7th order of PLP cepstral coefficients.

6.5.2. Shifted Delta Coefficients of Cepstral Features. The typical SDC that has been successfully and widely used in acoustic-based SLR systems is computed from the changes of the cepstral features across multiple frames. In this study, we used it to compare the discriminating ability with the proposed features.

In this experiment, we focused on two features which were computed from the SDC operation: SDCPLP (which is a result from applying the SDC operation to PLP features) and SDC + PLP (which is the concatenation of the PLP features and their SDC coefficients), as shown in Table 6.

6.5.3. Baseline Systems (i-Vector). Among the most popular modeling techniques used in acoustic systems, the *i*-vector

which is usually applied to model spectral features is the state-of-the-art in language recognition. To evaluate the effectiveness of the proposed system, we compared the performance between the proposed SLR system and the acoustic system based on i-vector.

The i-vector approach was motivated by the success of JFA, which models language and channel variability separately. In contrast with JFA, the i-vector concept utilizes the total variability subspace to model all variability in the same low dimensional subspace. That is, the total variability space contains the language and channel variability. The idea of total variability subspace is that the adapting of the Universal Background Model (UBM) to eigenvoice where the UBM is trained from all languages used in this experiment. The language-dependent and channel-dependent supervector M is a concatenation of all mean vectors of adapted GMM component. The GMM supervector M is obtained as (15) which is defined by the matrix T:

$$M = m + Tw, \tag{15}$$

where m is the UBM supervector (i.e., a language- and channel-independent component), T is a low rank rectangular matrix called the total variability, and w is the i-vector, which is random vector with a normal distribution $N(0, 1)$. The i-vector w is obtained for a given speech utterance.

To recognize language, the i-vector space, a log-likelihood ratio, is used to obtain a similarity score between a testing i-vector and the i-vectors of training class. There are many scoring methods for computing the similarity between two i-vectors w_1 and w_2. The most successful model for modeling i-vector in speaker recognition is the generative Probabilistic Linear Discriminant Analysis (pLDA) [44]; thus, the pLDA log-likelihood ratio is used to compute the distance score between two i-vectors in this paper. It is computed as follows:

$$\begin{aligned}\text{score}(w_1, w_2) = {} & \log N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{tot}} & B \\ B & \Sigma_{\text{tot}} \end{bmatrix}\right) \\ & - \log N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{tot}} & 0 \\ 0 & \Sigma_{\text{tot}} \end{bmatrix}\right),\end{aligned} \tag{16}$$

where w_1 and w_2 are the two i-vectors, $N(\cdot)$ is the normal Gaussian probability density function, and Σ_{tot} is the total covariance matrix of training i-vectors and B denotes the between-class covariance matrix of training i-vectors. The i-vectors within each class (i.e., language) are averaged and represented as one i-vector.

The feature extraction used in this experiment was similar to that employed in [5]. The 7 MFCCs extracted and the SDC with 7-1-3-7 configuration were obtained. The final 56-dimensional feature vectors, a concatenation of MFCCs and its SDC, were converted to an i-vector based on GMM with 2048-components by using ALIZE [45]. The components involving i-vector of training class such as the UBM, the T matrix, and total and between-class matrices were trained using the training set (as 5-fold in Section 6.3).

6.6. Experimental Setup. There were four experiments described as follows.

Experiment (1). This experiment was a comparison of the discriminative ability of PF attributes with other features. The purpose of this experiment was to compare the performance of language classifiers using proposed PF features and using conventional features. The 28 language classifiers were trained from different speech features and the capabilities of discrimination were evaluated for each feature.

Experiment (2). This experiment was a study of the performance of language classifiers when reducing the effect of confused target classes. The purpose of this experiment was to evaluate the classifier performance when the ambiguous languages, that is, Chinese-Min and Chinese-Wu, were removed. That is, we simplified the classification problem in language recognition by ignoring the varieties of Chinese language. We compared the performance of the language classifiers of the five target languages with the results from Experiment (1). Thus, the number of classifiers was represented as an elementary combination by $C(6, 2) = 15$.

Experiment (3). The purpose of this experiment was to demonstrate the superiority of the discriminative model by using the LDCRF model for the language recognition task. The PF attributes were used as the speech features for training and testing language classifiers in this experiment. In addition, the PLP features were used to construct classifiers based on the GMM model. In total, language classifiers trained on the LDCRF, CRF, and GMM models using PFs and language classifiers trained on PLP were applied.

Experiment (3.1). This was a comparison of the discriminative model with the generative model. The language classifiers using the LDCRF model were compared with the one using the GMM model. Firstly, the acoustic GMM system that employed the PLP feature was compared to LDCRF system with PLP features. Secondly, the LDCRF language classifiers using PF attributes were compared with the GMM system using PFs, in order to compare the performance between employing PFs with the discriminative model and employing PF with generative model. The GMM model was trained on speech utterances from each language using the expectation maximization (EM) algorithm. The best configuration contained 256 Gaussians and was initialized using 10 iterations with the maximum likelihood (ML) criterion.

Experiment (3.2). This was a comparison of the discriminative LDCRF model with the CRF model. The language classifiers based on LDCRF were compared with the classifiers based on CRF, which is another discriminative model technique in a sequential problem. We conducted the language classifiers trained from the CRF model using similar configuration parameters to the parameters used for training the LDCRF model.

Experiment (4). This was a study of the effect of dimension reduction on the PF features and a study of how the language

Table 7: Error of language classifiers using different speech feature sets (averaged across 28 pairs).

Feature	Error (%)
PF	**17.46**
SDPF	21.07
MFCC	25.48
PLP	19.51
SDCPLP	23.36
SDC + PLP	21.15

The least error result is displayed in bold.

classifiers were affected when missing PF attributes. The feature vector of PF attributes used in Experiments (1–3) consisted of 13 PF attributes. In this experiment, we iteratively removed one of the PF attributes from the feature vector. Thus, in each round, these 12 PF attributes were concatenated to 24-dimensional feature vector. The same configuration parameters as Experiment (1) were used to train the LDCRF model of language classifiers. The language classifiers were trained from the feature vector by removing one of the PF attributes. The results of this experiment were compared with the results of Experiment (1) which used a total of 13 PF attributes.

Experiment (5). The purpose of this experiment was to evaluate the performance of the proposed system with the state-of-the-art acoustic system, based on *i*-vector approach [5]. In this experiment, we used the *i*-vector space with pLDA scoring method. The concepts of *i*-vector approach and *i*-vector feature extraction are described in Section 6.5.3.

7. Results and Discussion

The performance of the experiments (as described in Section 6.6) was measured across the 28 language classifiers. The average values of accuracy and errors were computed from 5-fold cross-validation. The results from each experiment are described as follows.

For Experiment (1), Table 7 compares the performance of the language classifiers using PF, SDPF, MFCC, PLP, SDCPLP, and SDC + PLP as input features. It shows that the classifiers using the PF features were superior to the other speech features. The average error of the 28 language pairs using the PF attributes was 17.46% which was the smallest. From Table 7, the error of classifiers using different features can be represented in descending order as PF < PLP < SDPF < SDC + PLP < SDCPLP < MFCC, which indicates that it is less beneficial applying SDC to compute the SDPF and SDCPLP features, with a relative error increase of 17.13% and 16.48%, respectively.

Comparing the experiments using varying features of PLP, namely, PLP, SDCPLP, and SDC + PLP (a combination of PLP attributes and the SDC computed from PLP), the performance of PLP was better than SDC + PLP and SDCPLP. The SDCPLP option was better than SDC + PLP. However, there was no significant difference among these features

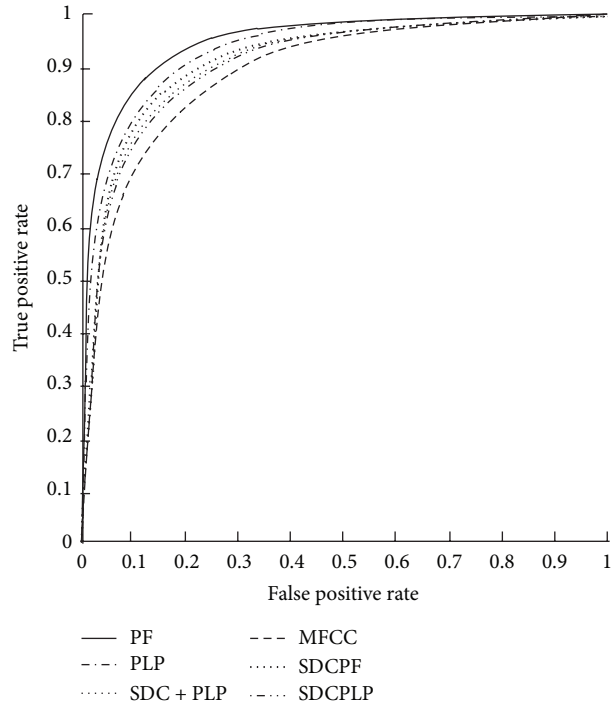

Figure 6: ROC curve classification results from 28 language pair classifiers using PF attributes versus others.

because they are highly correlated and they are derived from the same front-end feature, PLP.

Interestingly, the features performed relatively poorly when the SDC was applied to the original features, in contrast to the other research studies where applying the SDC to features achieved a better performance. This may have been caused as a result of these features being highly correlated with the original features and the LDCRF model that does not have strong dependence assumption and has the ability to incorporate the correlated observation.

Figure 6 compares the ROC curves of the classifiers using different speech features. The ROC curve of the classifier using PFs was higher than the classifier using conventional features and its SDC, which indicates that the classifier using PF features outperformed all the other classifiers.

Figure 7 compares the percentage of errors of the LDCRF language classifiers on different speech features. It demonstrates that the language classifiers using PF attributes outperformed those using the conventional features, but their performance was poorer when combining PF with the SDCs. From Figure 7, it can be noticed that different features in language classifiers also have the same trend and most of the classifiers using PFs achieved the best performance.

The performance of classifiers was also measured statistically using the paired *t*-test. There were 22 language classifiers using PFs that had better performance than the ones using PLP, three of which were significantly different: Rus_Can, Tha_Wu, and Urd_Wu. On the other hand, there were six language classifiers using PF attributes that were worse performers than using PLP: Min_Wu, Min_Urd, Arb_Urd,

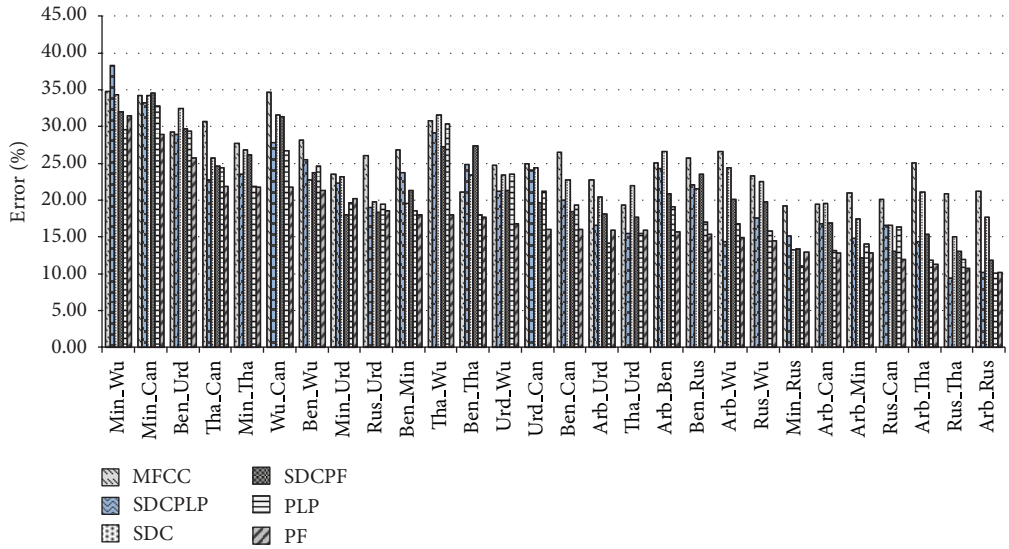

Figure 7: Errors for each of 28 language classifiers trained with LDCRF using PF attributes versus others.

Tha_Urd, Min_Rus, and Arb_Rus. Furthermore, there was no significant difference in language classifiers between using PF and PLP. There were 16 language classifiers using PFs that were significantly better than using SDCs. Additionally, there were 25 language classifiers using PFs that were better performers than SDC + PLP of which ten were significantly different: Arb_Ben, Ben_Min, Ben_Rus, Ben_Tha, ben_wuu, Min_Wu, Rus_Can, Tha_Wu, Urd_Wu, Urd_Can, and Wu_Can. Three PF language classifiers achieved better performance than using SDC such as Rus_Can, Tha_Wu, and Urd_Wu. Twenty-seven of the classifiers using PFs were better performers than SDPF, and eight of them were significantly different. There is one pair of language classifier using PFs, Min_Urd, was poorer than SDPF. The performances of all classifiers using PFs was better than using MFCC and 24 pairs had significant differences.

For Experiment (2), Table 8 shows the results of the language classifiers after removing the ambiguous Chinese languages. The results support that the discriminative ability of PF is superior to the other speech features. From comparing Tables 7 and 8, removing the confusing languages (Min Chinese and Wu Chinese) yielded better classification performance. The performance of language classifiers using PF, SDPF, MFCC, PLP, SDC, and SDC + PLP attributes achieved relative error reductions of 10.08%, 8.58%, 6.33%, 10.87%, 6.09%, and 9.99%, respectively. In future work, we plan to extend the analysis on feature reduction when removing confusing languages and to observe if there is any feature which helps improving the performance of some language classifier.

Table 9 compares the language classifier error from training with the LDCRF, CRF, and GMM models using PFs (as described in Experiment (3)). It shows that the LDCRF model achieved 17.46% error which was lower than for the CRF and GMM models. However, it was not significantly different between the LCDRF and CRF models.

Table 8: Language classifier error using different speech feature sets (averaged across 15 pairs).

Feature	Error (%)
PF	**15.70**
SDPF	19.26
MFCC	23.87
PLP	17.39
SDC	21.94
SDC + PLP	19.04

The least error result is displayed in bold.

Table 9: Error of language classifiers trained on LDCRF, CRF, and GMM models using PFs and classifiers trained on GMM using PLP (averaged across 28 pairs).

Model	Error (%)
LDCRF (PF)	**17.46**
CRF (PF)	19.11
GMM (PF)	30.63
GMM (PLP)	52.28

The least error result is displayed in bold.

Moreover, the classifiers modeled by GMMwere compared with using PFs and using the PLP as input features. The results of the experiment using PFs achieved better performance than from using the PLP, which could be due to the PFs having more discriminative ability than the PLP.

Figure 8 compares the confidence score results of CRF and LDCRF classifiers from Experiment (3) using different modeling techniques. The ROC curves of the LDCRF classifiers were slightly higher than for the CRF classifiers.

For Experiment (4), Table 10 shows the error of language classifiers trained from the feature vector that removed one of the PF attributes. The results showed that the classifiers

Table 10: Error of language classifiers trained from the feature vector after removal of one PF attribute (averaged across 28 pairs).

Removed attribute	Error (%)
None (Full PF)	**17.46**
High	17.79
Tense	17.80
Voiced	17.87
Strident	18.00
Anterior	18.02
Low	18.05
Vocalic	18.10
Continuant	18.12
Coronal	18.15
Nasal	18.17
Round	18.31
Consonantal	18.44
Back	18.60

The least error result is displayed in bold.

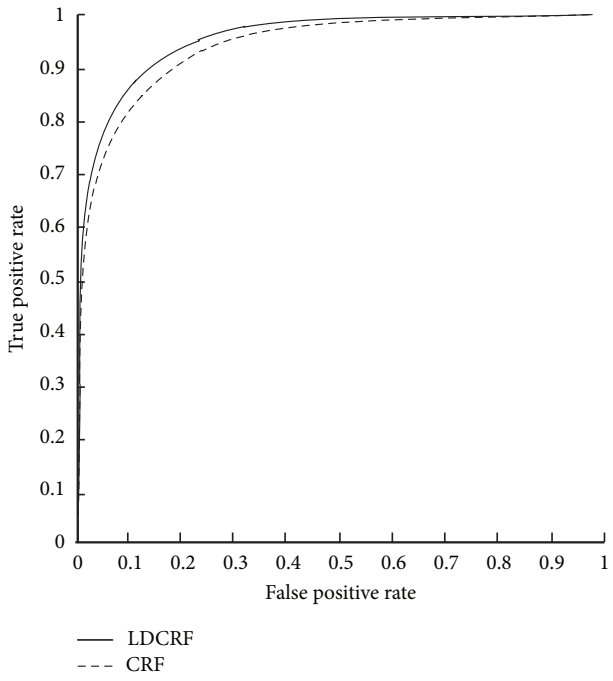

Figure 8: ROC curve classification results of LDCRF and CRF models (from 28 language classifiers).

achieved comparable performance with the one trained from the full set of PFs. Thus, removing one of the PF attributes did not significantly improve the performances of the classifiers.

The results of removing one PF attribute do not seem to be consistent. However, the overall classification results performed relatively poorly when one of PF attributes was removed from the feature vectors. This indicates that while the removal of a "*High*" attribute yielded a poorer performance in the experiment, performance was not as bad when a "*Back*" attribute was removed from the feature vector. In 11 out of 28 cases, the classifiers, with removing one PF attribute, were significantly different from the classifiers using the full set of PFs. Furthermore, there was a significant improvement when the "*Nasal*" attribute was removed from the feature vector (i.e., language classifier Arb_Urd).

From the results, we noticed that the removal of one attribute from the feature vector had a small effect on the overall accuracy. This was caused by the fact that the classifiers can use information from the other PF attributes which are in the same group of features. For example, "*High*" attribute can use information from "*Low*," "*Back*," and "*Round*" attributes (which are in the same group of vowel features). Another example is the removal of "*voiced*" attribute, which is correlated with "*Vocalic*" and "*Consonantal*" attributes, so the language can be classified by using other features instead. However, the removal of one attribute decreased the overall performance. Although each PF attribute was derived from separately-trained attribute detectors, the removal of PF attributes did not have a great impact on the performance. It can be concluded that the PF attributes in the feature vector are not truly independent.

Table 11: Error of language classifiers trained on LDCRF using PFs and acoustic based on *i*-vector approach.

Model	Error (%)
LDCRF (PF)	**17.46**
i-vector	19.94

The least error result is displayed in bold.

For Experiment (5), Table 11 shows the averaged error of language classifiers trained on LDCRF using PFs and the acoustic approach based on *i*-vector space (baseline system). The error of proposed system based on LDCRF using PFs was 17.46% and 19.94% for the baseline system. The result shows that the performance of the proposed system based on LDCRF with PFs was comparable to the performance of the acoustic system based on *i*-vector space. Even though ALIZE has been widely used to extract the *i*-vector, the tuning is needed to acquire the satisfying performance. However, it is not the main focus of this study. There were 21 language classifiers using LDCRF with PFs that had better performance than the baseline while two of them, Rus_Can and Rus_Wu, were significantly different. On the other hand, there were seven language classifiers using PF attributes that were worse performers than baseline system: Ben_Rus, Min_Rus, Rus_Tha, Tha_Wu, Urd_Wu, and Tha_Urd. However, they were not significantly different from the LCDRF models and baseline systems. The performance of the baseline system using *i*-vector was not good because it has not been fine tuned. In addition, it could be concluded from the several following reasons. Firstly, the *i*-vector approach is advantageous when sufficiently/large speech data across language is available for training the UBM supervector and adapting the language- and channel-dependent vectors; however, the speech corpus used in this paper is small; thus, it may be not suitable for the

i-vector approach. Secondly, the measurement, used to obtain a likelihood score of a testing *i*-vector is a generative pLDA model scoring. Another technique can be applied to improve the classifying of *i*-vector space. Further experiment has to employ discriminative classifiers such as SVM and Logistic Regression to classify language in *i*-vector space. Moreover, we will adapt the *i*-vector paradigm to model the proposed PF feature for SLR task in future work.

8. Conclusion

We have presented an acoustic approach SLR system applying the LDCRF model with PF attributes. From the experimental results, applying the SLR system with the discriminative LDCRF model showed significant improvement compared to the generative GMM model. In addition, employing LDCRF to learn the dynamic sequences of PF attributes for modeling the language can enhance the SLR performance. It can be an alternative discriminative method that pays attention to solving the problem of the independency among features. In terms of the discriminative ability of the speech attributes, the performance of the SLR system using PFs outperformed that using conventional speech features. The most notable conclusion that could be drawn from the experiment was that the PF attributes achieved better performance in language classification than the conventional speech features.

Conflict of Interests

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

Acknowledgments

This work was partially funded by Thailand Graduate Institute of Science and Technology (TGIST), NSTDA, and by CU. Graduate School Thesis Grant.

References

[1] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '02)*, pp. 33–36, 2002.

[2] E. Wong, J. Pelecanos, S. Myers, and S. Sridharan, "Language identification using efficient Gaussian mixture model analysis," in *Proceedings of the Australian International Conference on Speech Science and Technology*, p. 7.6, 2000.

[3] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[4] H. Li, B. Ma, and C.-H. Lee, "A Vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 271–284, 2007.

[5] O. P. B. L. Martínez, G. David, F. Luciana, and S. Nicolas, "Language recgontiion in iVector space," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '11)*, Florence, Italy, 2011.

[6] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr., "Language identification using Gaussian mixture model tokenization," in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing (ICASSP '02)*, pp. I-757–I-760, May 2002.

[7] K. E. Wong, *Automatic spoken language identification utilizing acoustic and phonetic speech information [Ph.D. thesis]*, Queensland University of Technology, 2004.

[8] C. C. L. Haipeng Wang, T. Lee, B. Ma, and H. Li, "Shifted-delta MLP features for spoken langauge recognition," *IEEE Signal Processing Letters*, vol. 20, pp. 15–18, 2013.

[9] O. S. B. Launay, O. Siohan, A. C. Surendran, and C.-H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing (ICASSP '02)*, pp. I-817–I-820, Orlando, Fla, USA, May 2002.

[10] K. Kirchhoff, *Robust speech recognition using articulatory information [Ph.D. thesis]*, Universität Bielefeld, 1999.

[11] L. Burget, P. Matějka, and J. Černocký, "Discriminative training techniques for acoustic language identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. I209–I212, May 2006.

[12] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language recognition with support vector machines," in *Proceedings of the Speaker and Language Recognition Workshop (ODYSSEY '04)*, 2004.

[13] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[14] P. A. Torres-Carrasquillo, D. Sturim, D. A. Reynolds, and A. McCree, "Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 723–726, September 2008.

[15] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[16] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Acoustic language identification using fast discriminative training," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07)*, pp. 389–392, August 2007.

[17] W. M. Campbell, "A covariance Kernel for SVM language recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4141–4144, April 2008.

[18] C. H. You, H. Li, and K. A. Lee, "A GMM-supervector approach to language recognition with adaptive relevance factor," in *Proceedings of the 18th European Signal Processing Conference*, pp. 1993–1997, 2010.

[19] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[20] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[21] B. L. Martínez González David, F. Luciana, and S. Nicolas, "Ivector-based prosodic system for language identification," in *Proceedings of the International Conference on Acoustics Speec (ICASSP '12)*, pp. 4861–4864, Kyoto, Japan, 2012.

[22] K. Kirchhoff, *Robust speech recognition using articulatory information [Ph.D. thesis]*, Universität Bielefeld, 1999.

[23] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.

[24] J. Li and C.-H. Lee, "On designing and evaluating speech event detectors," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 3365–3368, September 2005.

[25] S. Stüker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *Proceedings of the IEEE International Conference on Accoustics, Speech, and Signal Processing*, vol. 1, pp. I-144–I-147, April 2003.

[26] D. Harwath and M. Hasegawa-Johnson, "Phonetic landmark detection for automatic language identification," *Urbana*, vol. 51, 2010.

[27] U. V. Chaudhari and M. Picheny, "Articulatory feature detection with support vector machines for integration into ASR and phone recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '09)*, pp. 93–98, December 2009.

[28] S. Kanokphara, J. Macek, and J. Carson-Berndsen, "Comparative study: HMM and SVM for automatic articulatory feature extraction," in *Proceedings of the 19th International Conference on Advances in Applied Artificial Intelligence: Industrial, Engineering and other Applications of Applied Intelligent Systems (IEA/AIE '06)*, pp. 674–681, 2006.

[29] Y. H. Abdel-Haleem, *Conditional random fields for continuous speech recognition [Ph.D. thesis]*, University of Sheffield, 2006.

[30] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden conditional random fields for phone classification," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1117–1120, September 2005.

[31] D. Yu and L. Deng, "Deep-structured hidden conditional random fields for phonetic recognition," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH '10)*, pp. 2986–2989, September 2010.

[32] J. Morris and E. Fosler-Lussier, "Conditional random fields for integrating local discriminative classifiers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 617–628, 2008.

[33] D. Yu, S. Wang, Z. Karam, and L. Deng, "Language recognition using deep-structured conditional random fields," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 5030–5033, March 2010.

[34] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, June 2007.

[35] A. J. Quattoni, *Object recognition with latent conditional random fields [M.S. thesis]*, Massachusetts Institute of Technology, 2005.

[36] V. Deufemia, M. Risi, and G. Tortora, "Sketched symbol recognition with a latent-dynamic conditional model," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 1100–1103, August 2010.

[37] N. Chomsky and M. Halle, *The Sound Pattern of English*, 1968.

[38] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 8th International Conference on Machine Learning*, pp. 282–289, 2001.

[39] S. Kumar and M. Hebert, "Discriminative random fields: a discriminative framework for contextual interaction in classification," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp. 1150–1157, October 2003.

[40] N. Strom, "The NICO toolkit for artificial neural networks," http://www.speech.kth.se/NICO.

[41] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in *Proceedings of the IEEE 7th Workshop on Multimedia Signal Processing (MMSP '05)*, Shanghai, China, November 2005.

[42] P. Matějka, L. Burget, O. Glembek et al., "BUT language recognition system for NIST 2007 evaluations," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 739–742, September 2008.

[43] "NIST LRE-2007 Evaluation Plan," 2007, http://www.itl.nist.gov/iad/mig//tests/lre/2007/LRE07EvalPlan-v8b.pdf.

[44] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained Probabilistic Linear Discriminant Analysis for speaker verification," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, pp. 4832–4835, May 2011.

[45] A. Larcher, J. F. Bonastre, B. Fauve et al., "ALIZE 3.0—open source toolkit for state-of-the-art speaker recognition," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '13)*, 2013.