SUBSPACE-BASED PHONOTACTIC LANGUAGE RECOGNITION USING MULTIVARIATE DYNAMIC LINEAR MODELS

Hung-Shin Lee[1, 2], *Yu-Chin Shih*[2], *Hsin-Min Wang*[2], *Shyh-Kang Jeng*[1]

[1]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan

ABSTRACT

Phonotactics, dealing with permissible phone patterns and their frequencies of occurrence in a specific language, is acknowledged to be related to spoken language recognition (SLR) no matter the subject is a human or a machine. With the assistance of phone recognizers, each speech utterance can be decoded into an ordered sequence of phone vectors filled with likelihood scores contributed by all possible phone models. In this paper, we propose a novel approach to dig the concealed phonotactic structure out of the phone-likelihood vectors through a kind of multivariate time series analysis: dynamic linear models (DLM). In these models, treating the generation of phone patterns in each utterance as a dynamic system, the relationship between adjacent vectors is linearly and time-invariantly modeled, and unobserved states are introduced to capture a temporal coherence intrinsic in the system. Each utterance expressed by the DLM is further transformed into a fixed-dimensional linear subspace so that well-developed distance measures between two subspaces can be applied to linear discriminant analysis (LDA) in a dissimilarity-based fashion. The results of SLR experiments on the OGI-TS corpus demonstrate that the proposed framework outperforms the well-known vector space modeling (VSM)-based methods and achieves comparable performance to our previous subspace-based method.

Index Terms**—** phonotactic language recognition

1. INTRODUCTION

Nowadays, there are up to 7300 distinct languages spoken around the world. Due to the globalization, communication between different races or countries is becoming a more and more important issue. The springing up of a variety of multi-lingual services brought about the birth of automatic spoken language recognition (SLR), which is the process of identifying or verifying the language spoken in a speech utterance by means of a computer, and has appealed to many researchers to explore for more than twenty years. Typically, SLR techniques fall into two major categories according to the level of information that machines can use to distinguish one language from another [1]. At the acoustic level, a series of frames, each containing 80-200 ms temporal information, such as mel-frequency cepstral coefficients (MFCCs) or shifted delta cepstral (SDC) features [2], is derived from raw speech through speech parameterization processes. These acoustic frames are directly fed into back-end models, such as Gaussian mixture models (GMMs) [3], to form a vector-like representation for each utterance. In contrast, phonotactic approaches exploit phone recognizers to further convert acoustic frames of an speech utterance into a phone sequence to capture longer-term information. It is supposed that if phonotactic constraints, e.g., permissible syllable structures or phone pattern combinations, across languages for each speech utterance can be found out or stochastically described, the characteristics of each language will be well modeled [4], and much better recognition performance will be potentially gained.

Up to the present, various implementations to deal with phonotactic features have been proposed, from the use of several *n*-gram language model-based phonetic decoders for each single language (PPR) [5], the use of a single phonetic decoder followed by the computation of language dependent phone *n*-gram likelihoods (PRLM) [6], to the use of paralleled single-language phonetic decoders followed by a phone *n*-gram classifier (PPRLM) [7]. Some researchers have looked into a language-independent phone recognizer using a set of universal acoustic units or phones that is common for all languages. Along with the usage of the universal phone set (UPS), a new paradigm for modeling phonotactic constraints, namely vector space modeling (VSM), has been developed [8]. Stemming from the well-known VSM framework in the field of information retrieval (IR), Li *et al.* built a composite feature vector for each utterance by concatenating the vector-formed statistics from phonetic decoders, and applied support vector machines (SVM) to the composite vectors for classification [8]. In their work, each of phone or sound sequences was represented by a high dimensional phonotactic feature vector with the *n*-gram counts or term frequency-inverse document frequency (TF-IDF) weights, whose dimensionality is equal to the total number of phonotactic patterns needed to characterize the structure of the utterance given by a decoder. Moreover, Penagarikano *et al*. took time alignment information into account by considering time-synchronous cross-decoder phone co-occurrences [9]. They have thus defined a new concept of multi-phone labels, which attempts to integrate the contributions given by several decoders frame by frame and form a VSM-based label sequence different from the conventional *n*-gram patterns.

From the perspective of data representation, VSM has merits to well serve the back-end learning mechanism like SVM or GMM by transforming each varying-length phone sequence into a fixed-length vector. Nevertheless, more phonotactic attributes, such as bigram and trigram terms, have to be included to form a much higher dimensional vector, so as 1) to settle the order-losing issue that the sequential order of the phones in a decoded utterance is lost, and 2) to relax the assumption that unigrams (single phones) are statistically independent to some extent. Since the total number of *n*-gram patterns tends to increase exponentially with respect to *n*, *n* is often inevitably limited to 3 (trigram) or 4 (4-gram) with discriminative selection of the *n*-grams [10]. Practically, it is necessary to apply dimension reduction approaches, such as latent semantics analysis (LSA) [8] and principal component analysis (PCA) [11], on the original VSM-based feature vectors to make the classification task more efficient and to avoid "the curse of dimensionality" while training data are deficient. Recently, stemming from the idea of "i-vectors", which has provided superior performance in the speaker recognition field [12], some researchers have used probabilistic PCA (PPCA) to transform each high-dimensional vector filled with discrete features into a small-size set of latent variables corresponding to a high variability subspace [13-15, 30]. For example, in [13] and [14], Soufifar *et al.* used the subspace multinomial model, along with the maximum likelihood criterion, to effectively represent the information contained in the *n*-grams.

In this paper, as an extension of our previous subspace-based work in [16], we propose a new approach for data representation, in which the phonetic information as well as the contextual relationship can be more abundantly retrieved by likelihood computation and dynamic linear models, given a universal phone recognizer. In the VSM framework, the count or frequency of a phonotactic term is the only attribute that is concerned. In contrast, our approach enables us to look much farther and deeper through the decoder's eyes without much more memory. That is, not only can more information, such as likelihood scores of any phone segments, be captured, but also all possible phones can be taken into account instead of the single most likely one. The spirit is somewhat similar to the employment of phone lattices [17] or posteriogram-based *n*-gram counts [15]. Moreover, our representation also fits for the back-end classification. Under the assumption that the utterance representation can be approximately described by a collection of lower dimensional linear subspaces, a suitable dissimilarity-based learning algorithm along with the well-surveyed Projection metric are introduced for classification.

The remaining of this paper is organized as follows. In Section 2, we introduce the new representation of a speech utterance in the sense of dynamic linear models. In Section 3, we present the learning and scoring mechanisms for subspaces with the Projection metric. Section 4 gives the evaluation results and some discussions. Finally, conclusions and future work are outlined in Section 5.

2. SUBSPACE-BASED REPRESENTATION

According to the definition in [19, p. 3, 20], subspace-based learning is based on the extraction of the most conspicuous properties of each *class* separately, as represented or spanned by vector series expansions constructed from the feature vectors of each *class*. The learning mechanism focuses on how to measure the similarity between each subspace and a given data point. However, most widely-alleged subspace-based methods for the SLR tasks, such as i-vectors and PCA, do not fall into this definition. On the contrary, their goal is to derive a coordinate representation (or a vector-like point) for an utterance in a lower-dimensional linear space where all projected utterances are supposed to share the same set of bases. In contrast, we will introduce a new approach to represent each utterance as a subspace of the original feature space, where the salient structure of each utterance can be preserved.

2.1. Phone-likelihood vectors

Given the observed sequence of acoustic feature vectors $\mathbf{o}_1, \dots, \mathbf{o}_T$ derived from a speech utterance, a phone decoder singles out the best phone sequence $p_1, \dots, p_K$ based on Viterbi decoding which finds the most likely time alignment path through a huge probabilistic network. The main task of phonotactic-based language classifiers is to take advantage of the phone sequence $p_1, \dots, p_K$ as a basic unit for SLR. The basic idea behind our proposed phonotactic data representation is to take the single-best phone sequence given by the phone decoder as a kind of clusters toward the acoustic feature vectors in a phonotactic fashion. From the example in Figure 1, we can see that after phone decoding and time alignment, feature vectors $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}$ belong to top-1 phone u_1 while feature vectors $\{\mathbf{o}_4, \mathbf{o}_5, \mathbf{o}_6, \mathbf{o}_7\}$ belong to top-1 phone u_2. Along this vein, given phone boundaries, each set of acoustic vectors in its corresponding phone segment is further used to derive a more meaningful phonotactic feature vector built up with the average log-likelihood score for each phone, which is characterized by a hidden Markov model (HMM), through the Viterbi search algorithm. Consequently, each phone segment (or phone frame) p_i is expressed by a phone-likelihood vector $\mathbf{y}_i$, whose dimensionality is the size of the universal phone set $\{u_i\}$. Figure 1 also shows that the first phone p_1, which is most likely labeled as u_1, indeed has the highest log-likelihood score of -14 with respect to the first attribute u_1 due to the nature of dynamic programming contributed by the first three acoustic feature vectors $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}$. As for other attributes u_2, u_3, and u_4 in p_1, although their scores might be much smaller than the single-best u_1, and even u_3 might never appear in the single best phone sequence, they are not cast off but included into the phone-likelihood vector $\mathbf{y}_1$ to bring more uncertainty or information that might be helpful for classification.

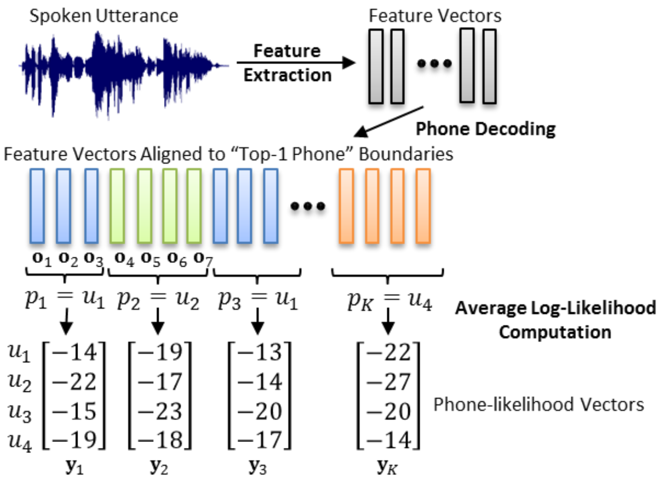
Spoken Utterance
Feature Extraction
Feature Vectors
Phone Decoding
Feature Vectors Aligned to "Top-1 Phone" Boundaries
$\mathbf{o}_1$ $\mathbf{o}_2$ $\mathbf{o}_3$ $\mathbf{o}_4$ $\mathbf{o}_5$ $\mathbf{o}_6$ $\mathbf{o}_7$
$p_1 = u_1$ $p_2 = u_2$ $p_3 = u_1$ $p_K = u_4$
Average Log-Likelihood Computation
u_1 u_2 u_3 u_4
−14 −22 −15 −19
−19 −17 −23 −18
−13 −14 −20 −17
−22 −27 −20 −14
Phone-likelihood Vectors
$\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$ $\mathbf{y}_K$

Figure 1. *Phone-likelihood vector extraction for an SLR system with four universal phones* $\{u_1, u_2, u_3, u_4\}$.

2.2. Dynamic linear models

After the aforementioned procedure, each utterance is expressed by a sequence of phone-likelihood vectors, which is more than just a set of vectors due to the temporal information, especially phonotactic constraints, contained in the sequence. To capture the temporal dynamics, we make a conjecture that each sequence was generated by a causal linear time-invariant (LTI) system, which might be a sub-system pertaining to some *language production* system. This conjecture is similar to the acoustic theory of speech production that assumes the speech production process to be a linear system, consisting of a source and filter [29]. A multivariate dynamic linear model (DLM), also known as a state space model, is one of causal LTI systems, which has been used to model moving human bodies or textures in computer vision [21] and signal analysis [22]. A simpler DLM to model the phone-likelihood sequence is described as follows.

Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1,\dots,K}, \mathbf{y}_i \in \mathcal{R}^D$ be a sequence of K phone-likelihood vectors for an utterance, where D denotes the size of the universal phone set, and $\{\mathbf{x}_i\}_{i=1,\dots,K}$, $\mathbf{x}_i \in \mathcal{R}^d$ be unobservable vectors representing the state of the system, where $d \le D$ denotes the system complexity or the state order. Then a DLM is specified by the following equations:

$$\mathbf{y}_i = \mathbf{C}\mathbf{x}_i + \mathbf{D}\mathbf{w}_i,\ \mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{W}); \tag{1}$$

$$\mathbf{x}_{i+1} = \mathbf{A}\mathbf{x}_i + \mathbf{B}\mathbf{v}_i,\ \mathbf{v}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{V}), \tag{2}$$

where $\mathbf{v}_i$'s and $\mathbf{w}_i$'s are evolution vectors and observation error vectors, respectively, which can be taken as additive noises with zero-mean Gaussian distributions. The matrices $\mathbf{A} \in \mathcal{R}^{d\times d}$ and $\mathbf{C} \in \mathcal{R}^{D\times d}$ are the system transfer operators that need to be estimated. To learn the dynamic system amounts to identifying the model parameters $\mathbf{A}$ and $\mathbf{C}$, while both $\mathbf{B}$ and $\mathbf{D}$ are often assumed to be identity matrices $\mathbf{I_B} \in \mathcal{R}^{d\times d}$ and $\mathbf{I_D} \in \mathcal{R}^{D\times d}$ for simplicity. It is commonly assumed that the distribution of the input sequence $\{\mathbf{y}_i\}_{i=1,\dots,K}$ is known, so that we can use maximum likelihood estimation (MLE) to infer the states $\{\mathbf{x}_i\}_{i=1,\dots,K}$ and to estimate the parameters from the observed sequence $\{\mathbf{y}_i\}_{i=1,\dots,K}$ by the expectation-maximization (EM) algorithm. However, in this paper, we adopt an alternative approach, which is based on reconstruction error minimization for (1) and (2) to estimate the parameters and can be proven to be asymptotically efficient approaching the ML solution.

Without any iterative procedures, the sub-optimal, closed-form, and fast solution starts with the singular value decomposition (SVD) of the observed sequence $\{\mathbf{y}_i\}_{i=1,\dots,K}$. Let $\mathbf{U\Sigma V}^T \approx \mathbf{Y}$ be the SVD of $\mathbf{Y}$, then $\mathbf{\Sigma}$ is a $d \times d$ diagonal matrix containing the largest d singular values $\sigma_1, \dots, \sigma_d$ while $\mathbf{U} \in \mathcal{R}^{D\times d}$ and $\mathbf{V} \in \mathcal{R}^{K\times d}$ are matrices, whose orthonormal columns $\{\mathbf{u}_i\}_{i=1,\dots,d}$ and $\{\mathbf{v}_i\}_{i=1,\dots d}$ approximately span the column and row spaces of $\mathbf{\Sigma}$, respectively. The unique parameters $\mathbf{A}$ and $\mathbf{C}$, and the states $\mathbf{X}_1^K = [\mathbf{x}_1, \dots, \mathbf{x}_K]$ can be sequentially estimated by

$$\hat{\mathbf{C}} = \mathbf{U}, \tag{3}$$

$$\hat{\mathbf{X}}_1^K = \mathbf{\Sigma V}^T, \tag{4}$$

$$\hat{\mathbf{A}} = arg\, min_{\mathbf{A}} \textstyle\sum_{i=1}^{K-1} \|\hat{\mathbf{x}}_{i+1} - \mathbf{A}\hat{\mathbf{x}}_i\|^2. \tag{5}$$

If $[\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{K-1}]$ and $[\hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_K]$ are expressed by $\hat{\mathbf{X}}_1^{K-1} \in \mathcal{R}^{d\times(K-1)}$ and $\hat{\mathbf{X}}_2^K \in \mathcal{R}^{d\times(K-1)}$, respectively, we can see that (5) is actually a problem of linear regression with least-squares estimation, which leads to a closed-form estimation of $\mathbf{A}$ as:

$$\hat{\mathbf{A}} = \hat{\mathbf{X}}_2^K \hat{\mathbf{X}}_1^{K-1^T} (\hat{\mathbf{X}}_1^{K-1} \hat{\mathbf{X}}_1^{K-1^T})^{-1}, \tag{6}$$

From (3) and (6), we derive $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ for each utterance, which can be seen as a compact and decoder-recognized representation.

2.3. Distance between two linear models

In [23], Cock and Moor described how to measure the distance between any two DLM models. For two utterances, given their corresponding DLM estimates $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$, their subspace-based representations can be respectively expressed by

$$\mathbf{S}_1 = \left[\hat{\mathbf{C}}_1^T, (\hat{\mathbf{C}}_1\hat{\mathbf{A}}_1)^T, (\hat{\mathbf{C}}_1\hat{\mathbf{A}}_1^2)^T, \dots, (\hat{\mathbf{C}}_1\hat{\mathbf{A}}_1^d)^T\right]^T, \tag{7}$$

$$\mathbf{S}_2 = \left[\hat{\mathbf{C}}_2^T, (\hat{\mathbf{C}}_2\hat{\mathbf{A}}_2)^T, (\hat{\mathbf{C}}_2\hat{\mathbf{A}}_2^2)^T, \dots, (\hat{\mathbf{C}}_2\hat{\mathbf{A}}_2^d)^T\right]^T, \tag{8}$$

where $\mathbf{S}_1, \mathbf{S}_2 \in \mathcal{R}^{(Dd)\times d}$ and d denotes not only the state order but also the size of the contextual window. For example, $d = 2$ means the subspace contains information of any three consecutive phones.

Note that the representation may not be unique since (1) and its solutions (3) and (4) are invariant to any coordinate transformations. Therefore, if we consider the linear span of the column vectors of $\mathbf{S}_i$ instead of the matrix $\mathbf{S}_i$ itself to represent the dynamical system, the distance measurement between $\mathbf{S}_1$ and $\mathbf{S}_2$ must be also invariant to such transformations. In this paper, under the assumption that $\mathbf{S}_1$ and $\mathbf{S}_2$ are equivalent if and only if $span(\mathbf{S}_1) = span(\mathbf{S}_2)$, we adopt the Projection metric as a reasonable distance measure between two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$ of two utterances [18], which is defined by

$$d_{Proj}(\mathbf{S}_1, \mathbf{S}_2) = \left(d - \textstyle\sum_{k=1}^{d} cos^2\, \theta_k\right). \tag{9}$$

$cos\,\theta_k$ is the cosine of the *k*-th principal angle between $span(\mathbf{S}_1)$ and $span(\mathbf{S}_2)$, also known as the *k*-th canonical correlation [24], which is further defined by

$$cos\,\theta_k = max_{\mathbf{a}_k \in span(\mathbf{S}_1), \mathbf{b}_k \in span(\mathbf{S}_2)} (\mathbf{a}_k \cdot \mathbf{b}_k), \tag{10}$$

subject to $\|\mathbf{a}_k\| = \|\mathbf{b}_k\| = 1$, and $\mathbf{a}_k \cdot \mathbf{a}_i = \mathbf{b}_k \cdot \mathbf{b}_i = 0$ $(i = 1, \dots, k-1, k \le d)$. Since $\mathbf{S}_1$ and $\mathbf{S}_2$ both have *linear independent* columns, $cos\,\theta_k$ can be easily derived through the SVD of $\mathbf{S}_1^T\mathbf{S}_2$, $\mathbf{S}_1^T\mathbf{S}_2 = \mathbf{U}(cos\,\mathbf{\Theta})\mathbf{V}^T$, where $cos\,\mathbf{\Theta} = diag(cos\,\theta_1, \dots, cos\,\theta_d)$ ([25], p. 604).

3. DISSIMILARITY-BASED LEARNING SCHEME

Since the subspace-based pattern (7) is not suitable for classifiers designed only for vectorial inputs, we need a dissimilarity-based learning algorithm that depends only on the distance metric (9) to discriminate utterances for the training and detection phases, which is briefly summarized as follows [26]:

1) In the training phase, we first construct the dissimilarity matrix $\mathbf{D} \in \mathcal{R}^{n\times n}$, where n denotes the total number of training utterances, and each entry $d_{i,j}$ corresponds to the dissimilarity between the pair of utterances (subspaces) i and j computed through the Projection metric (9). Thus, the *i*-th row of $\mathbf{D}$, $\mathbf{d}_i$, represents a new *n*-dimensional feature vector (called the dissimilarity vector) of utterance i $(i = 1, \dots, n)$.
2) In order to discriminate utterances of different languages, we perform linear discriminant analysis (LDA) on all the training dissimilarity vectors according to their language labels, and derive a projection matrix $\mathbf{A}$ and the transformed mean vector $\mathbf{m}_i$ for language L_i $(i = 1, \dots, C)$.
3) In the test phase, we represent each test utterance (subspace) as a dissimilarity vector by using the same measure against all the training utterances (subspaces).
4) We then project the test dissimilarity vector by $\mathbf{A}$ to form a lower-dimensional vector $\mathbf{w}$, and thus achieve the classification through invoking a classifier built in the dissimilarity space.

In the final scoring stage, assuming that the prior probabilities of all the target languages are equal, the decision score between a target language L_{tar} and the test utterance $\mathbf{w}_{test}$ can be expressed based on the log-posterior probability and computed by

$$score(L_{tar}, \mathbf{w}_{test}) = -G(\mathbf{w}_{test}, \mathbf{m}_{tar}) + \textstyle\sum_{i=1}^{C} G(\mathbf{w}_{test}, \mathbf{m}_i),$$

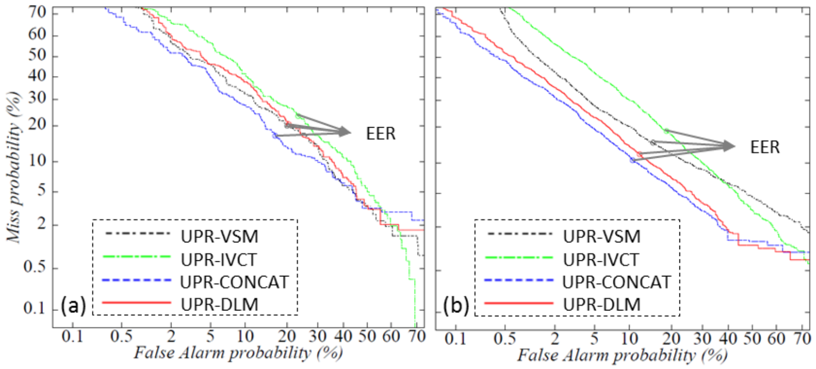

Figure 2. *DET plots for two VSM-based and two subspace-based approaches on (a) the 3-s and (b) the 1-to-50-s OGI-TS data sets.*

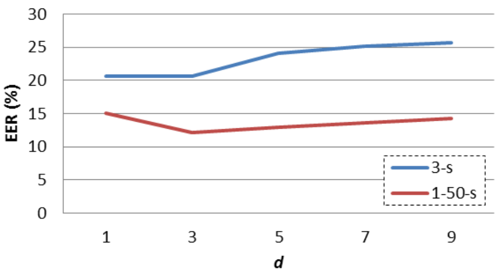

Figure 3. *EER with respect to d, the state order, in UPR-DLM on the 3-s and 1-to-50-s OGI-TS data sets.*

Table 1. *EER (%) for various phonotactic approaches on the 3-s and 1-to-50-s OGI-TS data sets.*

Methods	1-to-50-s data set	3-s data set
UPR-VSM	15.12	19.81
UPR-IVCT	18.62	23.48
UPR-CONCAT	10.68	16.78
UPR-DLM	**12.09**	**20.58**

where $G(\mathbf{w}_i, \mathbf{m}_j) = (\mathbf{w}_i - \mathbf{m}_j)\mathbf{S}_w^{-1}(\mathbf{w}_i - \mathbf{m}_j)$, and $\mathbf{S}_w$ denotes the within-class covariance matrix derived in the projective space.

4. EXPERIMENTS

We conducted the language verification task on the Oregon Graduate Institute Multi-language Telephone Speech (OGI-TS) Corpus [27], which contains the speech of 10 languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. The corpus is divided into three parts: 4650 utterances for training, 1899 utterances for development, and 1848 utterances for test. Some test utterances with lengths ranging from 2 to 4 seconds are culled to form the 3-s set to evaluate the system performance for short utterances, while all test utterances form the 1-to-50-s set. Besides, the corpus also includes 619 “story-before-tone” utterances, which have manually generated fine-phonetic transcriptions that can be used for supervised phone modeling for six languages.

A universal phone recognizer (UPR), which is composed of a set of language-independent context-independent phone models, is used in the experiments. Each phone model models a phone in the universal phone inventory of size 69, which is a union of the phones appearing in the 619 transcription files, with the phones sharing the same manner and place merged into one. Each phone model is a 3-state left-to-right CD-HMM with 32 Gaussian mixture components per state. The acoustic feature vector has 39 attributes comprised of 13 MFCCs including C0, along with their first and second order derivatives. According to the procedure shown in [8], all phone models were trained and refined from the 619 phone-transcribed utterances and the 4650 training utterances according to the maximum likelihood criterion, respectively.

Given the same UPR, we compared the proposed method with two well-known VSM-based methods, namely UPR-VSM [8] and UPR-IVCT [13], and one subspace-based method, UPR-CONCAT [16]. For UPR-VSM, a phone sequence is represented by a $(69 + 69^2 + 69^3)$ dimensional vector consisting of the TF-IDF values of unigram, bigram, and trigram phonotactic patterns. Latent semantic indexing (LSI) was further used for extracting 2000-dimensional key features needed for discriminating utterances from the statistics of some salient units and their co-occurrences. However, in UPR-IVCT, only unigram and bigram phonotactic patterns were used in the multinomial subspace model to train the 700-dimensional i-vector of each utterance [28]. Instead of the DLM mentioned in Section 2.2, UPR-CONCAT models the phonotactic information within an utterance by simply concatenating the phone-likelihood vectors belonging to a fixed-length sliding window centered on a current vector, whose size was set to be 3 in our experiment. All of the above parameter settings were determined by the experiments on the developing data set.

To make fair comparisons, in the back-end classification, we used LDA and its corresponding scoring technique mentioned in the end of Section 3 in all of the four methods. From Figure 2 and Table 1, we can see that when $d = 3$, UPR-DLM outperforms two VSM-based methods, UPR-VSM and UPR-IVCT, in terms of the equal error rate (EER) nearly on both data sets, and achieves comparable performance to UPR-CONCAT on the 1-to-50-s data set. Some possible reasons why UPR-DLM performs worse than UPR-CONCAT lie in that, 1) the language production process we attempt to model may have some nonlinearity effects that linear systems cannot fully describe; 2) the compact solutions of model parameters shown in (3) and (6) may not be close enough to the true solutions although the computation is efficient; 3) in UPR-CONCAT, SVD acts as a robust subspace generator that allows for high discrimination against noise contamination even when the phone decoder is not reliable, but UPR-DLM lacks this kind of operations.

Figure 3 shows the EER with respect to the state order (d). We can see that the minimal ERR is achieved when $d = 3$, which means that the phonotactic information contained in 4 consecutive phone-likelihood vectors are considered. Compared with the results of UPR-CONCAT, it seems to imply that the maximum size of phonotactic constraints can be set to 3 (trigram) or 4 (4-gram) in most of the phonotactic SLR tasks.

5. CONCLUSIONS

This paper presents a new phonotactic feature representation based on dynamic linear models and subspace formulation for automatic spoken language recognition. On the basis of the representation, the combination of the dissimilarity-based learning algorithm and LDA has been shown to perform well. In our future work, we plan to remedy the flaws found in the proposed framework and evaluate it on the NIST LRE corpora. Other nonlinear subspace-based methods will also be investigated, implemented, and compared in the experiments.

6. REFERENCES

[1] E. Ambikairajah *et al.*, “Language identification: a tutorial,” *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, 2011.

[2] P. A. Torres-Carassquilo *et al*., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in *Proc. ICSLP*, 2002.

[3] E. Wong and S. Sridharan, “Methods to improve Gaussian mixture model based language identification system,” in *Proc. ICSLP*, 2002.

[4] A. S. House and E. P. Neuburg, “Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations,” *J. Acoust. Soc. Amer.*, vol. 62, no. 3, 1977.

[5] Y. K. Muthusamy *et al*. “A comparison of approaches to automatic language identification using telephone speech,” in *Proc. Eurospeech*, 1993.

[6] T. J. Hazen and V. W. Zue, “Segment-based automatic language identification,” *J. Acoust. Soc. Amer.*, vol. 101, no. 4, 1997.

[7] M. A. Zissman and E. Singer, “Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling,” in *Proc. ICASSP*, 1994.

[8] H. Li *et al*., “A vector space modeling approach to spoken language identification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, 2007.

[9] M. Penagarikano *et al.*, “Improved modeling of cross-decoder phone co-occurrences in SVM-based phonotactic language recognition,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 19, no. 8, 2011.

[10] F. S. Richardson and W. M. Campbell, “Language recognition with discriminative keyword selection,” in *Proc. ICASSP*, 2008.

[11] T. Mikolov *et al.*, “PCA-based feature extraction for phonotactic language recognition,” in *Proc. Odyssey*, 2010.

[12] N. Dehak *et al.*, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 19, no. 4, 2011.

[13] M. Soufifar *et al.*, “iVector approach to phonotactic language recognition,” in *Proc. Interspeech*, 2011.

[14] M. Soufifar *et al.*, “Discriminative classifiers for phonotactic language recognition with iVectors,” in *Proc. ICASSP*, 2012.

[15] L. F. D’Haro et al., “Phonotactic language recognition using i-vectors and phoneme posteriogram counts,” in *Proc. Interspeech*, 2012.

[16] Y.-C. Shih *et al.*, “Subspace-based feature representation and learning for language recognition,” in *Proc. Interspeech*, 2012.

[17] J. L. Gauvain *et al*., “Language recognition using phone lattices,” in *Proc. ICSLP*, 2004.

[18] J. Hamm and D. D. Lee, “Grassmann discriminant analysis: a unifying view on subspace-based learning,” in *Proc. ICML*, 2008.

[19] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, 1983.

[20] E. Oja and T. Kohonen, “The subspace learning algorithm as a formalism for pattern recognition and neural networks,” in *Proc. IEEE Int. Conf. on Neural Networks*, vol. 1, 1988.

[21] G. Doretto *et al.*, “Dynamic Textures,” *International Journal of Computer Vision*, vol. 51, no. 2, 2003.

[22] A. Veen *et al.*, “Subspace based signal analysis using singular value decomposition,” *Proceedings IEEE*, vol. 81, 1993.

[23] K. D. Cock and B. D. Moor, “Subspace angles between ARMA models,” *Systems Control Lett.*, vol. 46, no. 4, 2002.

[24] A. Björck and G. H. Golub, “Numerical methods for computing angles between linear subspaces,” *Math. Computatio*n, vol. 27, no. 123, 1973.

[25] G. H. Golub and C. F. V. Loan, *Matrix Computations*, JHU Press, 1996.

[26] S.-W. Kim and R. P. W. Duin, “An empirical comparison of kernel-based and dissimilarity-based feature spaces,” Proc. SSPR&SPR, 2010.

[27] A. Y. K. Muthusamy *et al*., “The OGI multi-language telephone speech corpus,” in *Proc. ICSLP*, 1992.

[28] M. Kockmann *et al.*, “Prosodic speaker verification using subspace multinomial models with intersession compensation,” in *Proc. Interspeech*, 2010.

[29] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Signal Processing Series, Prentice-Hall Inc., Engelwood Cliffs, New Jersey, 1978.

[30] M. Tipping and C. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural Computation*, vol. 11, 1999.