SPOKEN LANGUAGE RECOGNITION USING SUPPORT VECTOR MACHINES WITH GENERATIVE FRONT-END

Kong-Aik Lee, Changhuai You, and Haizhou Li

Institute for Infocomm Research (I^{2}R)
Agency for Science, Technology and Research (A*STAR), Singapore
{kalee,echyou,hli}@i2r.a-star.edu.sg

ABSTRACT

This paper introduces a spoken language recognition system with a generative front-end and a discriminative backend. The generative front-end is built upon an ensemble of Gaussian densities. These Gaussian densities are trained to represent elementary speech sound units characterizing a wide variety of languages. We formulate the generative front-end in a form of sequence kernel. This sequence kernel transforms a spoken utterance into a feature vector with its attributes representing the occurrence statistics of the speech sound units. A discriminative support vector machine (SVM) then operates on the feature vectors to make classification decision. The proposed language recognition system demonstrates competitive performance on NIST 1996, 2003 and 2005 LRE corpora.

Index Terms— Language recognition, support vector machine, sequence kernel

1. INTRODUCTION

Automatic spoken language recognition is a process of determining the language spoken in an utterance. It has become an essential technique in applications, such as, multilingual speech recognition and spoken document retrieval [1]. One of the best known approaches to automatic spoken language recognition is the PPRLM (parallel phone recognition followed by language modeling) [1, 2]. In this approach, a set of phone recognizers are used to transcribe an input utterance into parallel sequences of phone tokens, one from each recognizer. Generative *n*-gram language models (LMs) are then derived for each target language to capture the statistics of the phone sequences generated by the PPR front-end. During recognition, a test utterance is first converted into phone sequences and then scored against the LMs. Another successful approach is the PPR-VSM (parallel phone recognition followed by vector space modeling) that replaces the generative *n*-gram LMs with support vector machines (SVMs) to exploit the advantages of a discriminative back-end [3].

The PPR front-end can be seen as a decoder that extracts relevant phonotactic information for language recognition. In this paper, we propose a simple front-end, where the elementary phonological units (i.e., speech sound categories) are now modeled with acoustically-defined Gaussian densities, instead of linguistically-defined phone recognizers modeled with hidden Markov models [1]. Our intention is to circumvent the need of laborious phonetic transcription in training the phone recognizers, while achieving competitive performance with the PPR approach. This idea was earlier explored in the framework of Gaussian mixture model (GMM) using the shifted-delta-cepstral (SDC) coefficients [4]. The long-term speech dynamic encoded in the SDC coefficients enables simple Gaussian densities in capturing phonotactic information essential for language discrimination.

In this paper, we formulate the generative front-end in a form of sequence kernel. Used as part of an SVM [5], the sequence kernel explicitly maps variable length speech utterances into higher-dimensional vectors for discriminative classification. In this sense, the operation of our kernel is similar to the generalized linear discriminant sequence (GLDS) kernel proposed in [6], except that the GLDS uses polynomial expansion at the front-end. Our system employs an ensemble of Gaussian densities at the front-end and adopts a discriminative back-end, leading potentially to better language discrimination.

We derive the sequence kernel from the generalized radial basis function (RBF) network and minimum squared-error (MSE) training criterion [7]. We have used similar approach to derive another sequence kernel for speaker recognition with modest success [8]. Here, we refine and tailor the approach for language recognition. By means of a self-organized ensemble of Gaussian densities, the resulting sequence kernel is found extremely effective in capturing language-dependent information. We also present a more rigorous formulation of sequence kernel SVM and a fast technique for sequence kernel computation in this paper.

2. SELF-ORGANIZED SOUND INVENTORY

Spoken languages differ in the inventory of speech sound units used to produce words [1]. Although the frequency of occurrence and the order of these sounds appear in a spoken utterance differs from one language to another, common speech sounds are shared considerably across languages [3]. In view of this, a universal inventory of speech sounds can be established by combining those from a predefined set of languages, which we refer to as the basis languages.

Consider that we have access to collections of speech samples for a predefined set of *K* basis languages, and the speech samples have been parameterized as sequences of feature vectors. We can train a Gaussian mixture model (GMM) for each basis language C_k, where $k = 1, 2, \ldots, K$, in the following form:

$$p(\mathbf{x} \mid C_k) = \sum_{j=1}^{M} p(\mathbf{x} \mid j) P(j \mid C_k), \qquad (1)$$

where M denotes the number of Gaussian components in the mixture, and $P(j \mid C_k)$ is the mixture weight for the jth Gaussian component $p(\mathbf{x} \mid j) \sim \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$ with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Assuming equal priors, we can pool the basis GMMs in (1) to obtain a composite GMM in the following form

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid C_k) P(C_k) = \sum_{j=1}^{KM} p(\mathbf{x} \mid j) P(j), \qquad (2)$$

where $P(j)$ are the mixture weights after renormalization such that all the weights sum to one, since we assumed $P(C_k) = 1/K$. The index j now ranges from 1 to KM as there are KM distinct components in the resulting mixture.

The GMM in (2) can be thought of to represent the underlying process that generates the multilingual data. The individual Gaussian components $p(\mathbf{x} \mid j)$ are trained to represent the underlying set of speech sound units (vowel, nasal, fricative etc.) in a self-organized manner. The weight $P(j)$ represents the *a priori* probability of occurrence of a particular sound in a spoken utterance. This viewpoint has long been postulated in the literature [9, pp. 719] and, in our case, further supported by the SDC coefficients that capture long-term spectral information across a large number of frames [4]. This is consistent with the PPR front-end where phonotactic tokenization is performed over multiple frames.

3. SEQUENCE KERNEL DERIVATION

In (2), we obtain an ensemble of Gaussian densities defining a universal inventory of speech sounds. We are now in the position to derive a sequence kernel that utilizes these density functions to characterize speech utterances. We motivate the approach based upon the generalized radial basis function (RBF) network and minimum squared-error (MSE) training criterion [7].

3.1. Generalized RBF network

Using Bayes' theorem, the set of KM Gaussian density functions in (2) can be written in normalized form as

$$\gamma_j(\mathbf{x}) = \frac{p(\mathbf{x} \mid j) P(j)}{\sum_{j'=1}^{KM} p(\mathbf{x} \mid j') P(j')} \quad \text{for } j = 1, 2, \ldots, KM. \qquad (3)$$

Using this set of normalized Gaussian basis functions we form a generalized RBF network [7] as shown in Fig. 1. The output of the network can be represented in a compact form as

$$f(\mathbf{x}_n) = \sum_{j=1}^{KM} w_j \gamma_j(\mathbf{x}_n) = \mathbf{w}^T \boldsymbol{\gamma}(\mathbf{x}_n), \qquad (4)$$

where $\mathbf{w} \equiv [w_1, w_2, \ldots, w_{KM}]^T$ is the network weights vector, and $\boldsymbol{\gamma}(\mathbf{x}_n) \equiv [\gamma_1(\mathbf{x}_n), \gamma_2(\mathbf{x}_n), \ldots, \gamma_{KM}(\mathbf{x}_n)]^T$ is the vector of normalized Gaussian basis functions. The activations of the basis functions indicate the probabilities of the presence of corresponding speech sounds given an observation vector $\mathbf{x}_n$. Since each language has its own inventory of sound units, difference subsets of basis functions will be activated by feature vectors belonging to different languages. The network in Fig. 1 can therefore be trained to discriminate a target language from some competing languages depending on the activations of its basis functions.

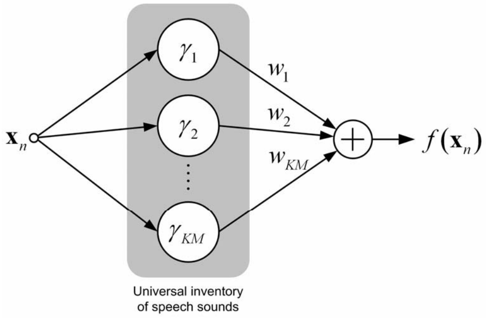

Fig. 1. A generalized radial basis function (RBF) network. Each basis function represents elementary sound units characterizing speech utterances.

3.2. MSE discriminative training

Consider a binary classification problem. The network weights $\mathbf{w}_{\text{tgt}}$ can be determined for a target language class by minimizing the following squared-error function [7]

$$J(\mathbf{w}) = \sum_{n=1}^{N_X} \left[\mathbf{w}^T \boldsymbol{\gamma}(\mathbf{x}_n) - t\right]^2 + \sum_{n=1}^{N_Z} \left[\mathbf{w}^T \boldsymbol{\gamma}(\mathbf{z}_n) - 0\right]^2, \qquad (5)$$

where the target language data $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X}\}$ is given a desired output value of $t = (N_X + N_Z)/N_X$, and the competing languages data $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_Z}\}$ a desired output value of 0. Taking the derivative of (5) and set it equal to zero, we obtain the following MSE solution

$$\mathbf{w}_{\text{tgt}} = \left[\frac{\boldsymbol{\Gamma}^T \boldsymbol{\Gamma}}{(N_X + N_Z)}\right]^{-1} \left[\frac{1}{N_X} \sum_{n=1}^{N_X} \boldsymbol{\gamma}(\mathbf{x}_n)\right], \qquad (6)$$

where

$$\boldsymbol{\Gamma} \equiv \left[\boldsymbol{\gamma}(\mathbf{x}_1), \ldots, \boldsymbol{\gamma}(\mathbf{x}_{N_X}), \boldsymbol{\gamma}(\mathbf{z}_1), \ldots, \boldsymbol{\gamma}(\mathbf{z}_{N_Z})\right]^T \qquad (7)$$

is the $(N_X + N_Z) \times KM$ data matrix with each row represents the activations of the KM basis functions in response to a given feature vector.

The optimum weights $\mathbf{w}_{\text{tgt}}$ are now fully determined by the training data $\{X, Z\}$ and the basis functions $\boldsymbol{\gamma}$. Given a test sequence $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_Y}\}$, the output of the network averaged over the entire sequence is given by

$$f(Y) = \frac{1}{N_Y} \sum_{n=1}^{N_Y} \mathbf{w}_{\text{tgt}}^T \boldsymbol{\gamma}(\mathbf{y}_n) = \mathbf{w}_{\text{tgt}}^T \left[\frac{1}{N_Y} \sum_{n=1}^{N_Y} \boldsymbol{\gamma}(\mathbf{y}_n)\right]. \qquad (8)$$

Substituting (6) in (8), the output f can be expressed as a function of two sequences, X and Y, as follows

$$f(X, Y) = \boldsymbol{\rho}_X^T \boldsymbol{\rho}_Y, \qquad (9)$$

where

$$\boldsymbol{\rho}_X = \boldsymbol{\Lambda}^{-1/2}\left[\frac{1}{N_X}\sum_{n=1}^{N_X}\boldsymbol{\gamma}(\mathbf{x}_n)\right] \text{ and } \boldsymbol{\rho}_Y = \boldsymbol{\Lambda}^{-1/2}\left[\frac{1}{N_Y}\sum_{n=1}^{N_Y}\boldsymbol{\gamma}(\mathbf{y}_n)\right] \quad (10)$$

are referred to as the *characteristic vectors* for the observation sequences X and Y, respectively. The matrix $\boldsymbol{\Lambda}$ is a diagonal approximation of a correlation matrix in the following form

$$\boldsymbol{\Lambda} \approx diag\left[\frac{\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}}{(N_X + N_Z)}\right]. \quad (11)$$

This approximation assumes that the outputs of the basis functions γ_j are uncorrelated, and greatly simplifies the matrix inversion and multiplication operations in (10).

The elements of a characteristic vector collectively represent a histogram describing the occurrence statistics of speech sound units for a given speech utterance. Different languages would exhibit different patterns of histogram. This phonotactic information is used to discriminate a target language from other competing languages in our system.

3.3. Probabilistic sequence kernel (PSK) SVM

Equation (9) indicates that the generalized RBF network measures the similarity of a test sequence *Y* to a reference sequence *X* by computing the inner product of their characteristic vectors. Instead of simple inner product, an SVM can be used to define a hyperplane that separates $\boldsymbol{\rho}_X$ of the target language from those $\boldsymbol{\rho}_Z$ of the competing languages. Such a hyperplane is described by a set of support vectors in the following form:

$$g(\boldsymbol{\rho}_Y) = \sum_{l=1}^{L}\alpha_l t_l \boldsymbol{\rho}_l^T \boldsymbol{\rho}_Y + b, \quad (12)$$

where *L* denotes the number of support vectors $\boldsymbol{\rho}_l$, b is the bias, and the term $\alpha_l t_l$ indicates the weight of the support vector $\boldsymbol{\rho}_l$ in characterizing the hyperplane. Using (9) in (12), the sequence-comparing functionality pertaining to the RBF network can be written as part of the SVM, as follows

$$g(Y) = \sum_{l=1}^{L}\alpha_l t_l f(X_l, Y) + b, \quad (13)$$

where the function $f(X_l, Y)$ is now referred to as the probabilistic sequence kernel (PSK), the purpose of which is to transform variable length speech utterances into characteristic vectors (*KM*-dimension in this case) for SVM classification.

4. LANGUAGE RECOGNITION USING PSK-SVM

The probabilistic sequence kernel (PSK) SVM, as shown in Fig. 2, consists of two major elements, namely, (i) front-end set of Gaussian basis functions, (ii) SVM model for each target language. To construct the front-end, we first identify a set of basis languages and train a GMM for each of them. The basis GMMs are then pooled together to form the front-end bases as in (3). It should be mentioned that we select the basis languages according to the availability of training data, and may not necessarily need to include all the target languages to be recognized. This flexibility facilitates the system to be easily scaled-up for new target languages.

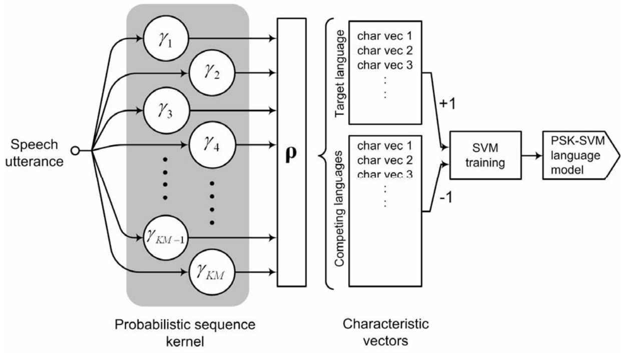

Fig. 2. Training of probabilistic sequence kernel (PSK) SVM for language recognition.

To construct the back-end classifier, we first compute the normalization matrix $\boldsymbol{\Lambda}$ using speech utterances from all the target languages, and then transform all the utterances into characteristic vectors according to (10) and (11). We use a *one* vs. *rest* strategy in training the SVM models. As shown in Fig. 2, the characteristic vectors are assigned with appropriate label (i.e., +1 for target language, -1 for competing languages) for SVM training [5]. Notice that all SVM models are trained using the same set of characteristic vectors (with different labels assigned), which greatly reduces the computations. Further computation reduction is achieved by (i) assuming diagonal covariance matrices $\boldsymbol{\Sigma}_j$ for the Gaussian densities, (ii) evaluating only the top scoring Gaussian densities. This fast technique for the sequence kernel computation is detailed in the next section.

5. EXPERIMENTS

We conduct the experiments on the NIST 1996, 2003 and 2005 language recognition evaluation (LRE) data for 30-seconds trials. The task of the evaluation is to detect the presence of a hypothesized target language given a recorded telephony speech. There are 12 target languages for the 1996 and 2003 tasks, and 8 for the 2005 task. The training data is drawn from the CallFriend corpus available from the Linguistic Data Consortium (LDC).

We compare four systems, namely, GLDS-SVM, GMM, PSK-SVM, and PPRLM. The first three systems use SDC with (7, 1, 3, 7) configuration [10], which results in SDC vectors of 49 dimension. We follow closely the system setup in [6, 10] for the GMM and GLDS-SVM. Briefly, each target language is modeled with two gender-dependent GMMs with 2048 mixtures. We use polynomial expansion up to third order for the GLDS-SVM. The PPRLM uses three phone recognizers (Czech, Hungarian, and Russian) based on long temporal context developed by Brno University of Technology [2]. We used similar setup as in [2], except that the output scores are merged from the individual PRLM subsystems by taking their average.

For the PSK-SVM system, we attain the front-end Gaussians from twelve basis languages, which are readily available from the CallFriend corpus. We first train an *initial*

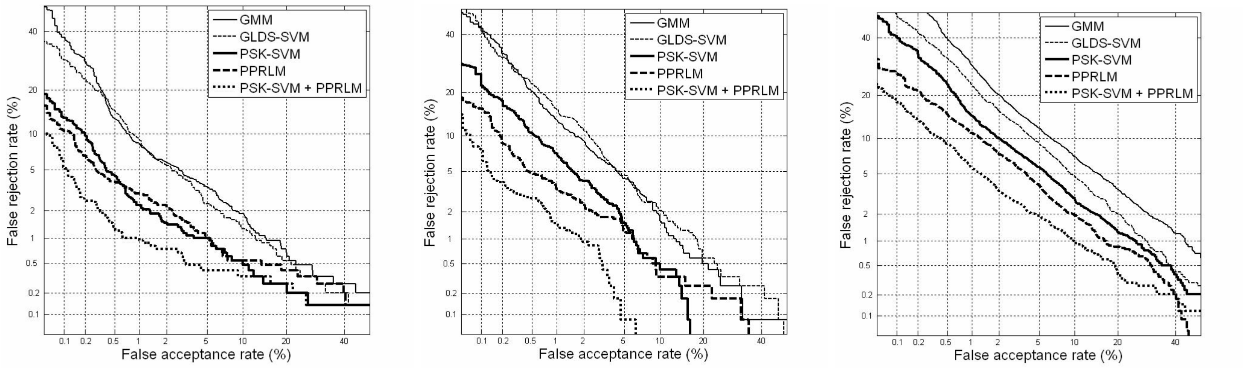

Fig. 3. Performance of four systems on the NIST 1996 (left), 2003 (middle) and 2005 (right) LRE for 30-seconds trials.

GMM with 2048 mixtures using a small fraction of data. The parameters of this GMM are then used to initialize the expectation-maximization (EM) algorithm in adapting to the twelve basis GMMs, one at a time. Since the basis GMMs are initialized with the same parameters, there exist one-to-one correspondences among their Gaussian components. This property allows a fast computation procedure as follows: For each feature vector, the *initial* GMM is used to determine the top N Gaussians with higher likelihoods. Using this information on the basis GMMs, we evaluate the top N Gaussians, while the remaining Gaussians are assumed to have zero activation. We use $N = 100$ (approximately 5% of 2048) in the experiments. For this case, we achieve a faster computation with a factor of $(12\times2048)/(12\times100 + 2048) \approx 7$.

The performance of the four systems is listed in Table I in terms of equal error rate (EER). The detection error tradeoff (DET) curves are shown in Fig. 3. It is evident that the performance of the PSK-SVM is consistently better than the GMM and GLDS-SVM on 1996, 2003 and 2005 LRE for the same SDC coefficients used. From Fig. 3, it can be observed that the PSK-SVM exhibits competitive performance to that of the PPRLM, and even outperforms the PPRLM at certain decision thresholds. The plots also show the fusion of the PSK-SVM and PPRLM. We use simple weighted sum rule in the score fusion. The optimum weights are first selected on the 1996 LRE, and used for 2003 and 2005 LRE. The fused system work extremely well. It gives relative EER improvements of 42%, 41%, and 36% over the best system for 1996, 2003 and 2005 LRE, respectively.

6. CONCLUSIONS

We have proposed a novel generative front-end for characterizing speech utterances using an ensemble of Gaussian densities for language recognition. By exploiting the long-term speech dynamic of the SDC coefficients, the Gaussian densities represent underlying set of speech sounds characterizing different languages. Used as the bases in a sequence kernel, the Gaussian density functions map speech utterances into characteristic vectors for SVM classification. We have also introduced the concept of basis languages in constructing the generative front-end and a fast technique for sequence kernel computation. Language recognition experiments showed that the proposed method exhibits good performance consistently across the NIST 1996, 2003 and 2005 LRE tasks.

Table I: EER (%) performance of four systems on the NIST 1996, 2003 and 2005 LRE for 30 seconds test durations.

System	1996	2003	2005
GMM	4.00	4.61	8.37
GLDS-SVM	3.62	4.70	6.94
PSK-SVM	1.72	2.98	5.48
PPRLM	2.14	2.17	4.38
PSK-SVM + PPRLM	0.99	1.29	2.79

7. REFERENCES

[1] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31-44, Jan. 1996.

[2] P. Matějka, P. Schwarz, J. Černocký, and P. Chytil, “Phonotactic language identification using high quality phoneme recognition,” in *Proc. Eurospeech*, pp. 2237-2240, 2005.

[3] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification", *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 1, pp. 271-284, Jan. 2007.

[4] P. A Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J.R. Deller, Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in *Proc. ICSLP*, pp. 89-92, 2002.

[5] R. Collobert and S. Bengio, “SVMTorch: support vector machines for large-scale regression problems,” *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[6] 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.

[7] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1995.

[8] K. A. Lee, C. You, H. Li, and T. Kinnunen, “A GMM-based probabilistic sequence kernel for speaker recognition,” in *Proc. Interspeech*, pp. 294-297, 2007.

[9] T. F. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Upper-Sadder River, NJ: Prentice-Hall, 2002.

[10] E. Singer, P. A Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, “Acoustic, phonetic, and discriminative approaches to automatic language identification,” in *Proc. Eurospeech*, pp. 1345-1348, 2003.