Characterizing Speech Utterances for Speaker Verification with Sequence Kernel SVM

Kong-Aik Lee[1]*, Changhuai You*[1]*, Haizhou Li*[1]*, Tomi Kinnunen*[2]*, and Donglai Zhu*[1]

[1]Institute for Infocomm Research (I^{2}R),
Agency for Science, Technology and Research (A*STAR), Singapore
[2]Department of Computer Science and Statistics, University of Joensuu, Finland
kalee@i2r.a-star.edu.sg

Abstract

Support vector machine (SVM) equipped with sequence kernel has been proven to be a powerful technique for speaker verification. A number of sequence kernels have been recently proposed, each being motivated from different perspectives with diverse mathematical derivations. Analytical comparison of kernels becomes difficult. To facilitate such comparisons, we propose a generic structure showing how different levels of cues conveyed by speech utterances, ranging from low-level acoustic features to high-level speaker cues, are being characterized within a sequence kernel. We then identify the similarities and differences between the popular generalized linear discriminant sequence (GLDS) and GMM supervector kernels, as well as our own probabilistic sequence kernel (PSK). Furthermore, we enhance the PSK in terms of accuracy and computational complexity. The enhanced PSK gives competitive accuracy with the other two kernels. Fusing all the three kernels yields an EER of 4.83% on the 2006 NIST SRE core test.

Index Terms: speaker verification, characteristic vector, support vector machine, sequence kernel

1. Introduction

Modern speaker verification systems consist of two main components: feature extraction and speaker modeling. At the feature-extraction front-end, speech utterances are usually parameterized as short-term cepstral features. For speaker modeling, the classical approaches of Gaussian mixture modeling (GMM) [1, 2] and vector quantization (VQ) [3] are commonly used. In these *generative* approaches the training feature vectors are assumed to be drawn from a speaker-specific probability distribution and the training process consists of estimating the parameters of the underlying density function. To this end, *maximum a posteriori* (MAP) adaptation of a so-called *universal background model* (UBM) is commonly used [2].

Recent advances in speaker verification [4, 5, 6] have largely relied on the *discriminative* learning mechanism of support vector machines (SVMs) [7] for boosting classification accuracy. Rather than modeling within-class distributions of the target and background speakers, SVM aims at modeling the decision boundary between them. To this end, one practical issue is to represent variable-length speech utterances into suitable form for SVMs. A common approach is to transform speech utterances explicitly into fixed- and high-dimensional vectors via a so-called *sequence kernel* or *dynamic kernel* [4, 5, 6]. SVM classifiers then operate on these expanded vectors to make classification decision.

A number of sequence kernels have been proposed by different authors. The kernels differ in their feature expansion mechanism (basis function selection), feature transformations and normalizations, and in other kernel-specific steps. For instance, the popular *generalized linear discriminant sequence* (GLDS) kernel [4] performs feature expansion with monomial bases, followed by averaging and variance normalization of the expanded vectors. Other kernels include GMM supervector kernel [5], probabilistic sequence kernel (PSK) [6, 8], Fisher kernel [9], MLLR supervector kernel [10], and incomplete Cholesky kernel [11] just to mention a few. Rank normalization [12] and within-class covariance normalization [13] are examples of kernel normalizations.

The abovementioned sequence kernels were motivated and derived from different perspectives in their own right. Due to the inherent differences of the kernels, an analytical comparison between their underlying mechanisms is always a difficult task. In this paper, our primary aim is to identify and compare the typical steps encapsulated within the sequence kernel SVMs. In particular, we show how the GLDS kernel [4] and GMM supervector kernel [5], as well as our own PSK kernel [6] can be reformulated such that they conform to a generic structure in characterizing speech utterances for text-independent speaker verification. Having the kernels analyzed under a common platform makes it easier to identify their affinities and differences, and to discover the reasons *why* certain kernels perform better than others.

In addition, we present some recent improvements to our PSK kernel [6]. We improve on the basis function selection, which brings along a fast computation technique in evaluating the kernel as well as better classification accuracy compared to our previous proposal. We also equip our kernel with the *nuisance attribute projection* (NAP) technique [14], which has been proven to be successful in dealing with channel variability.

2. Characterizing Speech Utterances

Speech utterances can be represented in various forms with different levels of compactness targeted for a specific application. Figure 1 illustrates the idea of characterizing speech utterances for text-independent speaker verification. At the first level, the acoustic speech signal is a measure of the changes in acoustic sound pressure level due to the movements of articulators (tongue, lips, etc.). At the next level of the structure are the cepstral feature vectors. They are usually computed either via mel-frequency filterbank, leading to *mel-frequency cepstral coefficients* (MFCCs), or via linear prediction, yielding the *linear predictive cepstral coefficients* (LPCCs) [15]. The cepstral feature vectors give a compact representation of the short-term spectra correlating with the changes in the vocal tract shape over time.

The cepstral features convey both speaker-specific cues as well as linguistic information (the words). In speaker verification, our intention is to suppress other factors influencing the cepstral features, leaving only the speaker-

dependent component [1]. As shown in Fig. 1, we use a set of *basis functions* to analyze and summarize the speaker-specific cues into a compact vector $\boldsymbol{\rho}$ dubbed as the *characteristic vector* of the speaker. Representing the utterances as fixed-dimensional vectors rather than variable length vector sequences greatly simplifies the subsequent speaker modeling and matching tasks.

Let $\varphi_j(\mathbf{x})$ for j = 1, 2,…, M, be the set of M basis functions. A sequence $X = \{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ of N cepstral feature vectors of dimension D is converted into a characteristic vector $\boldsymbol{\rho}$ via the following two operations:

(i) Feature expansion

$$\mathbf{x} \mapsto \boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \ldots, \varphi_M(\mathbf{x})]^T, \quad (1)$$

(ii) Statistical analysis

$$\boldsymbol{\rho} = f\{\boldsymbol{\varphi}(\mathbf{x}); \mathbf{x} \in X\}, \quad (2)$$

where the superscript T denotes matrix transposition and $f\{\cdot\}$ represents some functions of the expanded features $\boldsymbol{\varphi}(\mathbf{x})$, such as the arithmetic mean. With these two operations, the number of feature coefficients is significantly reduced from the original $N \times D$ coefficients of X down to the M coefficients of $\boldsymbol{\rho}$, since $N \gg M$ for long utterance. The characteristic vector therefore gives a compact representation of the vocal characteristic of the speaker provided that the basis functions are properly defined.

Furthermore, if the dimensionality of the characteristic vector can be made sufficiently high, i.e., $M \gg D$, the speaker verification problem would be more likely to be linearly separable one according to the Cover's theorem on separability of patterns [16, pp. 257]. Linear classifiers like SVM can then be used for discriminating between speakers based on their characteristic vectors. Since the expansion (1) is explicit, we can also be sure that the Mercer's condition [7] is automatically satisfied in the characteristic feature space and admissible for used with a linear-kernel SVM, as follows:

$$g(\boldsymbol{\rho}) = \sum_{l=1}^{L} \alpha_l t_l \boldsymbol{\rho}_l^T \boldsymbol{\rho} + \beta. \quad (3)$$

Here, $\{\boldsymbol{\rho}_l\}$, l = 1, 2 ,…, L are the L support vectors, β is the bias, and the term $\alpha_l t_l$ indicates the weight of the lth support vector. As depicted in Fig. 1, we usually need to normalize the characteristic vectors prior to SVM training and classification. Feature normalization is crucial since SVM is not invariant to feature scaling [7].

3. Basis Function Selection

A number of sequence kernels can be shown to follow the generic structure of Fig. 1 with the major differences being the types of basis functions used. In the following, we show how the GLDS kernel [4] and GMM supervector kernel [5] can be rewritten to coincide with (1) and (2) so that their common attributes can be identified.

3.1. GLDS Kernel

In the GLDS kernel [4], the expansion (1) of a cepstral feature vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$ is defined using the monomials of its elements. For instance, the feature expansion consisting of the monomials up to the second order is

$$\boldsymbol{\varphi}_{\text{GLDS}}(\mathbf{x}) = \left[1, x_1, x_2, \ldots, x_D, x_1^2, x_1x_2, \ldots, x_1x_D, x_2^2, \ldots, x_2x_D, \ldots, x_D^2\right]^T. \quad (4)$$

The sequence $X = \{\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_N\}$ is then represented as a characteristic vector by taking the arithmetic mean of the expansions:

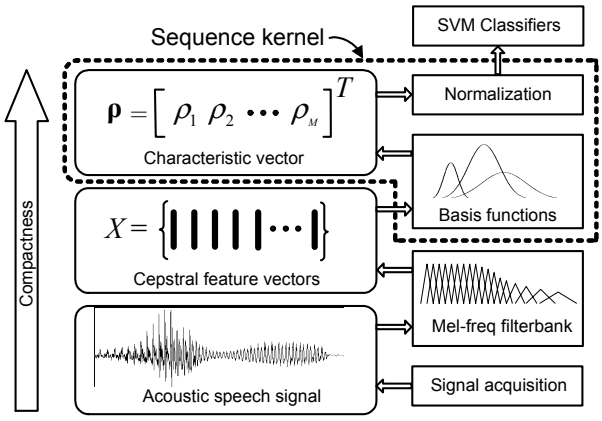

Figure 1*: A generic structure for characterizing speech utterances for speaker verification. The representation improves into a more compact format when we move from the bottom to the top of the structure.*

$$\boldsymbol{\rho}_{\text{GLDS}} = \frac{1}{N} \sum_{\mathbf{x} \in X} \boldsymbol{\varphi}_{\text{GLDS}}(\mathbf{x}). \quad (5)$$

The averaging operation yields the estimates of the first- and second-order moments: $E\{x_i\}$, $E\{x_i x_j\}$ for $i, j = 1,2,\ldots,D$, where E denotes the statistical expectation. The characteristic vector in (5) therefore captures the sample means and correlations (higher-order statistics are obtained for orders above two) of cepstral features.

The characteristic vector $\boldsymbol{\rho}_{\text{GLDS}}$ has dimensionality M = ((*D*+*k*)!)/(*D*!*k*!), where k is the maximum order of the monomials. In practice, only monomials up to the third order have shown to be useful for speaker and language recognition purposes [4]. The dimensionality becomes unfeasible for k > 3. The practical virtue of the monomial bases is their simplicity in terms of (i) computation, and (ii) no training is required unlike in [5, 6]. This is beneficial when there is a limited amount of data available for training. However, basis functions that exploit the underlying acoustic structure learned from a sufficient amount of training data would generally exhibit better performance, as we shall demonstrate experimentally in Section 5.

3.2. GMM Supervector Kernel

The GMM supervector [5] is formed by stacking the mean vectors of a GMM adapted from the UBM with the MAP criterion [2]. Denoting the number of Gaussians in the UBM by M, we define M vector-valued basis functions as

$$\boldsymbol{\varphi}_j(\mathbf{x}) = \left\{\frac{p(j \mid \mathbf{x})}{n_j}\right\}\mathbf{x} \quad \text{for } j = 1,2,\ldots,M, \quad (6)$$

where

$$p(j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid j)P(j)}{\sum_{i=1}^{M} p(\mathbf{x} \mid i)P(i)} \quad (7)$$

denotes the posterior probability of the jth Gaussian component $p(\mathbf{x} \mid j) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ of the UBM. The probabilistic count

$$n_j = \sum_{\mathbf{x} \in X} p(j \mid \mathbf{x}) \quad (8)$$

at the denominator in (6) is determined for the jth Gaussian component by evaluating the total occupancy for the sequence X into that component. Stacking together the M vector-valued bases $\boldsymbol{\varphi}_j(\mathbf{x})$, we form the MD-dimensional expansion as

$$\boldsymbol{\varphi}_{\mathrm{SV}}(\mathbf{x}) = \left[\boldsymbol{\varphi}_1^T(\mathbf{x}), \boldsymbol{\varphi}_2^T(\mathbf{x}), \ldots, \boldsymbol{\varphi}_M^T(\mathbf{x})\right]^T . \quad (9)$$

Finally, the GMM supervector (i.e., the characteristic vector) is obtained by taking the *sum* of the expansions, followed by regularization with *a priori* information (i.e., the supervector $\boldsymbol{\mu}$ of the UBM), as follows:

$$\boldsymbol{\rho}_{\mathrm{SV}} = \boldsymbol{\Lambda}\left[\sum_{\mathbf{x}\in X} \boldsymbol{\varphi}_{\mathrm{SV}}(\mathbf{x})\right] + (I - \boldsymbol{\Lambda})\boldsymbol{\mu} . \quad (10)$$

Here, $\boldsymbol{\Lambda}$ is a diagonal matrix controlling the relative contributions of the *a posteriori* and *a priori* information. The characteristic vector $\boldsymbol{\rho}_{\mathrm{SV}}$ therefore captures the speaker information based on the positions of the Gaussian components adapted from the UBM.

Equation (10) gives the MAP adaptation [2] of mean vectors in vector-matrix notation. The control parameters at the diagonal of $\boldsymbol{\Lambda}$ are set according to the occupancy n_j and the desired rate of adaptation [2]. Letting $\mathbf{1}$ denote a $1\times D$ vector of all ones, the control matrix is given by

$$\boldsymbol{\Lambda} = \mathrm{diag}\{[\lambda_1\mathbf{1}, \lambda_2\mathbf{1}, \ldots, \lambda_M\mathbf{1}]\}, \quad (11)$$

where $\lambda_j = n_j/(n_j + r)$ and r is the relevance factor taken usually in the range 8 ~ 20, and $\mathrm{diag}\{\cdot\}$ denotes the operation of transforming its vector argument into a diagonal matrix.

4. Enhancing the Probabilistic Sequence Kernel (PSK)

The basis functions of the PSK kernel [6] are defined based on an ensemble of Gaussian densities, $p(\mathbf{x}\,|\,j) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $j = 1,2,\ldots,M$, which serves as a decoder in identifying the characteristic sound patterns representing speaker-specific cues. In particular, the expansion is given by

$$\boldsymbol{\varphi}_{\mathrm{PSK}}(\mathbf{x}) = [\,p(j=1\,|\,\mathbf{x}),\ p(j=2\,|\,\mathbf{x}),\ \ldots,\ p(j=M\,|\,\mathbf{x})]^T , \quad (12)$$

where $p(j\,|\,\mathbf{x})$ denotes the posterior probability of the jth Gaussian, as given by (7). Each element of the expansion $\boldsymbol{\varphi}_{\mathrm{PSK}}(\mathbf{x})$ gives the probability of occurrence of the jth acoustic class evaluated for a given feature vector $\mathbf{x}$. The average probabilistic count across the entire sequence X is given by

$$\boldsymbol{\rho}_{\mathrm{PSK}} = \frac{1}{N}\sum_{\mathbf{x}\in X} \boldsymbol{\varphi}_{\mathrm{PSK}}(\mathbf{x}) . \quad (13)$$

The characteristic vector $\boldsymbol{\rho}_{\mathrm{PSK}}$ can therefore be interpreted as an M-bin histogram indicating the probabilities of occurrence of various acoustic sound classes observed in the given speech utterance X.

The GMM supervector in (10) captures the speaker information based on the *positions* of Gaussian components, whereas the PSK expansion (12) hinges on the posterior probabilities of Gaussian components. These Gaussian components represent general vocal tract configurations in producing various speech sounds: the mean $\boldsymbol{\mu}_j$ represents the average spectral shape of the jth acoustic classes, the covariance matrix $\boldsymbol{\Sigma}_j$ represents the variations of the average spectral shape, and the weight $P(j)$ represents the probability of occurrence of the jth sound class.

In our initial study [6], we established the Gaussian bases by aggregating the UBM and an adapted GMM into an ensemble of $2M$ Gaussians. Therefore, each target speaker has its own set of bases. The major drawback of this approach is high computational load when we scale up the system to include more target speakers; the expansion of a test utterance has to be performed repeatedly for each speaker using different set of bases. This is undesirable in practice, where we usually need to score multiple speaker's models for a given test utterance.

In this paper, we use a common set of bases for all speakers instead of the speaker-dependent kernel used in [6]. This is achieved by aggregating the GMMs of K speakers selected from a background dataset. Each GMM contributes Q components, resulting in $M = KQ$ Gaussian bases. We first train a root GMM with Q mixtures from the background data using the expectation-maximization (EM) algorithm [1]. We then train a GMM for all the speakers in the background using the root GMM as initial models. In our implementation, we use five iterations of EM in training the root GMM and another five iterations of EM for the background speakers. We then select a subset of K speakers that gives the largest scattering measure [17] from the background. The scattering measure is defined as the average distance between all pairs of GMMs in a subset. The more scattered the speakers, the resulting KQ Gaussian bases would therefore cover a richer set of acoustic sound classes.

The background speakers' GMMs are pooled together with equal weights to form an ensemble of $M = KQ$ Gaussians. An identical set of bases can then be used for all target speakers. It should be mentioned that the purpose of the root GMM is to allow a fast computation procedure in evaluating the expansion (12). For an input vector, the root GMM is first used to determine the top S Gaussians with higher likelihoods (we found that $S = 10$ is sufficient for $Q = 256$). Using this information, we evaluate only the top S Gaussians in each of the background GMMs, while the remaining Gaussians are assumed to have zero probabilities. We have successfully used similar techniques for spoken language recognition in [8].

5. Experiments

We evaluated the performance of the three characteristic vector representations described earlier for (i) GLDS kernel [4], (ii) GMM supervector [5], and (iii) probabilistic sequence kernel (PSK) [6]. The experiments were carried out on the 2006 NIST SRE [18]. The core test consists of 3,612 genuine and 47,836 imposter trials. There are 810 target speakers each enrolled with one side of a 5 minutes conversation. Our background data includes over 3000 speech utterances of 2.5 minutes duration from 310 speakers drawn from the 2004 NIST SRE data. All speech samples are first pre-processed to remove silence and converted into sequences of 36-dimensional MFCCs (with deltas and double deltas) and finally represented as characteristic vectors.

For the GLDS kernel, we used all monomials up to third order. For the GMM supervector, the UBM consists of 512 mixtures. For PSK, we formed the Gaussian bases by aggregating $K = 72$ GMMs selected from the background, each with $Q = 256$ mixtures. The characteristic vectors are normalized with different schemes before they can be used with SVM. We use variance normalization for GLDS kernel, Kullback-Leibler divergence normalization [5] for GMM supervector, and rank normalization [12] for PSK.

The value $K = 72$ was selected for the PSK considering that there are 310 speakers in the background dataset, and by having $Q = 256$, the characteristic vectors representation for PSK has the same dimensionality ($72\times256 = 18{,}432$) with the the GMM supervector ($512\times36 = 18{,}432$). This demonstrates the flexibility of the PSK expansion in controlling the capacity of the SVM classifier. The capacity of the GLDS kernel, on the other hand, is hardly controllable since the dimensionality of the kernel grows exponentially with increased monomial order. Its performance (not shown in this paper) degrades when we increased the order beyond three.

Table 1. *EER of the PSK system (with and without NAP) evaluated on the 2006 NIST SRE core test.*

System	EER (%)		
	Male	Female	All
PSK	7.15	9.48	8.50
PSK + NAP	4.76	6.31	5.95

Table 2. *EER and Min DCF of (i) GLDS kernel, (ii) GMM supervector, (iii) PSK, and their fusion evaluated on the 2006 NIST SRE core test.*

System	EER (%)	Min DCF (×100)
GLDS + NAP	6.38	3.18
GMM supervector + NAP	5.62	2.68
PSK + NAP	5.95	2.76
Fuse all	4.83	2.53

We investigated the performance of the PSK system with and without *nuisance attribute projection* (NAP) [14]. The NAP projection matrix has a rank of 40 and was derived from the 2004 NIST SRE data. The equal error rates (EERs) for all trials and for both genders are shown in Table 1 (Note that NIST SRE tasks do not include cross-gender trials). A relative EER improvement of 30% is obtained by introducing NAP into the PSK. Though the EERs are higher for female trials, the relative improvements introduced by NAP are almost similar for both genders.

The performances of the three systems are shown in Table 2 in terms EER and minimum detection cost function (Min DCF) [18]. The detection error tradeoff (DET) curves are plotted in Fig. 2. The PSK and GMM supervector kernels systematically outperform the GLDS kernel. This observation confirms our claim that better characterization of speaker information is obtained by incorporating additional knowledge regarding the underlying acoustic structure into the bases as in the PSK and GMM supervector. Finally, we fused the three system using equal weights considering that their scores are approximately in the same range. The fused system gives relative EER improvement of 14% over the best system, suggesting that the different kernels capture complementary aspects of the same feature space.

6. Conclusions

We have shown a generic structure for characterizing speech utterances for speaker verification with two major elements: (i) cepstral feature extraction followed by (ii) speaker information extraction via basis functions. The effectiveness of the characterization depends on the ability of the basis functions in capturing speaker-specific information. Based on the proposed structure, we systematically analyzed the GLDS kernel and GMM supervector kernel, and revamped our previously proposed probabilistic sequence kernel (PSK) for improved accuracy and reduced computational complexity. Experiments on the 2006 NIST speaker detection task showed that the enhanced PSK exhibits good performance and it fuses nicely with the GLDS and GMM supervector.

7. References

[1] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, Jan. 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

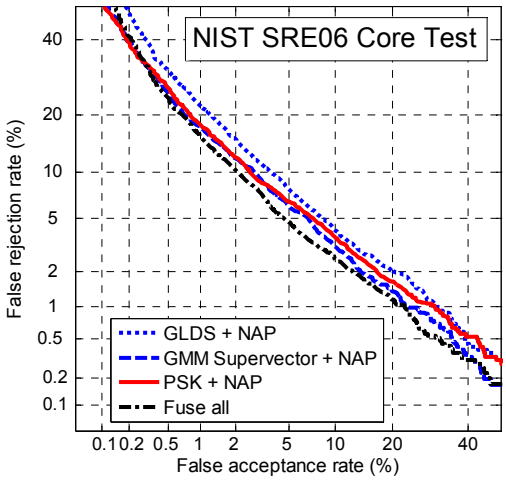

Figure 2: *Detection error tradeoff (DET) curves of the three speaker verification systems and their fusion evaluated on the 2006 NIST SRE core test.*

[3] V. Hautamäki, T. Kinnunen, I. Kärkkäinen, M. Tuononen, J. Saastamoinen, P. Fränti, "Maximum *a Posteriori* Estimation of the Centroid Model for Speaker Verification", *IEEE Signal Processing Lett.*, vol. 15, pp. 162-165, 2008.

[4] 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.

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

[6] 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.

[7] N. Cristianini and J. Shawe-Taylor, An *Introduction to Support Vector Machines*. Cambridge: Cambridge University Press, 2000.

[8] K. A. Lee, C. You, and H. Li, "Spoken language recognition using support vector machines with generative front-end," in *Proc. IEEE ICASSP*, 2008, pp. 4153-4156.

[9] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 203-210, Mar. 2005.

[10] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 1987-1998, Sep. 2007.

[11] J. Louradour, K. Daoudi, F. Bach, "SVM speaker verification using an incomplete Cholesky decomposition sequence kernel," in *Proc. IEEE Odyssey*, 2006, pp. 1-5.

[12] E. Shriberg, L. Ferrer, A. Venkataraman, and A. Kajarekar, "SVM modeling of SNERF-Grams for speaker recognition," in *Proc. ICSLP*, 2004, pp 1409-1412.

[13] A. O. Hatch, S. Kajarekar, A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Interspeech*, 2006, pp. 1471-1474.

[14] A. Solomonof, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE ICASSP*, 2005, pp. 629-632.

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

[16] S. Haykin, *Neural Network*: *A Comprehensive Foundation*. NJ: Prentice-Hall, 1999.

[17] Y. Mami and D. Charlet, "Speaker recognition by location in the space of reference speakers," Speech communication, vol. 48, no. 2, pp. 127-141, 2006.

[18] *The NIST Year 2006 Speaker Recognition Evaluation Plan*, National Institute of Standards and Technology, Mar. 2006.