Shifted-Delta MLP Features for Spoken Language Recognition

Article *in* IEEE Signal Processing Letters · January 2013
DOI: 10.1109/LSP.2012.2227312

5 authors, including:

Haipeng Wang
The Chinese University of Hong Kong
13 PUBLICATIONS **229** CITATIONS
SEE PROFILE

Cheung-Chi Leung
Institute for Infocomm Research
68 PUBLICATIONS **558** CITATIONS
SEE PROFILE

Bin Ma
Institute for Infocomm Research
213 PUBLICATIONS **1,919** CITATIONS
SEE PROFILE

Haizhou Li
National University of Singapore
624 PUBLICATIONS **6,972** CITATIONS
SEE PROFILE

Shifted-Delta MLP Features for Spoken Language Recognition

Haipeng Wang, *Student Member, IEEE*, Cheung-Chi Leung, *Member, IEEE*, Tan Lee, *Member, IEEE*, Bin Ma, *Senior Member, IEEE*, and Haizhou Li, *Senior Member, IEEE*

***Abstract*—This letter presents our study of applying phoneme posterior features for spoken language recognition (SLR). In our work, phoneme posterior features are estimated from a multilayer perceptron (MLP) based phoneme recognizer, and are further processed through transformations including taking logarithm, PCA transformation, and appending shifted delta coefficients. The resulting shifted-delta MLP (SDMLP) features show similar distribution as conventional shifted-delta cepstral (SDC) features, and are more robust compared to the SDC features. Experiments on the NIST LRE2005 dataset show that the SDMLP features fit well with the state-of-the-art GMM-based SLR systems, and SDMLP features outperform SDC features significantly.**

***Index Terms*—Feature robustness, shifted-delta cepstral features, shifted-delta MLP features, spoken language recognition.**

I. Introduction

SPOKEN language recognition (SLR) refers to the process of automatically determining the language identity given a speech utterance. A variety of speech characteristics are useful for SLR tasks, ranging from high-level word syntactic information to low-level acoustic signal properties [1]. There have been two representative approaches to SLR, which are based on phonotactic features and spectral features respectively. Phonotactic features are derived from the outputs of phone recognizers and usually modeled by N-gram model or vector space model [2]. Spectral features are captured directly from acoustic signals and modeled typically by Gaussian mixture models (GMM) [3].

The most commonly used spectral features in SLR systems are the shifted-delta cepstral (SDC) features, which have been justified by extensive experiments [3]–[5]. Delta coefficients generally refer to the time derivatives of static coefficients of successive frames. Shifted-delta coefficients are the delta coefficients computed over several consecutive blocks of frames, which involve a relatively longer time span. It is believed that the long-time temporal information plays an important role in capturing language-specific spectral properties. However, like standard cepstral features, the SDC features are highly sensitive to speaker and environmental variations. In [6], it was proposed to transform cepstral-domain features into posterior probability features using a multi-layer perceptron (MLP). The MLP features have been found to be more robust for speech recognition than cepstral-domain features [7].

In this study, we investigate the application of shifted-delta MLP (SDMLP) features for SLR tasks. The SDMLP features are obtained by applying shifted-delta operation to the MLP features produced by an MLP-based phone recognizer. The resulted features can be incorporated straightforwardly into the state-of-the-art GMM-based SLR systems in the same way as the SDC features. It is expected that the proposed features leverage the robustness of MLP features and the benefit of long time span of shifted-delta operation. Moreover, the SDMLP features contain information about temporal variation of the MLP features, which can be interpreted as the evolution within a phoneme or the transition between connecting phonemes. Such phoneme contextual information of speech contributes greatly to SLR [2]. Our work is different from the previous work [8], [9] on MLP-based features for SLR tasks. We do not directly model the posterior features, but make transformations to derive the SDMLP features, which can be easily used by the existing GMM-based SLR systems.

The performance of the proposed SDMLP features was evaluated using the 2005 NIST Language Recognition Evaluation (LRE) dataset. Experimental results show that the SDMLP features outperform the SDC features significantly, and provide complementary benefits to existing acoustic and phonotactic approaches in SLR.

II. Shifted-Delta MLP Feature

A. Feature Extraction Procedure

Fig. 1 illustrates the proposed process of SDMLP feature extraction. An MLP phoneme recognizer is used to take in acoustic features from input speech and classify the respective frames into different phonemes. The output of the MLP can be interpreted as a measure of phoneme posterior probability at the frame level. The MLP posterior features are highly sparse in nature. They are transformed into logarithmic scale such that the resulted Log-MLP features are more suitable for Gaussian modeling. Subsequently principal component analysis (PCA) is applied to the Log-MLP features for decorrelation and feature dimension reduction. Lastly the shifted-delta operation is performed on the PCA outputs.

We follow the standard procedure of shifted-delta operation for SLR applications [4]. Let $\boldsymbol{o}(t)$ be the static feature vector

Manuscript received May 31, 2012; revised October 12, 2012; accepted October 24, 2012. Date of current version November 15, 2012. This work was supported in part by the General Research Funds (Refs. 414010 and 413811) from the Hong Kong Research Grants Council. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Constantine L. Kotropoulos.

H. Wang and T. Lee are with Department of Electrical Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: hpwang@ee.cuhk.edu.hk; tanlee@ee.cuhk.edu.hk).

C.-C. Leung, B. Ma, and H. Li are with the Institute for Infocomm Research, A*STAR, Singapore (e-mail: ccleung@i2r.a-star.edu.sg; mabin@i2r.a-star.edu.sg; hli@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2012.2227312

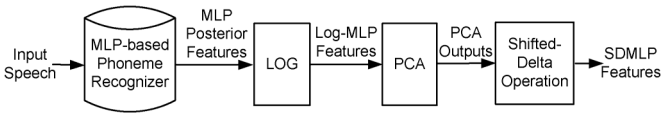

Fig. 1. Block diagram of SDMLP feature extraction. Post-processing like VAD or/and feature transformations can be applied to the SDMLP features.

that is composed of N feature elements derived at frame t. The delta feature vector at frame t is computed as

$$\Delta(t) = \boldsymbol{o}(t+d) - \boldsymbol{o}(t-d). \tag{1}$$

which covers a block of $2d+1$ frames. The shifted-delta feature vector $\widetilde{\boldsymbol{\Delta}}(t)$ is obtained by concatenating the delta features from k consecutive blocks, i.e.,

$$\widetilde{\boldsymbol{\Delta}}(t) = \left[\Delta(t)^T \quad \Delta(t+P)^T \quad \ldots \quad \Delta\left(t+(k-1)P\right)^T\right]^T, \tag{2}$$

where P is the time shift (in frames) between two neighboring blocks. After $\widetilde{\boldsymbol{\Delta}}(t)$ is computed, it is appended to the static feature vector $\boldsymbol{o}(t)$ to form the final feature vector, which is of size $(k+1)\times N$. The parameter setting for the shifted-delta computation is summarized as $N-d-P-k$. Among the four parameters, $d-P-K$ define the whole temporal span involved in the shifted-delta computation. In the standard SDC feature extraction, N is the dimension of static cepstral feature vector. For SDMLP feature extraction, N is the number of the principal components used in the PCA transformation.

In this study, the SDC features are derived from Mel-frequency cepstral coefficients (MFCC) by setting $N-d-P-k$ to 7-1-3-7. This is the standard setting validated by extensive experiments [3]. The dimension of SDC features is 56. For SDMLP feature extraction, N is set to 30 in the feature analysis and evaluated as a variable in the experiments. $d-P-k$ are still set to 1-3-7. The MLP posterior features are derived from a Hungarian phoneme recognizer [10] developed by Brno University of Technology (BUT). The phoneme recognizer was trained using the Hungarian part of the SpeechDAT-E corpus to recognize 61 phonemes. It uses the split temporal context network structure [11], which splits temporal vectors of critical bank energies into left-context and right-context parts, builds an MLP for each part, and uses a merging MLP to deliver the posterior probabilities. For this phoneme recognizer, the input involves a 310 ms long temporal context.

B. Feature Analysis

We carried out quantitative analysis of the statistical distribution and the robustness of different features. The purpose of analyzing the feature distribution is to see whether the SDMLP features can be used in the existing GMM-based SLR systems. The average skewness and the average kurtosis [12] are used to describe the feature distributions. For a random variable x, the skewness is defined as

$$\mathrm{Skew}(x) = \mathrm{E}\left[\left(\frac{x-\bar{x}}{\sigma}\right)^3\right], \tag{3}$$

TABLE I
AVERAGE SKEWNESS AND AVERAGE KURTOSIS ON TIMIT

Feature	SDC	MLP	Log-MLP	PCA outputs	SDMLP
Avg. Skew	-1.10	5.18	-1.45	-1.39	-1.32
Avg. Kurt	1.55	85.64	-0.23	0.086	0.43

where $\bar{x}$ and δ denote the mean and standard deviation of x. The skewness measures the degree of symmetry of the distribution. The Kurtosis of x is defined as

$$\mathrm{Kurt}(x) = \mathrm{E}\left[\left(\frac{x-\bar{x}}{\sigma}\right)^4\right] - 3. \tag{4}$$

It is a measure of the peakedness of the distribution. For a normal distribution, both skewness and kurtosis are zero. The average skewness and average kurtosis are the means of the skewness values and kurtosis values over all feature elements.

Table I shows the average skewness and kurtosis of different features derived from the TIMIT corpus. TIMIT, as a small-size dataset with rich phonetic and speaker variability, is suitable for quick analysis and obtaining detailed insights into the feature distribution. As can be seen, the average skewness and kurtosis of the MLP posterior features are quite different from those of the SDC features. This is due to sharp peakedness and high asymmetry of the sparse distribution of the MLP posterior features. Relatively, the distribution of the SDMLP features is much more similar to the SDC features, and the average kurtosis of the SDMLP features is even closer to zero than that of the SDC features. This indicates that the distribution of the SDMLP features fits well with the existing GMM-based SLR systems.

We compare the feature robustness against speaker variation using the SA1 and SA2 utterances in the TIMIT corpus. All the SA1 utterances contain the same spoken content produced by different speakers, and so do the SA2 utterances. Let V_{spk} denote the quantitative measure of the acoustic variation within the SA1 utterances and within the SA2 utterances, and let V_{total} denote the similar measure for the combined set of SA1 and SA2 utterances. V_{spk} is mainly due to speaker variation, while V_{total} involves both speaker and linguistic variations. We use the ratio $V_{\mathrm{spk}}/V_{\mathrm{total}}$ to indicate the degree of feature robustness. The smaller the ratio, the more robust the feature.

In our work, V_{spk} and V_{total} are computed based on the i-vector representation [13]. Given an utterance, let $\boldsymbol{M}$ denote the GMM supervector which is created by stacking all the mean vectors of the adapted GMM of this utterance. It is assumed the variability of the GMM supervectors is captured by a single space $\boldsymbol{T}$ which contains both speaker variability and linguistic variability. Then $\boldsymbol{M}$ can be modeled as

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}, \tag{5}$$

where $\boldsymbol{m}$ is the universal GMM supervector, and the coordinate vector $\boldsymbol{w}$ is the i-vector. Note that in this work, i-vector is used only in feature analysis, but not in our SLR systems. With the i-vector representation, the speaker variation V_{spk} is

$$V_{\mathrm{spk}} = \sum_{i\in\{SA1\}} \|\boldsymbol{w}_i - \bar{\boldsymbol{w}}_1\|_2^2 + \sum_{j\in\{SA2\}} \|\boldsymbol{w}_j - \bar{\boldsymbol{w}}_2\|_2^2, \tag{6}$$

where $\{SA1\}$ and $\{SA2\}$ are the index sets of SA1 and SA2 utterances respectively, and $\boldsymbol{w}_i$ is the i-vector for the i_{th} utter-

TABLE II
$V_{\text{spk}}/V_{\text{total}}$ OF TIMIT SA1 AND SA2 UTTERANCES

i-vector dimension	5	10	15	20	25
SDC	0.51	0.71	0.84	0.86	0.92
Log-MLP	0.33	0.54	0.66	0.72	0.76
PCA outputs	0.31	0.49	0.57	0.65	0.69
SDMLP	0.39	0.62	0.71	0.78	0.86

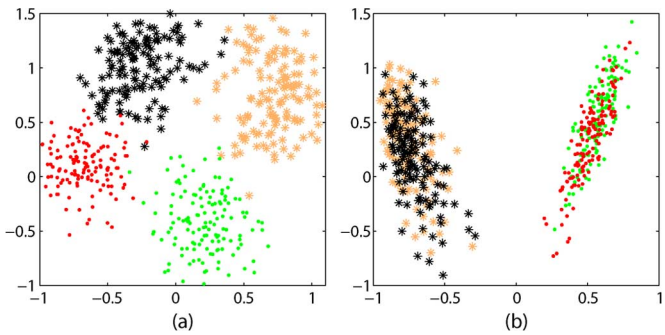

Fig. 2. Two-dimensional i-vector representations of TIMIT utterances. Black $*$ = female SA1 utterances, red $\cdot$ = female SA2 utterances, brown $*$ = male SA1 utterances, green $\cdot$ = male SA2 utterances. (a) SDC; (b) SDMLP.

ance. $\bar{\boldsymbol{w}}_1$ and $\bar{\boldsymbol{w}}_2$ are the mean vectors of SA1 and SA2 i-vectors respectively. The total variation V_{total} is specified by

$$V_{\text{total}} = \sum_{i \in \{SA1\}} \|\boldsymbol{w}_i - \bar{\boldsymbol{w}}\|_2^2 + \sum_{j \in \{SA2\}} \|\boldsymbol{w}_j - \bar{\boldsymbol{w}}\|_2^2, \quad (7)$$

where $\bar{\boldsymbol{w}}$ is the mean vector of all the SA1 and SA2 i-vectors.

Table II shows the $V_{\text{spk}}/V_{\text{total}}$ values when different i-vector dimensions are used. The proportion of speaker variation of MLP-based features are consistently smaller than that of SDC features. In other words, the MLP-related features are more robust against speaker variation. Conceptually, the MLP phoneme posterior probabilities are speaker independent if the MLP training involves enough balanced speakers. In practice the Log-MLP features still exhibit some speaker variation, which is further reduced by the PCA transformation. The SDMLP features have larger $V_{\text{spk}}/V_{\text{total}}$ values than the PCA outputs because the temporal dynamic information involves speaker variation, but this information is important for SLR. To visualize the data distribution, we represent TIMIT SA1 and SA2 utterances by two-dimensional i-vectors as shown in Fig. 2. It is clear that the speaker variation of the SDMLP features is significantly reduced compared to the SDC features.

III. LANGUAGE RECOGNITION EXPERIMENTS

A. System Implementation

The SDMLP features are used in the same way as the SDC features in SLR systems. In this work, we evaluated the performances with three acoustic SLR systems: GMM system, GMM-SVM system [14], and Model Pushing (MP) system [15]. A language-independent universal background model (UBM) with 1024 mixtures was first trained using all the training data. In the GMM system, language-specific GMM models were trained by MAP adaptation from the UBM. Both the GMM-SVM system and the MP system are based on the GMM mean supervectors. The GMM-SVM system used the Kullback-Leibler (KL) kernel and one-versus-rest training strategy. For the MP system, we first obtained the SVM language models using the GMM supervectors with the one-versus-rest strategy. Then the supporting vectors were normalized by the respective supporting vector weights, and transferred back into the standard GMM model [15].

TABLE III
PERFORMANCES VERSUS PCA RANK IN BASIC GMM SYSTEM WITHOUT fLFA AND WITHOUT SCORE BACKEND. RESULTS ARE GIVEN IN EER% ON 30S CONDITIONS OF DIFFERENT DATASETS

PCA rank	SDMLP			MLP+Δ+ΔΔ		
	LRE96	LRE03	LRE05	LRE96	LRE03	LRE05
7	3.80	5.54	9.90	5.47	7.19	12.50
10	3.06	4.68	9.21	3.75	5.51	10.96
20	2.15	3.69	**8.89**	2.80	4.34	9.90
30	2.05	**3.59**	9.03	2.80	**4.20**	**9.39**
40	**2.01**	4.71	9.24	2.78	4.51	9.68
50	2.15	4.47	9.26	**2.37**	4.58	9.98

B. Experimental Setup

We conducted SLR experiments on the NIST LRE2005 closed-set task. The test set covers 7 target languages (English, Hindi, Japanese, Korean, Mandarin, Spanish, and Tamil), and contains three nominal durations (3 seconds, 10 seconds, and 30 seconds). The training set was from the CallFriend corpus and the OGI multi-language corpus. The NIST LRE1996 corpus and the NIST LRE2003 corpus were used as the development set. The SDC and SDMLP features were extracted as introduced in Section II-A. Both features were processed through a voice activity detection (VAD) module, and normalized to zero mean and unit variance. Feature-domain latent factor analysis (fLFA) [17] was applied to compensate for the speaker and channel variations. The fLFA subspace was estimated on the training set and its rank was set to 40. The average equal error rate (EER) was the evaluation criterion.

C. Experimental Results

We first evaluated the performance of the SDMLP features as a function of the PCA rank, which is denoted by N in the $N-d-P-k$ parameter setting in the SDMLP feature extraction. The basic GMM system without fLFA and without score backend was used for this purpose. Corresponding results are shown in Table III. As the PCA rank goes from 7 to 50, the best overall performance on the development set was achieved when the PCA rank is equal to 30. Note that if the PCA rank is switched from 30 to 40, the performance on LRE1996 corpus becomes slightly better (1.95% relative EER reduction), but the performance on LRE2003 corpus degrades severely (31.2% relative EER degradation). So we chose 30 as the PCA rank in the following experiments, and thereby the dimension of the following SDMLP features would be 240.

Table III also shows the performances of $\text{MLP}+\Delta+\Delta\Delta$ features that are the PCA outputs augmented by the delta and acceleration vectors. We followed the standard way to compute the delta and acceleration vectors as commonly used for speech recognition. The superiority of the shifted delta vector for SLR can be observed on both development set and test set. We also tested the performances of PCA outputs without deltas when PCA rank was set to 30. The corresponding EERs were 3.69% for LRE96, 5.52% for LRE03, and 10.95% for LRE05, which were even worse than the performances of $\text{MLP}+\Delta+\Delta\Delta$ features. This demonstrated the importance of temporal dynamic information for SLR.

TABLE IV
EER% PERFORMANCES OF SDMLP FEATURES AND SDC FEATURES ON LRE05 DATASET. fLFA AND LDA+GAUSSIAN BACKEND ARE USED

Features	Systems	3s	10s	30s
SDC	GMM	25.1	17.4	13.6
	GMM-SVM	30.2	13.8	4.88
	MP	**22.6**	11.6	4.83
SDC+fLFA	GMM	23.2	14.4	10.4
	GMM-SVM	30.9	13.7	4.70
	MP	23.1	**11.1**	**4.48**
SDMLP	GMM	21.2	11.4	7.77
	GMM-SVM	30.5	12.0	3.87
	MP	18.3	8.18	3.67
SDMLP+fLFA	GMM	19.1	8.92	5.33
	GMM-SVM	32.2	12.5	3.69
	MP	**18.2**	**7.88**	**3.33**

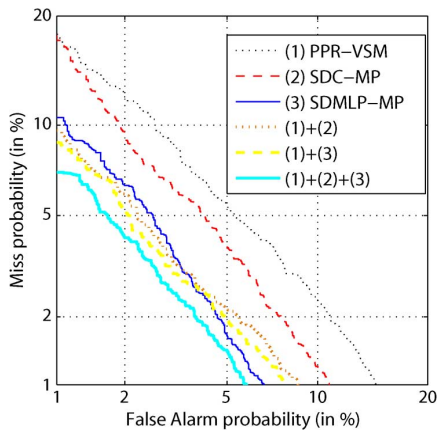

Fig. 3. DET curves of systems on LRE05 30s task.

We then evaluated the performances of SDMLP features in different systems with comparison to the SDC features. For each system, an LDA+Gaussian score backend was applied. This backend first transforms score vectors with linear discriminant analysis and then uses tied covariance models to model the resulting score vectors. The development set and part of Call-Friend Corpus were used to train the backend. Results are shown in Table IV. The proposed SDMLP features perform consistently better than the SDC features. To test the statistical significance, we conducted McNemar's test [16] between the language recognition performances of SDC features and SDMLP features when fLFA and MP systems were used. The improvement of the SDMLP features over the SDC features was verified to be statistically significant (p-values $< 1 \times 10^{-4}$ for 3s, 10s and 30s conditions). Another interesting observation is that although the non-linguistic variations have been reduced in the SDMLP features, both fLFA and discriminative training work well with the SDMLP features. This is consistent with the observations in [7], which reported that the MLP-based features worked well with MLLR adaptation and MMI training in speech recognition task.

The SDMLP features are computed from the frame-level outputs of the MLP-based phoneme recognizer. Frame-level information is utilized by acoustic systems, and phoneme recognizers are utilized by phonotactic systems. So one more question is whether the performances of the SDMLP features are complementary or redundant to the fusion performances of the existing acoustic and phonotactic systems. To address this concern, we built a PPR-VSM [2] system using the three BUT phoneme recognizers (Czech, Hungarian and Russian) [10], and made linear fusions of the SDMLP model pushing (SDMLP-MP) system, the SDC model pushing (SDC-MP) system and the PPR-VSM system on the 30s tasks. Corresponding results are shown in Fig. 3. The SDMLP-MP system complements the SDC-MP system and the PPR-VSM system very well. The fusion between the SDMLP-MP system and the PPR-VSM system leads the performance to 3.07% in EER, which is a 7.80% relative reduction compared to the performance of the SDMLP-MP system. The fusion of these three systems achieves 2.81% in EER, which is a 13.8% relative reduction compared to the fusion performance of the SDC-MP system and the PPR-VSM system.

IV. CONCLUSIONS AND FUTURE WORK

In this letter, a new feature named SDMLP is proposed for SLR tasks. The extraction of SDMLP features includes generating MLP outputs, taking logarithm, PCA transformation, and shifted-delta operation. Quantitative analysis shows that the proposed SDMLP features exhibit similar distribution as the SDC features and are more robust against speaker variations. Experiments on LRE05 dataset show that the SDMLP features outperform the SDC features significantly and complement with the existing acoustic and phonotactic approaches. Future work may include feature dimension reduction of the SDMLP features and feature-level combination between the SDMLP features and the SDC features.

REFERENCES

[1] R. Tong *et al.*, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *Proc. ICASSP*, 2006, pp. 205–208.
[2] H. Li *et al.*, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 271–284, 2007.
[3] E. Singer *et al.*, "Acoustic, phonetic, discriminative approaches to automatic language identification," in *Proc. EUROSPEECH*, 2003, pp. 1345–1348.
[4] P. A. Torres-Carrasquillo *et al.*, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP*, 2002, pp. 89–92.
[5] P. Matejka *et al.*, "Brno university of technology system for NIST 2005 language recognition evaluation," in *Proc. IEEE Odyssey Speaker and Language Recognition Workshop*, 2006, pp. 1–7.
[6] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.
[7] Q. Zhu *et al.*, "On using MLP features in LVCSR," in *Proc. ICSLP*, 2004, pp. 921–924.
[8] D. Imseng *et al.*, "Hierarchical multilayer perceptron based language identification," in *Proc. INTERSPEECH*, 2010, pp. 2722–2725.
[9] L. F. D'Haro *et al.*, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," in *Proc. INTERSPEECH*, 2012, pp. 1–4.
[10] [Online]. Available: http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context
[11] P. Schwarz *et al.*, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, 2006, pp. 325–328.
[12] L. T. DeCarlo, "On the meaning and use of kurtosis," *Psychol. Meth.*, vol. 2, no. 3, pp. 292–307, 1997.
[13] N. Dehak *et al.*, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.
[14] F. Castaldo *et al.*, "Acoustic language identification using fast discriminative training," in *Proc. INTERSPEECH*, 2007, pp. 346–349.
[15] W. M. Campbell, "A covariance kernel for SVM language recognition," in *Proc. ICASSP*, 2008, pp. 4141–4144.
[16] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.
[17] W. M. Campbell *et al.*, "A comparison of subspace feature-domain methods for language recognition," in *Proc. INTERSPEECH*, pp. 309–312.