Feature Normalization Using MVAW Processing for Spoken Language Recognition

Chien-Lin Huang, Shigeki Matsuda, and Chiori Hori
National Institute of Information and Communications Technology, Kyoto, Japan
E-mail: {chien-lin.huang, shigeki.matsuda, chiori.hori}@nict.go.jp Tel:+81-774986316

Abstract— **This study presents a noise robust front-end post-processing technology. After cepstral feature analysis, the feature normalization is usually applied for noisy reduction in spoken language recognition. We investigate a highly effective MVAW processing based on standard MFCC and SDC features on NIST-LRE 2007 tasks. The procedure includes mean subtraction, variance normalization, auto-regression moving-average filtering and feature warping. Experiments were conducted on a common GMM-UBM system. The results indicated significant improvements in recognition accuracy.**

I. INTRODUCTION

Due to the trend of globalization, the ability to identify the multiple spoken languages has become important. There are many spoken language recognition applications such as multilingual speech recognition, machine understanding systems and call service [1]. The state-of-the-art spoken language recognition uses signal processing and statistical modeling techniques to characterize language and reduce the speakers, channels and noisy effects at the same time.

Acoustic and phonotactic systems are two main categories in spoken language recognition. They are often implemented respectively and then combined for the final recognition decision [2]. The acoustic system is derived from the speech signal such as Mel-frequency cepstral coefficients (MFCC). Gaussian mixture models (GMM) [3] and support vector machine (SVM) [4] shown the useful modeling techniques to reflect the statistical patterns. The phonotactic system is motivated by languages differ in the arrangement of sound tokens. The sound tokens can be any acoustically meaningful segments such as phones, syllables or lexical words. Sound tokenizers decode the input speech into a sound token sequence and further derive the phonotactic features with n-gram approach [5]. Many studies adopted phone units as the sound tokens such as in the parallel phone recognizer (PPR) front-end [1].

Spoken language recognition can be divided into three components including feature analysis, statistical modeling and evaluation. Due to the variability of speaker, channel, gender and environment, the compensation of cepstral features for mismatch has been critical for good performance in many speech applications. Recently, model compensation approaches are successfully applied for speaker and language recognition. Eigenchannels used in GMM considers the various channel factors that provide the good solution for channel mismatch [6]. Nuisance attribute projection (NAP) that removes the irrelevant expansion to speaker recognition is used for channel compensation [7]. Common methods can be used for the feature domain compensation [8]. Several front-end analysis methods are provided to solve noisy, channel and speaker mismatch problems such as RelAtive SpecTrA (RASTA), cepstral mean subtraction (CMS), cepstral variance normalization (CVN) and vocal tract length normalization (VTLN). RASTA processing [9] is used for noise reduction. In feature analysis, cepstral mean subtraction and cepstral variance normalization are used to compensate for the linear channel variations [10]. Many efforts have been devoted to advance the performance of spoken language recognition in the past years.

Noise-robustness is one of the most important problems in spoken language recognition. This paper furthers the feature studies by proposing a robust front-end analysis approach. We applied the MVAW front-end prost-processing on MFCC and shifted delta cepstra (SDC) features. M means the cepstral mean subtraction. V is the cepstral variance normalization. A denotes the auto-regression moving-average filtering. W means the feature warping. This study applies an auto-regression moving-average (ARMA) filter in the cepstral domain. The mean subtraction and variance normalization of time sequences in each frame is further processed by an ARMA filter and feature warping. The combination of CMS, CVN, ARMA filtering and feature warping achieves better performance than mean subtraction and variance normalization alone in the conventional spoken language recognition. The outline of this paper is in the following. Section II presents the proposed robust front-end analysis scheme for spoken language recognition. In Section III, we describe the experimental setup and report the experiment results. Finally, Section IV concludes this work.

II. THE PROPOSED SCHEME

As shown in Fig. 1, there are several procedures of the front-end analysis for speech applications. In speech recognition, Mel-cepstral analysis is processed with MVN (CMS and CVN), and ARMA filtering for feature normalization [13]. In speaker recognition, the front-end processing includes Mel-cepstral analysis, RASTA, MVN, and feature warping. Finally, the process of VTLN, Mel-cepstral analysis, RASTA, SDC, and MVN is usually applied for spoken language recognition [2]. This study proposed a procedure of robust front-end analysis for spoken language

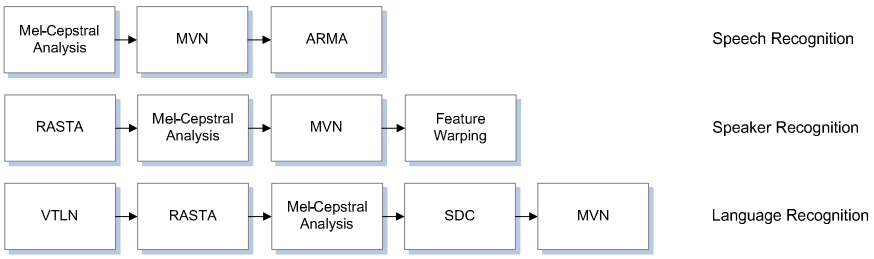

Figure 1. Procedures of the front-end analysis for speech recognition, speaker recognition and language recognition.

Figure 2. Procedures of robust front-end analysis for language recognition.

recognition. We investigate the feature normalization processing technique consisting of mean subtraction, variance normalization, ARMA filtering and feature warping as Fig 2.

A. MVAW Processing

MFCC was used as the short-time frequency feature. After the log-amplitude of the magnitude spectrum, frequency bins are smoothed with the perceptually motivated Mel-frequency scaling.

$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \qquad (1)$$

The feature vectors used in this study were the 7 dimensions of raw MFCCs and SDC with the 7-1-3-7 parameter configuration. Each frame of the speech data is represented by a 56-dimensional feature vector. The frame rate of 16 ms frame size (128 samples at 8k Hz sampling rate and 64-sample shift) and 8 ms frame shift is applied in this study. The Mel-frequency cepstral coefficient is estimated with

$$\mathbf{c}_m = \sum_{n=1}^{N} \mathbf{e}_n \times \mathbf{R}_{m,n}, \quad m = 1, 2, ..., M, \qquad (2)$$

where $M = 7$ is the number of cepstral coefficients and $N = 24$ is the number of filters in the Mel-filter bank. The log filter-bank energies $\mathbf{e}_n$ represent in the spectral domain. Discrete cosine transformation (DCT) matrix $\mathbf{R}_{m,n}$ is used for the spectral-cepstral transformation.

$$\mathbf{R}_{m,n} = \sqrt{\frac{2}{N}} \cos\left(\frac{\pi m}{N}(n - 0.5)\right). \qquad (3)$$

The SDC are estimated by stacking delta-cepstral across multiple frames. There are four parameters {*M, d, P, k*} specified in SDC processing. *d* is the advance and delay for the delta computation. *P* is the time shift between consecutive blocks. *k* is the number of blocks whose delta-coefficients are concatenated to form the final feature vector.

$$\mathbf{c}'(t) = \mathbf{c}(t + iP + d) - \mathbf{c}(t + iP - d). \qquad (4)$$

After SDC processing, speech activity detection is applied to remove silences in the utterance. The accuracy of speech activity detection is important for reliable and robust language recognition. This study applied a hybrid endpoint detector [13]. The strategy is to find endpoints using a three-pass approach in which energy pulses were located and edited, and the endpoint pairs were scored in the order of most likely candidates.

The feature normalization is applied to reduce the noisy and channel effects after cepstral feature analysis. The first step of MVAW processing is mean subtraction defined as

$$\bar{\mathbf{c}}_t = \mathbf{c}'_t - \boldsymbol{\mu}, \quad \boldsymbol{\mu} = \sum_{t=1}^{T} \mathbf{c}'_t / T, \qquad (5)$$

where $\boldsymbol{\mu}$ denotes a mean vector. Moreover, variance normalization is estimated as follows

$$\hat{\mathbf{c}}_t = \bar{\mathbf{c}}_t / \boldsymbol{\sigma}, \quad \boldsymbol{\sigma} = \sqrt{\sum_{t=1}^{T} \left(\mathbf{c}'_t - \boldsymbol{\mu}\right)^2 / T}, \qquad (6)$$

where $\boldsymbol{\sigma}$ is an estimate of the standard deviation. Additionally, the cepstral mean subtraction and cepstral variance normalization are applied for slowly varying convolutive noises [10].

After mean subtraction and variance normalization, the auto-regression and moving average filtering is applied for further noisy reduction and defined by

$$\tilde{\mathbf{c}}_t = \left(\sum_{a=-A}^{t-1} \tilde{\mathbf{c}}_a + \sum_{a=t}^{A} \hat{\mathbf{c}}_a\right) / (2A + 1), \qquad (7)$$

TABLE I
Training data of 14 target languages from different corpus & dialects

Languages	Corpus & Dialects
Arabic	CallFriend, LRE07dev
Bengali	LRE07dev
Farsi	CallFriend
German	CallFriend, OHSU
Japanese	CallFriend, OHSU
Korean	CallFriend, OHSU
Russian	LRE07dev
Tamil	CallFriend, OHSU
Thai	LRE07dev
Vietnamese	CallFriend
Chinese	CallFriend-Ch, CallFriend-Mandarin, LRE07dev-cfr, LRE07dev-wuu, LRE07dev-yuh, OHSU-MM, OHSU-TM
English	CallFriend-USEng-NSD, CallFriend-USEng-SD, OHSU-IE OHSU-AE
Hindustani	CallFriend, LRE07dev, OHSU
Spanish	CallFriend-Spanish-CD, CallFriend-Spanish-NCD, OHSU

where A is the order of the ARMA filter. Besides shifted delta cepstra (SDC) approach in the time sequence filtering, an auto-regression moving-average filtering is directly applied in the cepstral domain. The ARMA filter is a low-pass filter, smoothing out any spikes in the time sequence [12]. Moreover, feature warping is applied for normalizing features to a standard normal distribution. Feature warping is used to reduce the additive noise and channel effects. The details can be found in [11].

B. Acoustic Spoken Language Recognition

The Gaussian mixture modeling classifier is used in this study which is assumed to consist of a mixture of a specific number of multivariate Gaussian distributions. The gender-dependent universal background model (UBM) and the language GMMs consist of mixture of 256 Gaussians. The iterative EM algorithm is used to estimate the parameters of Gaussian components. A log-likelihood ratio (LLR) based evaluation function is applied for the testing utterance $\mathrm{O} = \{o_1, ..., o_T\}$.

$$\Lambda = \frac{1}{T}\sum_{t=1}^{T}[\log p(o_t \mid \lambda^s) - \log \sum_{\forall q} p(o_t \mid \lambda^q)], \qquad (8)$$

where $\lambda_j = \{\mu_j, \Sigma_j, w_j\}$ and T is the number of frames,

$$p(o_t \mid \lambda_j) = w_j \times \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_j|}} \times \exp[-\frac{1}{2}(o_i - u_j)^T \Sigma_j^{-1}(o_i - u_j)], \qquad (9)$$

where v_i is the i-th test feature; u_j is the mean vector of model j-th Gaussian component; Σ_j represents the covariance matrix, and d denotes the dimension of the mean vector u_j. w_j is the mixture weight. λ^s and λ^q are GMMs corresponding to the target language s and possible background languages q, respectively. If the log-likelihood score is higher than the threshold $\Lambda > \theta$, the claimed spoken language will be accepted, else rejected.

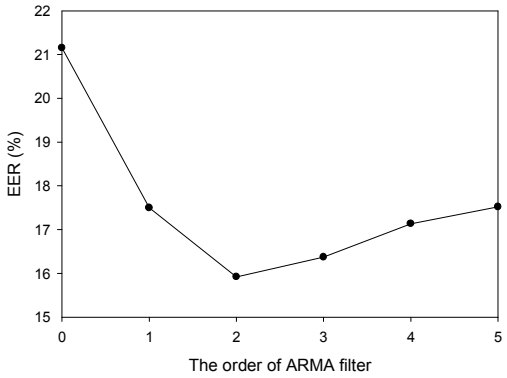

Figure 3. EER evaluations with the order of ARMA filter.

TABLE II
IMPROVEMENTS IN PERFORMANCE OBTAINED SUCCESSIVELY WITH DIFFERENT NORMALIZATION TECHNIQUES

System	NIST-LRE07, EER		
	30s	10s	3s
Baseline	36.58 %	37.96 %	42.54 %
with MVN	28.57 %	30.35 %	32.91 %
with MVW	21.15 %	23.97 %	30.83 %
with MVAW	15.92 %	19.66 %	26.86 %

III. EXPERIMENTS

The following experiments were conducted on the 2007 NIST Language Recognition Evaluation (LRE[1]) including 30 seconds, 10 seconds and 3 seconds trails. All the results presented in this study are closed-set language recognition including, Arabic, Bengali, Chinese, English, Farsi, German, Hindustani, Japanese, Korean, Russian, Spanish, Thai, Tamil, Vietnamese. Spoken language models were trained on the CallFriend[2], OHSU, and NIST-LRE 2007 development data. In this study, there were 32 spoken language models trained for the identification of 14 target languages according to different corpus and dialects. Details were shown in Table I. After VAD processing, there are about 36.58 minute data for each gender-dependent language models.

A. Evaluation of the Order of ARMA Filter

Two types of errors, false acceptance and false rejection, occur in spoken language recognition. The results of spoken language recognition were evaluated by the equal error rate (EER) in this study. EER reports the system performance when the false acceptance and false rejection rates are equal. The results of spoken language recognition were both

[1] http://www.nist.gov/speech/tests/lre/
[2] http://www.ldc.upenn.edu/

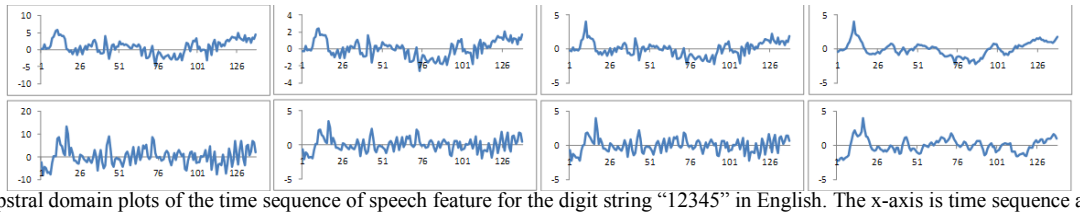

Figure 4. Cepstral domain plots of the time sequence of speech feature for the digit string "12345" in English. The x-axis is time sequence and the y-axis means the log magnitude. The first row is $\mathbf{c}^5$ and it presents MFCC features. The second row is $\mathbf{c}^{10}$ and it denotes SDC features. Within each box, original features (first column), MVN processing (second column), MVW processing (third column), MVAW processing (fourth column).

evaluated by the EER in this study. This study evaluated the variety of windows to obtain a best result for ARMA filtering as shown in Fig. 3. The case of $A = 0$ means no ARMA filtering. Based on experiments, the best case of $A = 2$ was selected in this study.

B. Evaluation of Baseline System and MVAW

Table II shows the improvements in performance obtained by using different techniques. The baseline system was GMM with 256 Gaussian mixture components, features were 7 dimension MFCCs and 49 dimension SDCs. The error rate of the baseline system is very high. The following feature normalization approaches can be greatly improved by CMS, MVN, ARMA filtering and feature warping step by step. Compared with baseline system, the EER score of feature normalization of MVN, MVW and MVAW had significantly reduced in the 30 second trials with 21.90%, 42.18% and 56.48% relative EER reduction, respectively. There was 44.28% relative EER reduction from the conventional MVN normalization (28.57%) to the proposed novel MVAW processing (15.92%). Figure 4 shows the visualization of MVN, MVW and MVAW processing. We see the more smoother and the well noise reduction between original and MVAW processing cases.

IV. Conclusions

This study presented the robust front-end post-processing for spoken language recognition. The MVAW post-processing yields good results to solve the noise and mismatched training/testing conditions. The feature normalization of MVAW shows significant to front-end processing in the experiments. Results in this paper demonstrate the progress has been made by MVAW. MVAW achieved about 24.73% relative EER reduction from 21.15% (MVW processing) to 15.92% (MVAW processing). There was still room for further improving the performance with the Maximum Mutual Information (MMI) training, eigenchannel algorithm and increasing the number of Gaussian components in place [14].

Acknowledgment

The authors would like to thank Bin Ma for his indispensable advice, support, and comments.

References

[1] M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[2] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, "Brno University of Technology System for NIST2005 Language Recognition Evaluation," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 1–7, 2006.

[3] 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.

[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] R. Tong, B. Ma, H. Li, and E.-S. Chng, "Target-Oriented Phone Selection from Universal Phone Set for Spoken Language Recognition," in *Proc. Interspeech*, 2008.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp.1435–1447, 2007.

[7] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advance in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, pp. 629–632, 2005.

[8] W. M. Campbell, D. E. Sturim, P. Torres-Carrasquillo, and D. A. Reynolds, "A Comparison of Subspace Feature-Domain Methods for Language Recognition," in *Proc. Interspeech*, pp. 309–312, 2008.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.

[10] A. Viikki, and K. Laurila, "Segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[11] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. 2001: A Speaker Odyssey*, pp. 213–218, 2001.

[12] C.-P. Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Trans Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[13] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 29, no. 4, pp. 777–785, 1981.

[14] V. Hubeika, L. Burget, P. Matejka, and P. Schwarz, "Discriminative Training and Channel Compensation for Acoustic Language Recognition," in *Proc. Interspeech*, pp. 301–304, 2008.