ROBUST FEATURE FRONT-END FOR SPEAKER IDENTIFICATION

*Gang Liu, Yun Lei and John H.L. Hansen**

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA
{gxl083000, yxl059200, John.Hansen}@utdallas.edu

Abstract

One important challenge for speaker identification (SID) system is sustained performance in diverse conditions. This study presents a novel front-end feature extraction method for SID in clean, noisy, and channel-mismatched acoustic conditions. To address the problem, the perceptual minimum variance distortionless response (PMVDR) feature is employed. While PMVDR has been successfully used for noisy ASR, it has not been considered for SID. We also incorporate longer temporal speaker knowledge based on the shifted delta cepstral (SDC) algorithm. The evaluation over YOHO and another new diversified Robust Open-Set Speaker Identification (ROSSI) database show that both PMVDR and the union with SDC can improve performance significantly. Compared with traditional feature extraction, PMVDR and PMVDR-SDC always give improvement across diverse adverse conditions. Also, PMVDR-SDC can contribute additional improvement in the presence of noise and channel mismatch.

Index Terms: PMVDR, SDC, speaker identification, noise, robustness

1. Introduction

Speaker identification (SID) systems are increasingly employed in real-world applications such as voice authentication, forensics, and surveillance. The idea behind any SID system is to identify the inherent differences in the different speakers' articulatory organs (the structure of the vocal tract, the size of the nasal cavity, and vocal cord characteristics) and the manner of speaking (wording, repetition, prosodic traits).

In SID, although system derived from clean speaker speech data can be recognized usually with high accuracy, recognition performance decreases dramatically for noisy and channel mismatched speech. The adverse condition from application poses real challenge to any practical SID system. Significant amount of research has been conducted in finding speech features that would yield maximum information about the identity of the speakers, thereby increasing the accuracy of the SID system. Most published works in the areas of speech recognition and speaker recognition focus on speech under the noiseless environments and few published works focus on speech under noisy conditions [1, 2, 3, 4].

Mel frequency cepstral coefficients (MFCCs) have proven to be one of the most effective feature sets for speech processes, especially automatic speech recognition (ASR) and SID. They are computed by applying a Mel-scaled filter-bank either to the short-term FFT magnitude spectrum or to the short-term LPC-based spectrum to obtain a perceptually meaningful smoothed gross spectrum. Both the FFT and LPC-based spectrum, however, have limited ability to remove undesired harmonic structures, especially for high pitch speech [5], which may affect speaker representation. Furthermore, studies have shown that FFT-based MFCCs are less effective to suppress noise disruption than other feature front-end [6].

Perceptual minimum variance distortionless response (PMVDR) feature front-end, on the other hand, can directly warps the FFT power spectrum of speech during the feature estimation process, removing the traditional Mel-scaled filterbank as a perceptually motivated frequency partitioning [8]. It can provide a better approximation of the perceptual scales. Another advantage is that PMVDR can effectively model medium and high-pitch speech and track the upper envelope. Therefore, it has the potential to provide more details about speaker excitation information and yield higher accurate recognition. Although researchers have already shown that PMVDR can provide superior performance in ASR, dialect identification (DID) and emotion identification (EID), little research has ever explored its application for speaker identification.

In language identification, shifted delta cepstrum (SDC) approach [7] is widely used. SDC algorithm can incorporate additional temporal information into the feature vector. Although we can also try to integrate longer temporal details by increasing windowing length before FFT process, it is limited by the short-term stationary assumption behind all speech processing techniques. This study, also the major contribution, will mainly rely on PMVDR and SDC to explore the robust speaker identification tasks.

The remainder of this paper is organized as follows: Sec. 2 describes the two databases that are used to develop and evaluate the system. The baseline system is introduced in Sec. 3. Sec. 4 presents the different feature extraction schemes. Sec. 5 provides the SID experiment results and analysis and Sec. 6 presents conclusion and future work.

2. Evaluation Corpora

To verify performance, we work on two corpora. One is clean and another is realistic with diversified mismatched condition.

2.1. YOHO

The YOHO Database consists of 138 speakers, 30 of them female and 108 female. The data was collected over a three month period, with approximately 3 day verification intervals. The speech data consists of a series of combination-lock phrases, for example 24-52-78. For each speaker, there are 4 *enrollment* sessions (each contains 24 phrases) and 10

*This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

verification sessions (each contains 4 phrases). The data was recorded at 8kHz with a 3.8kHz bandwidth at 16 bits per sample. This paper fully follows the corpus structure (Tab. 1). Although this database contains some recording environment noise, we call it "clean" and will artificially introduce some noise to produce different noisy versions of the database to test the robust performance of different feature extraction front-end in the matched acoustic conditions.

Table 1. *YOHO Database.*

	# speaker	# session/speaker	# total session	Avg. duration (sec.)
Training	138	96	13248	4.05
Testing	138	40	5520	4.14

2.2. ROSSI

We also use the Robust Open-Set Speaker Identification (ROSSI) database, which is designed to test and evaluate the robustness of both closed-set and open-set SID systems in various mismatched conditions. There are 10 evaluation sets total, two (set 9 and 10) of which are blind and can only be evaluated by the corpus owner. Aside from these blind sets, each evaluation set contains 100 in-set speakers (data for both training and testing) and 100 out-of-set speakers; here we only use in-set speakers for this research (Tab. 2). Each set has different channel with or without different background noise, which aims to capture real situations of SID and no artificial noise is added. From Tab. 2, we can see that all the data set are very noisy (SNR<10dB) except that set 1 is relative clean.

Table 2. *ROSSI Database. (Mic=Microphone, Cell=Cellphone, "Various" represents the case which has different channel, noise mixed conditions seen in Set 1 through 6.).*

Set	Train	SNR (dB)	Test	SNR (dB)	Type
1	Table-Mic	34.6	Lapel-Mic	27.2	Mic Physical Separation
2	Cell Public	6.1	Cell Public	5.8	Noisy (Channel & Background Variation)
3	Cell Public	6.1	Cell Vehicle	5.9	
4	Landline Office	7.7	Cell Office	8.0	
5	Landline Office	7.7	Cell Vehicle	5.9	
6	Cell Roadside	5.8	Landlin e Office	7.7	
7	Cell Office	7.9	Various	6.6	Mixed (channel & noise)
8	Cell Vehicle	5.9	Various	6.4	

3. GMM-Baseline System

The Gaussian Mixture Models (GMM) classifier is a popular method for text independent SID. We use this approach as our baseline system (We note that GMM based classifier is not state-of-the-art SID system, the reason for this choice is that the focus here is front-end and complicated backend system may make it difficult to single out contribution only due to the feature-end difference.). Figure 1 shows the block diagram of the baseline GMM training/testing system. The noise reduction module is implemented by using extended spectral subtraction to mitigate the noise disturbance [11]. The feature extraction module is the focus of this work and will be supplied with different schemes discussed in Sec.4. The feature warping module is used to Gaussianlize the extracted feature to approach normal distribution and thus better match the GMM modeling assumption. Another benefit is that it has noise robustness [9]. Then speaker dependent GMMs are trained. While testing, the incoming audio is classified as a particular speaker based on the maximum posterior probability measure over all the GMM candidates. Except the feature extraction module, all the other modules are fixed to provide a fair comparison.

The mainstream feature for SID is MFCC and therefore is used as baseline feature. In our study, an analysis window of 20msec duration is used, with 10msec frame update rate. We use traditional 36-dimensional feature vector together with the GMM classifier to provide a benchmark system.

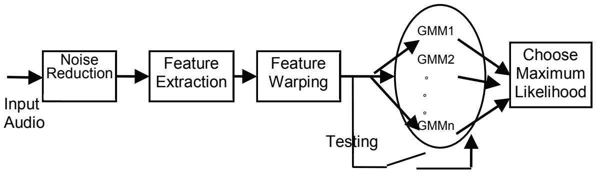

Figure 1: *Baseline GMM based SID system.*

4. Feature Extraction Front-end

4.1. PMVDR

Previous research [8] showed that PMVDRs are better able to model the upper spectral envelope at the perceptually important harmonics, which may include important speaker clues. Unlike MFCC parameters, PMVDRs do not require an explicit filterbank analysis of the speech signal. We have found this new feature representation provides not only robustness against noise in speech recognition, but also higher accuracy in clean speech tasks. Here, we propose to test this feature in the context of SID. A block diagram of the PMVDR feature extraction [8] is shown in Figure 2.

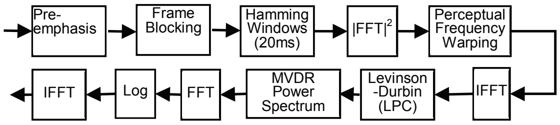

Figure 2: *PMVDR feature extraction process.*

It has been shown that implementing the perceptual scales through the use of a first order all-pass system is feasible. In fact, both Mel and Bark scales are determined by changing the only parameter, *α*, of the system. The filter, *H(z)*, and the warped frequency, $\hat{\omega}$, are given as

$$H(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, | \alpha | < 1 \tag{1}$$

$$\hat{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{2}$$

where ω represents the linear frequency. Here the value of α controls the warping degree. We will optimize the warping factor first before we run any PMVDR-based experiment.

Utilizing direct warping on the FFT power spectrum by removing the filterbank processing step leads to the preservation of almost all the information in the short-term speech spectrum. We can now summarize the remainder of the proposed PMVDR algorithm as follows:

1) Obtain the perceptually warped FFT power spectrum,

2) Compute the "perceptual autocorrelations" by utilizing the IFFT on the warped power spectrum,

3) Perform a i^{th} order LP analysis via Levinson-Durbin recursion using perceptual autocorrelation lags,

4) Calculate the i^{th} order MVDR spectrum from the LP coefficients according to Eq.(1) in [8],

5) Obtain the final cepstrum coefficients using the straightforward FFT-based approach.

Finally, we use 36-dimensional PMVDR features and each feature vector contains 12 statics, deltas and delta-deltas. We use the same windowing and frame skipping as in MFCC before further processing. Cepstral mean normalization is also utilized. Since PMVDR removes the filterbank processing, we can avoid the demanding computation and noise sensitivity incurred by filterbank processing. This is crucial to realistic SID system.

4.2. SDC

The aim of including shifted delta cepstrum (SDC) in the context of SID is to incorporate additional temporal information into the feature vector. The SDC is in fact *k* blocks delta cepstrum coefficients [7]. Suppose the basic set of cepstrum coefficients, $\{c_j(t)$, $j = 0, 1, \ldots, N-1\}$, is available (which are PMVDR statics in this study) at frame *t*, where *j* is dimension index and *N* the number of cepstrum coefficients. The SDC feature can be expressed as following:

$$s_{(iN+j)}(t) = c_j(t + iP + d) - c_j(t + iP - d), \quad i = 0,1,\ldots,k-1 \tag{3}$$

where *d* is the time difference between frames, *P* is the time shift between blocks, and *k* is the total number of blocks. The SDC coefficients can be concatenated with the basic cepstrum coefficients. Thus, we can obtain the feature vector as $\{c_j(t)$, j=0,1, …, N-1; $s_{(iN+j)}(t)$, j=0, 1, …, N-1, i=0, 1, …, k-1$\}$, which is the SDC version of features.

The parameter configuration of SDC *N-d-P-k* in language identification is 7-1-3-7. In our SID task, we fix the optimal configuration at 10-1-3-3 based on hill-climbing searching.

5. Experiment Results

To evaluate the proposed front-end feature schemes, we will compare PMVDR, PMVDR-SDC with MFCC-based SID system. To explore the robustness of the different feature extraction schemes in noisy conditions, we also introduce additive white Gaussian noise (AWGN) with the varying SNR (signal noise ratio) level from -5dB to 20dB into training set and testing set of the YOHO database. To further explore the potential benefit of the proposed schemes, a similar experimental set-up is also carried out on the more realistic ROSSI corpus.

5.1. Warping Factor Optimization

Before exploring the performance of PMVDR on SID system, we need to optimize the warping factor. The range of warping factor is [0, 1], we search the space with the step 0.1 and summarize the results in Figure 3. From the searching results, we can see [0.1, 0.5] is an ideal searching space, we can get optimal value at 0.4 for the PMVDR warping factor and will use this optimal value in the rest of this study for PMVDR-involved feature extraction front-end.

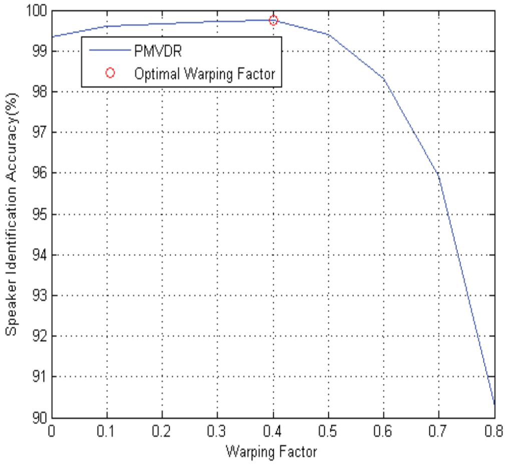

Figure 3: *PMVDR Warping Factor Optimization.*

5.2. Feature Extractions Performance

Now we consider an evaluation of the effectiveness of the proposed various schemes. All GMM-based classifiers backend are based on the same experiment setup as in Sec. 3, and MFCCs are the feature for the baseline system. The noise robustness performance of the features under different SNR level on the noisy versions of YOHO is summarized in Tab. 3. This pilot experiment is only to verify the effectiveness of proposed schemes in a single variation: noise. The noise level is always matched, namely, the train and test set have the same SNR level, so we also perform the evaluation in a more realistic corpus, ROSSI, and check the performance of different front-end in channel, noise mismatch condition, which is reported in Figure 4.

Table 3. *SID performance on YOHO Database.*

Error Rate (%)	-5dB	0dB	10dB	20dB	clean
MFCC	53.06	21.49	3.33	1.00	0.36
PMVDR	45.45	17.63	2.99	0.96	**0.25**
PMVDR+SDC	**38.71**	**14.46**	**2.37**	**0.71**	0.27

5.3. Analysis

From Tab. 3 we can see that in clean condition, both the performance of PMVDR and the union of PMVDR and SDC are better than MFCC with the error rate decreased by 31% and 25%. Since MFCC+SDC is always worse than MFCC, so we do not test MFCC+SDC in the rest of this study.

This demonstrates PMVDR-based feature extraction can better capture the speech characteristics of different speakers. Although MFCCs perform well in less noisy condition, they

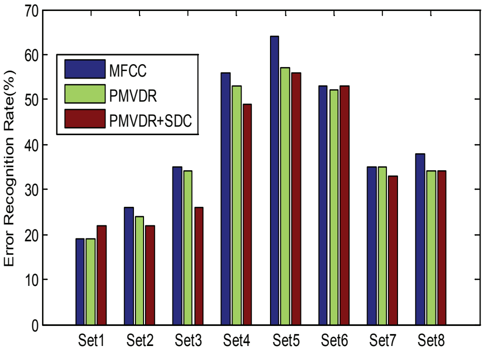

Figure 4: *SID performance on ROSSI Database.*

are much worse than other feature front-end in strong noise disturbance. When the SNR is -5dB, PMVDR and PMVDR-SDC can decrease the error rate by 14% and 27%, respectively. When SNR is varied from -5dB to 20dB, we can observe consistently that the noise robustness of PMVDR-SDC is always the best and that of MFCC is always inferior to the other two.

From Figure 4, PMVDR performs better than MFCC in all evaluation sets except set 1 and 7, on which their performance are similar. PMVDR-SDC can give further improvement in all evaluation sets except in set 1 and 6. On set 6 it gives the similar performance as MFCC and on Set 1 (with only microphone-channel mismatched situation), the introduction of SDC is backfired and deserve further analysis. In mixed cases, set 7 and 8, we can see clearly that PMVDR-SDC always give consistent better performance.

Compared with MFCC, PMVDR does not require an explicit filterbank analysis and thus are less sensitive to noise disturbance, similar results was also reported in other publication in ASR[8] and EID[6]. The introduction of SDC is mean to incorporate additional temporal information into the feature vector and can bring consistent robust improvement to SID in noisy match/ mismatch, and channel mismatch condition. This benefit comes from the cepstrum substraction in Eq.(3). Although we can incorporate longer speech information by enlarging the windowing length during speech processing, it is strictly limited by the inherent short-term stationary premise of speech processing since all the human being's speech related organs can be regarded as stationary only in a very short period.

6. Conclusions

To improve system robustness, signal processing or model adaptation for noise and channel is needed to ensure consistent performance of SID, which is challenging in real applications. We approached this issue from the feature extraction front-end, the very first step in all SID system. The PMVDR feature extraction, which has not been investigated in the context of SID, outperforms the baseline system in both clean (with error rate decreased by 31%, relatively) and strong noise disturbance (with error rate decreased by 14%, relatively). To further improve the system performance, we also explore the application of the SDC algorithm, which can give further relative system improvement up to 50% in clean condition and 31% in strong noisy condition.

The experiment on the more diversified and realistic corpus also verified the effectiveness of PMVDR across different conditions. PMVDR-SDC, especially, in the presence of noise and channel variation, can always give better performance than MFCC. One of the benefits of the introduction of SDC is that it can overcome the short-term stationary assumption to a degree and incorporate longer temporal structure information of speech to give more robustness to the SID system.

Although PMVDR can demonstrate better speech/speaker characterization ability and noise robustness, its warping factor need to be tuned to achieve optimal performance. In general, [0.1, 0.5] is an ideal searching space.

This is only a preliminary exploration in the front-end aiming to propose a good alternative to the popular MFCC. To fully confirm the validity of the proposed schemes, evaluation on more comprehensive tasks such as NIST SRE and state-of-the-art back-end system (for example, Joint Factory Analysis or iVector) will be the future work of this study.

7. References

[1] D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Transactions on SAP, Vol. 2, 1994, pp. 639-643.

[2] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using nonlinear transformation for robust speech recognition on the Aurora database," Proc. ICASSP2000.pp. 1117-1120

[3] D. Wu, A.C. Morris and J. Koreman,"MLP Internal Representation as Discriminant Features for Improved Speaker Recognition," Proc. NOLISP2005, Barcelona, Spain, 2005, pp.25-33.

[4] Y. Konig, L. Heck, M. Weintraub and K. Sonmez, "Non-linear discriminant feature extraction for robust text-independent speaker recognition," Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications, 1998, pp.72-75.

[5] L. Gu and K. Rose, "Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition", Proc. ICSLP'00

[6] G. Liu, Y. Lei, and J.H.L. Hansen, "A Novel Feature Ex-traction Strategy for Multi-stream Robust Emotion Identification," in INTERSPECCH2010, pp482-485.

[7] B. Bielefeld, "Language identification using shifted delta cepstrum," In 14th Annual Speech Research Symposium, 1994.

[8] U. H. Yapanel, J.H.L. Hansen. "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition". Speech Communication 50 (2008) 142–152.

[9] J. Pelecanos and S. Sridharan. Feature Warping for Robust Speaker Verification. In Proceedings of Speaker Odyssey Conference, Crete, Greece, 2001

[10] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," IEEE Trans. on Speech and Audio Proc., pp. 221–239, May 2000.

[11] P. Sovka, P. Pollak, and J. Kybic, "Extended spectral subtraction," in Proceedings of European Signal Processing Conference (EUSIPCO '96), pp. 963-966, Trieste, Italy, September 1996.