Robust Speaker Verification Using Short-Time Frequency with Long-Time Window and Fusion of Multi-Resolutions

Chien-Lin Huang[1], *Bin Ma*[2], *Chung-Hsien Wu*[1], *Brian Mak*[3] *and Haizhou Li*[2]

[1] CSIE Department, National Cheng Kung University, Tainan, Taiwan
[2] Institute for Infocomm Research, Singapore
[3] CSE Department, Hong Kong University of Science and Technology, Hong Kong

Abstract

This study presents a novel approach of feature analysis to speaker verification. There are two main contributions in this paper. First, the feature analysis of short-time frequency with long-time window (SFLW) is a compact feature for the efficiency of speaker verification. The purpose of SFLW is to take account of short-time frequency characteristics and long-time resolution at the same time. Secondly, the fusion of multi-resolutions is used for the effectiveness of robust speaker verification. The speaker verification system can be further improved using multi-resolution features. The experimental results indicate that the proposed approaches not only speed up the processing time but also improve the performance of speaker verification.

Index Terms: speaker verification, short-time frequency with long-time window, fusion of multi-resolutions

1. Introduction

Nowadays, voice biometrics has become increasing popular in telephony applications [1]. The state-of-the-art text-independent speaker verification uses signal processing and statistical modeling techniques to characterize speakers. Typically, speech activity detection is first applied [2], then spectral features that are robust to noise and channel effects are extracted. After speech feature extraction, a speaker verification system makes a decision with Gaussian mixture model (GMM) or support vector machines (SVM) classifier using the criterion of log-likelihood ratio (LLR).

Speaker verification is a pattern recognition problem and the overall procedure can be divided into three components: feature analysis, statistical modeling and evaluation. In feature analysis, cepstral mean subtraction and cepstral variance normalization are used to compensate for the linear channel variations [3]. RASTA (RelAtive SpecTrA) processing [4] is used for noise reduction. Moreover, feature warping is robust to additive noise and linear channel effects [5]. In statistical modeling, maximum a posteriori (MAP) algorithm is the basic approach to speaker adaptation [1]. Eigenvoice provides a rapid speaker adaptation under the condition of sparse training data [6]. Eigenchannels used in GMM considers the various channel factors that provide the good solution for channel mismatch [7]. Nuisance attribute projection (NAP) that removes the irrelevant expansion to speaker recognition is used for channel compensation [8]. In evaluation, the various score normalization approaches are successfully applied for robust speaker verification. The test normalization (Tnorm) of the likelihood score is an online procedure. The input test speech utterance is computed by cohort models to obtain the normalization scores using mean and standard deviation [9]. The zero normalization (Znorm) of the likelihood score is an offline procedure. Every speaker models are tested with imposter speech utterances to obtain the mean and standard deviation scores of normalization [10].

Many efforts have been devoted to advance the performance of speaker verification in the past years. This paper furthers the feature studies by proposing a novel feature analysis approach. To consider the short-time frequency characteristics and long-time resolution at the same time, the short-time frequency with long-time window (SFLW) approach is adopted, which greatly reduces the computation cost. Moreover, this study proposes the fusion of multi-resolution of feature analysis that is different from a combination of several subsystems. The goal of fusion of multi-resolutions is to improve the performance of a single system in various frequency resolutions. The speaker recognition experiments are made using the 2006 NIST Speaker Recognition Evaluation (SRE) core condition test trials. The experimental results indicate that the speaker verification performance has been improved in both the effectiveness and efficiency. The outline of this paper is in the following. Section 2 presents the proposed robust speaker verification scheme using short-time frequency with a long-time window and the fusion of multi-resolutions approaches. Section 3 provides some experimental results and discussions. Finally, Section 4 concludes this work.

2. The Proposed Scheme

As shown in Fig. 1, there are several steps in the scheme of robust speaker verification: First, this study employs the speech activity detection and shot-time frequency feature extraction. Secondly, we analyze the sequence of short-time frequency feature vector using a long-time window. Thirdly, the GMM and eigenchannels are performed to build the universal background models (UBM) and speaker models. The log-likelihood ratio and ZTnorm score normalization are used for the evaluation. Finally, the fusion of multi-resolutions is applied for robust speaker verification.

2.1. The Baseline System of Speaker Verification

The accuracy of speech activity detection is important for reliable and robust speaker verification. This study applied a hybrid endpoint detector [2]. 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. Mel-frequency cepstral coefficient (MFCC) was used as the short-time frequency feature. Each frame of the speech data is represented by a 36-dimensional feature vector, consisting of 12 MFCCs, along with their deltas, and double-deltas as the raw features. After the log-amplitude of the magnitude spectrum, frequency bins are smoothed with the perceptually motivated Mel-frequency scaling. Additionally, the cepstral mean subtraction (CMS)

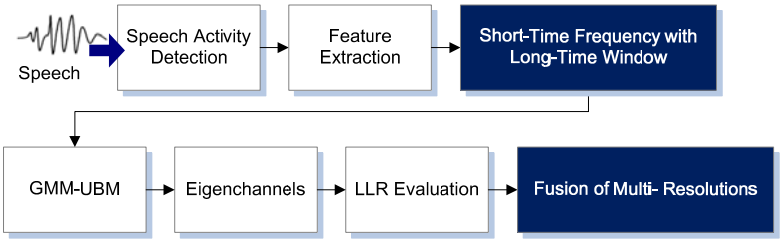

Figure 1: Scheme of the robust speaker verification.

and cepstral variance normalization (CVN) are applied for slowly varying convolutive noises.

The GMM classifier is used in this study which is assumed to consist of a mixture of a specific number of multivariate Gaussian distributions. The iterative EM algorithm is used to estimate the parameters of Gaussian components [11]. There are different types of channel information in the NIST speaker evaluation such as cellular, cordless and land-line telephone callers. One of the key points in NIST SRE is to find a solution for the problem of channel mismatch. Kenny et al. proposed the eigenchannel approach to solve this problem [7]. In eigenchannels, many different channel utterances of speakers were used to estimate UBM and speaker models for the channel mismatch in order to incorporate the channel information into speaker models.

A log-likelihood ratio (LLR) based evaluation function is applied for testing the trials.

$$\Lambda = \frac{1}{T}\sum_{t=1}^{T}[\log p(v_t \mid \lambda_{SPK}) - \log p(v_t \mid \lambda_{UBM})]. \quad (1)$$

If the log-likelihood score is higher than the threshold $\Lambda > \theta$, the claimed speaker will be accepted, else rejected. The log-likelihood score is estimated from the multivariate Gaussian pdf as follows:

$$p(v_t \mid \lambda) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_j|}} \times \exp[-\frac{1}{2}(v_i - u_j)^T \Sigma_j^{-1}(v_i - u_j)], \quad (2)$$

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.

2.2. Short-Time Frequency with Long-Time Window and Fusion of Multi-Resolutions

In feature extraction, speech is broken into small segments for short-time analysis. These segments have to be small enough to ensure the frequency characteristics of the magnitude spectrum are relatively stable. However, the sensation of a sound arises as the result of multiple short-time spectrums with different characteristics, such as vowel and consonant sections [12]. In order to capture the long term nature of signal, the actual features are estimated as the mean of long-time window of the extracted short-time features.

A sequence of feature vectors $X = \{x_1, x_2, ..., x_T\}$ is estimated as $V = \{v_1, v_2, ..., v_N\}$ according to short-time frequency with long-time window analysis (SFLW). *T* denotes the number of short-time frequency frames. *N* denotes the compacted size after the process of SFLW which can be denoted as *N=T/M*. *M* indicates the size of long-time window. Figure 2 shows the example of short-time frequency with long-time window.

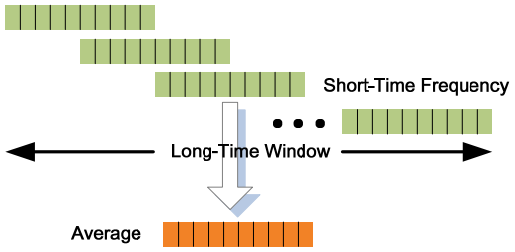

Figure 2: Short-time frequency with long-time window.

The purpose of this transformation is to obtain a new representation of feature analysis which is more compact and suitable for statistical modeling. Due to the frame size reduction, SFLW can speed up the speaker verification process, especially in the steps of evaluation and score normalizations. Two resolutions are applied for the further fusion analysis in this study, including

1) the short-time frequency analysis of 16 ms (128 samples at 8k Hz sampling rate and 64-sample shift) with long-time window of 80 ms containing 8 short-time frequency frames and

2) the short-time frequency analysis of 8 ms (64 samples at 8k Hz sampling rate and 32-sample shift) with long-time window of 72 ms containing 16 short-time frequency frames.

Furthermore, this study proposes a novel fusion of multi-resolutions in feature analysis based on one GMM-UBM system. The fusion of multi-resolutions differs from a combination of different kinds of subsystems which is usually applied for speaker verification, such as GMM-UBM, GMM-SVM and MLLR-SVM [13]. Two different resolutions are applied based on the SFLW analysis, including 128 Length / 64 Shift with 8 Frame and 64 Length / 32 Shift with 16 Frame, and simply fused the verification scores of these two multi-resolution subsystems. The fusion weight was equally set as 0.5. The advantage of fusion of multi-resolutions is to improve the performance of a single speaker verification system by using the various feature resolutions.

3. Experiments

For the NIST SRE-2006 experiments, performance comparisons were made in different training and test conditions.

Table 1. The different size of frame and shift in the baseline system and SFLW.

	Male	Female	Average
64Length / 32Shift	12.36%	14.10%	13.23%
128Length / 64Shift	10.35%	11.87%	11.11%
256Length / 128Shift	9.96%	12.70%	11.33%
512Length / 256Shift	10.10%	12.12%	11.11%
1024Length / 512Shift	11.14%	14.13%	12.64%
64Length / 32Shift with 16Frame	9.60%	12.02%	10.81%
128Length / 64Shift with 8Frame	9.59%	11.42%	**10.51%**

Table 2. Number of feature samples in the baseline system and SFLW.

	Male	Female	Average
64Length / 32Shift	4,281,792	4,619,556	4,450,674
128Length / 64Shift	2,296,120	2,445,205	2,370,663
256Length / 128Shift	1,220,677	1,274,314	1,247,496
512Length / 256Shift	627,712	652,288	640,000
1024Length / 512Shift	294,321	292,233	293,277
64Length / 32Shift with 16Frame	534,971	577,187	**556,079**
128Length / 64Shift with 8Frame	611,149	573,756	**592,453**

Table 3. EER reports with eigenchannels in the baseline systems.

	Male	Female	Average
128Length / 64Shift	5.52%	5.64%	**5.58%**
256Length / 128Shift	5.59%	6.35%	5.97%
512Length / 256Shift	6.91%	6.87%	6.89%
1024Length / 512Shift	8.91%	10.87%	9.89%

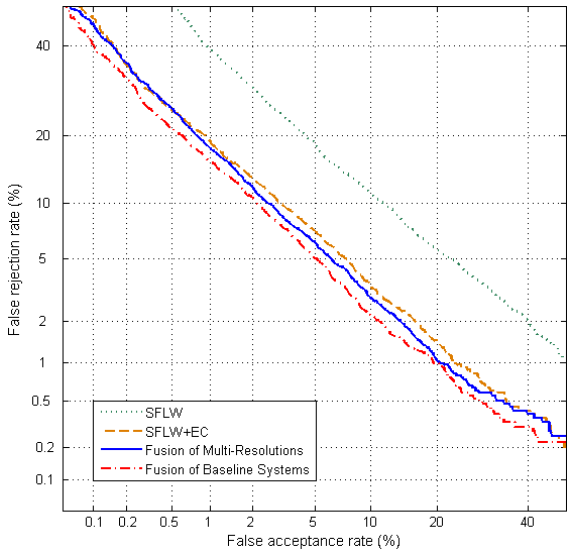

Figure 3: Comparison of improvement with SFLW (the setup of 128 Length / 64 Shift with 8 Frame), SFLW+EC, fusion of multi-resolutions and fusion of baseline systems.

The effected factors, such as language, telephone transmission type, and microphone type were examined in the NIST SRE [14]. The NIST SRE-2004 1side data was used to train a gender-dependent GMM-UBM with 512 Gaussian mixtures. The NIST SRE-2004 training and test data were provided by 168 female and 168 male speakers. The following experiments were measured on the NIST SRE-2006 1conv4w-1conv4w with 51,448 trails [14]. There are 810 enrolled speakers including 461 female and 349 male.

3.1. Evaluation of Baseline System and SFLW

Two types of errors, false acceptance and false rejection, occur in speaker verification. The results of speaker verification 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. Table 1 shows the GMM-UBM performance of the baseline system and SFLW for the different sizes of frame and shift.

The number of FFT sample points is usually a power of 2. It is advantageous computationally to have smaller frame size but too small or too big in frame size will hurt the performances in baseline systems, such as 64 Length / 32 Shift and 1024 Length / 512 Shift. The SFLW with a smaller frame size (128 Length / 64 Shift with 8 Frame) achieved the best performance with an EER of 10.51% in the GMM-UBM experiment.

The number of feature samples was also reported for different settings in GMM-UBM training. As shown in Table 2, the number of feature samples of the SFLW approach is smaller than most of conditions of the baseline system but 1024 Length / 512 Shift. It is obvious that SFLW provides a much compact and effective feature database, especially when compared with the conditions of 256 Length / 128 Shift, 128 Length / 64 Shift and 64 Length / 32 Shift.

3.2. Evaluation of Eigenchannels and Fusion of Multi-Resolutions

To solve the various types of channel effects in NIST-SRE dataset, the eigenchannel approach (EC) was used to improve the performance. The number of eigenchannels was chosen to be EC=30 in this study. The eigenchannel evaluation of the baseline systems was shown in Table 3. It is observed that EER decreases as frame size is reduced. The setup of 128 Length / 64 Shift achieved the best result, EER=5.58%. It is also noted that the smaller frame size leads to higher computation cost according to Table 2. Therefore, the setup of 64 Length / 32 Shift was not examined due to the high computation cost.

The eigenchannel evaluation of SFLW (the setup of 128 Length / 64 Shift with 8 Window) is shown in Table 4. EER has been greatly reduced from 10.51% to 5.87% after the eigenchannel process. The fusion of multi-resolutions was applied for the further improvement. The results show that the multi-resolution features are complementary. The proposed SFLW and the fusion of multi-resolutions approaches obtained EER=5.45%. The proposed approach also reduced 0.13% EER compared with the best case of baseline system 128 Length / 64 Shift. The number of feature samples of fusion of multi-resolutions was 1,148,532 and is only one-half of feature samples compared with the best case of baseline system 128 Length / 64 Shift.

If we don't care about the computation cost, the performance of speaker verification can be further improved by the fusion of SFLW and baseline systems without the setup of 1024 Length / 512 Shift, shown in Table 5. The fusion weights were equal. There is a significant improvement in the single system with different feature resolutions.

Table 4. Evaluation in the fusion of multi-resolutions.

	Male	Female	Average
64Length / 32Shift with 16Frame	5.83%	6.53%	6.18%
128Length / 64Shift with 8Frame	5.45%	6.28%	**5.87%**
Fusion of Multi-Resolutions	5.26%	5.63%	**5.45%**

Table 5. Fusion of SFLW and baseline systems.

	Male	Female	Average
EER	4.84%	5.03%	**4.94%**
Feature Samples	4,755,658	4,945,563	**4,850,611**

Table 6. Evaluation in the fusion of multi-resolutions with ZTnorm.

	Male	Female	Average
64Length / 32Shift with 16Frame	5.79%	6.43%	6.11%
128Length / 64Shift with 8Frame	5.51%	6.08%	5.80%
Fusion of Multi-Resolutions	5.10%	5.49%	**5.30%**

However, the disadvantage is expensive in the computation cost. Finally, the above evaluations are plotted with the well-know Detection Error Tradeoff (DET) curves in Fig. 3. Figure 3 showed that the line of Fusion of Multi-Resolutions means the result of Table 4 and the line of Fusion of Baseline Systems denotes the result of Table 5

3.3. Evaluation of Score Normalization

The tuning of decision thresholds is very tricky in speaker verification due to the score uncertainty caused by the intraspeaker and interspeaker variability [1]. Therefore, score normalization is usually applied in the speaker recognition task. ZTnorm was applied for the score normalization in this study. ZTnorm provides the score normalization both in the speech feature and speaker model domains. The procedure of ZTnorm is that Tnorm is firstly applied and then Znorm speaker models are tested by imposters' speech utterances to find the mean μ_{ZT} and standard deviation σ_{ZT} of the normalization scores.

$$\tilde{\Lambda} = \frac{\Lambda - \mu_{ZT}}{\sigma_{ZT}}. \quad (3)$$

Comparing the experiment setups in Table 4, the log-likelihood ratio was normalized with the ZTnorm scores for the robust speaker verification as shown in Table 6. The results showed that 0.15% EER reduction from 5.45% to 5.30% for the fusion of multi-resolutions. It should be noted that the Tnorm process is considered in the Znorm process at the same time.

4. Conclusions and Future Work

This study has presented short-time frequency analysis with a long-time window and the fusion of multi-resolutions approaches for robust speaker verification. SFLW considers the short-time frequency characteristics and long-time resolution at the same time. It represents a faster and more efficient speaker verification system. The computation cost is greatly reduced in SFLW. The fusion of multi-resolutions further improves the performance of speaker verification system with the different feature resolutions. Experimental results showed that the proposed approaches achieved a satisfactory performance as well as effectiveness and efficiency.

In future work we plan to explore the relations between speech activity detection and short-time frequency with long-time window approaches. Moreover, SFLW would be incorporated with the eigenvoice approach to evaluate the short test of 10 sec of NIST SRE. These results would be interesting. Finally, all results did not contain RASTA and feature warping.

5. Acknowledgements

The authors would like to thank the National Science Council (NSC), Taiwan, Republic of China, for financially supporting this research under Contract No. NSC-096-2917-I-006-121.

6. References

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I.Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk-Delacrtaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Processing*, vol. 4, pp. 430–451, 2004.

[2] Lamel, L., Rabiner, L., Rosenberg, A. and Wilpon, J., "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 777-785, 1981.

[3] Vikki, A. and K. Laurila, "Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Communication*, vol. 25, pp. 133-147, 1998.

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

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

[6] Roland Kuhn, Jean-Claude Junqua, Patrick Nguyen, and Nancy Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695-707, 2000.

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

[8] Alex Solomonoff, W.M. Campbell and Ian Boardman, "Advance in channel compensation for SVM speaker recognition" in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 629–632, 2005.

[9] C. Auckenthaler and Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10 no 1-3, 2000.

[10] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 595–598, 1988.

[11] Pelecanos, J., Myers, S., Sridharan, S. and Chandran, V., "Vector quantization based Gaussian modeling for speaker verification," *International Conference on Pattern Recognition Proceedings*, vol. 3, pp. 294-297, 2000.

[12] Tzanetakis, G. and Cook, P., "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002.

[13] Brummer, N., Burget, L., Cernocky, J.H., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D.A., Matejka, P., Schwarz, P. and Strasheim, A., "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15 no. 7, pp. 2072–2084, 2007.

[14] NIST SRE (Online) http://www.nist.gov/speech/tests/spk/