Noise Robust Speaker Verification Using Mel-Frequency Discrete Wavelet Coefficients and Parallel Model Compensation

Zekeriya Tufekci and Sabri Gurbuz[†‡]

Department of Electrical and Electronics Engineering
Izmir Institute of Technology, Urla-Izmir 35430, Turkey

[†] Harran University EE Department, Sanliurfa 63200, Turkey
[‡]ATR Human Information Science Laboratories, Kyoto 619-0288, Japan
E-mail:zekeriyatufekci@iyte.edu.tr, sabrig@ieee.org

Abstract

Interfering noise severely degrades the performance of a speaker verification system. The Parallel Model Combination (PMC) technique is one of the most efficient techniques for dealing with such noise. Another method is to use features local in the frequency domain. Recently, Mel-Frequency Discrete Wavelet Coefficients (MFDWCs) [1, 2] were proposed as speech features local in frequency domain. In this paper, we discuss using PMC along with MFDWCs features to take advantage of both noise compensation and local features (MFDWCs) to decrease the effect of noise on speaker verification performance. We evaluate the performance of MFDWCs using the NIST 1998 speaker recognition and NOISEX-92 databases for various noise types and noise levels. We also compare the performance of these versus MFCCs and both using PMC for dealing with additive noise. The experimental results show significant performance improvements for MFDWCs versus MFCCs after compensating the Gaussian Mixture Models (GMMs) using the PMC technique. The MFDWCs gave 5.24 and 3.23 points performance improvement on average over MFCCs for -6 dB and 0 dB SNR values, respectively. These correspond to 26.44% and 23.73% relative reductions in equal error rate (EER), respectively.

1. Introduction

Research on speaker verification [3] has been an active area for decades. The goal of speaker verification system is to determine from a voice of sample if a person is whom he or she claims. The speech can be constrained to be known phrase (text-dependent) or totaly unconstrained (text-independent). This study is concerned with the text-independent speaker verification. The GMMs [4] recently have become dominant approach in text-independent speaker verification. In this paper, speakers were modeled using GMMs.

Real world applications require that speaker verification systems be robust to interfering noise. The performance of a speech recognition or speaker verification system drops dramatically when there is a mismatch between training and testing conditions. Many different approaches have been studied to decrease the effect of noise on the performance [5]. One of the most effective and popular model-based techniques for dealing with noisy speech is Parallel Model Combination [6–10]. This technique attempts to estimate matched speech models (noisy speech models) given the clean speech models and a noise model.

In addition to the method mentioned above, recognition systems based on features local in the frequency domain, such as multiband [11–13] and multiresolution [1, 2, 14] techniques, have received great attention for dealing with noisy speech. In this paper, the speech feature vector will be referred to as local if some of the coefficients of the vector represent local information in the frequency domain, even though the other coefficients do not.

Conventional feature extraction methods use the entire frequency band to extract speech features for speech recognition. However, as pointed out by Fletcher [15] (and reviewed by Allen in [16]), the Human Speech Recognition (HSR) system works with partial recognition information across frequencies, probably in the form of speech features that are local in frequency. Fletcher's work [15] led to the subband-based speech recognizer [11, 12].

There are three main motivations for local (in frequency domain) feature-based recognizers:

1. Some subbands of the speech spectrum may be inherently more relevant than others for the task of speech recognition or speaker verification. Therefore, the contribution of each subband to the overall recognition decision can be weighted based on the information that each subband conveys.
2. Transitions between more stationary segments of speech do not necessarily occur at the same time across the different frequency bands. The local feature-based approach may have the potential of relaxing the synchrony inherent in current HMM systems.
3. The local features will be affected differently in noisy environments. When one frequency band is corrupted by noise, only a few coefficients will be affected if the coefficients represent local information; otherwise, the noise will affect all coefficients. Even if the whole frequency band is corrupted by noise, the SNR will be different for each subband (each coefficient). This is the most important property of local features. It allows us to weight the contribution of each coefficient in the global score based on the SNR for each coefficient.

It is known that the recognizer is optimal [9] if the training and testing conditions are identical. A practical method for approaching an optimal recognizer for different noise conditions is PMC. This technique allows for estimating the HMM parameters for new environments. Since the features are local, the estimated HMM parameters for the new environment will represent local information. This is very important because the estimated parameters for a particular coefficient will be affected only if that particular coefficient is corrupted by noise. In the previous work [17] the PMC and MFDWCs were successfuly used for noise robust speech recognition. In this paper the PMC technique was implemented along with MFDWCs to test MFDWCs performance for noisy conditions on a text-independent speaker verification task.

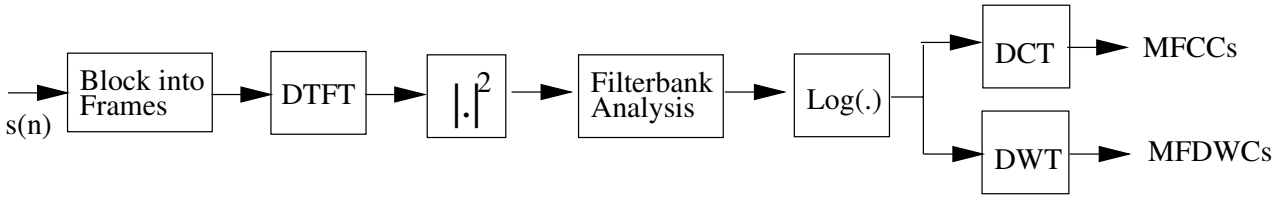

Figure 1: Extraction of the MFCCs and MFDWCs

2. Wavelet transform and MFDWCs

MFDWCs are obtained by applying the DWT to the mel-scaled log-filterbank energies of a speech frame. The WT uses short basis functions to measure the high frequency content of the signal and long basis functions to measure the low frequency content of the signal. This property of the WT makes the WT different from the Short Time Fourier Transform(STFT) and the Fourier Transform(FT).

A wavelet is a function $\psi(t) \in L^2(\mathbb{R})$ (space of square-integrable functions) of zero average and unit norm which satisfies

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \quad (1)$$

and

$$\|\psi(t)\| = 1. \quad (2)$$

$\psi(t)$ is called **mother wavelet**. The analysis function of wavelet transform at scale s and translation u is given by:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}). \quad (3)$$

The wavelet transform of a function $f(t) \in L^2(\mathbb{R})$ at the time u and scale s is given by:

$$F(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*(\frac{t-u}{s})dt = \int_{-\infty}^{+\infty} f(t)\psi^*_{u,s}(t)dt. \quad (4)$$

where * denotes complex conjugate. Parseval's theorem states that

$$F(u,s) = \int_{-\infty}^{+\infty} f(t)\psi^*_{u,s}(t)dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\Omega)\Psi^*_{u,s}(\Omega)d\Omega \quad (5)$$

where $F(\Omega)$ and $\Psi_{u,s}(\Omega)$ are Fourier Transforms (FT) of $f(t)$ and $\psi_{u,s}(t)$, respectively. As seen, the transform of the signal depends on both $\psi_{u,s}(t)$ and FT of $\psi_{u,s}(t)$. So the locality of $f(t)$ in the time and frequency domains depends on the spread of $\psi_{u,s}(t)$ in time and frequency, respectively. When the signal is corrupted by noise local in time and/or in frequency, this noise affects only a few coefficients if the coefficients represent local information in time and frequency. Therefore, we can decrease the contribution of noise corrupted coefficients to the overall recognition score depending on the SNR for noise corrupted coefficients.

Theoretically, any function with zero mean and finite energy can be a wavelet. There are many criteria, though, by which to choose a wavelet. Since we cannot implement a wavelet of infinite duration, we need compactly supported wavelets for practical applications. Decay of the wavelet in the frequency and time domains is important. We want the wavelet to decay quickly in time and frequency in order to have good locality in time and frequency. Filterbank-based wavelets can be implemented efficiently. Since our signals are of finite length, the wavelet coefficients will have unwanted large variations at the borders because of the discontinuities at the frame borders. We can use folded wavelets that require symmetric or anti-symmetric wavelets such as the spline wavelet to decrease the effect of discontinuities at the borders, or we can use border wavelets. When considering all the conditions given above, the options for choosing a wavelet are limited. In addition we use the Discrete Wavelet Transform (DWT) instead of the Continuous Wavelet Transform since our signal is discrete. Figure 2 shows the spreads of the basis function of the wavelet (used in this paper) in the time and frequency domains. We want the spread of the basis functions to be well concentrated in the time and frequency domains for noise robust speaker verification. For additional information about the Wavelet Transform(WT) and implementations of the WT, interested readers may refer to [18, 19].

Figure 1 illustrates extraction of the MFCCs and MFDWCs. The first five steps are the same for both as shown in Figure 1. Only the last step is different in that we take the Discrete Cosine Transform (DCT) of the log-filterbank energies to calculate MFCCs or the DWT of the log-filterbank energies to calculate MFDWCs. The first step is to divide the speech signal into blocks using overlapping smooth windows such as Hammimg, Hanning, etc. The next step is to take the Discrete Time Fourier Transform (DTFT) of the windowed signal. Then the square of the DTFT of the windowed signal is calculated. The outputs of the fourth step are the mel-scaled filterbank energies. The fifth step involves calculating the logarithm of the mel-scaled filterbank energies.

3. The PMC technique applied with MFCCs and MFDWCs

The recognition system is optimal when there is no mismatch between training and testing conditions. The simple solution to get the optimal recognition system is to retrain the system for the new test environment. However, it is not practical to retrain the system since we need the entire training data for the new environment. Even if we have the training data for the new environment, it is a very time consuming process to retrain the system.

The PMC [6–9] technique was proposed by Gales and Young to deal with new testing conditions by estimating the noisy speech model using clean speech and noise models. The PMC technique is very effective and less time consuming compared to retraining the system using the training data for the new environment.

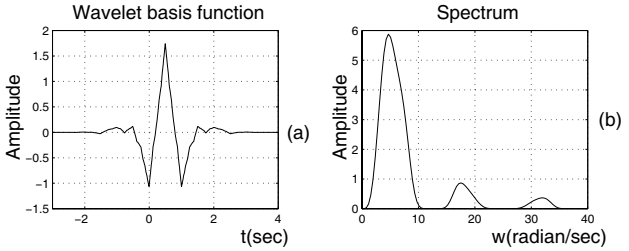

Figure 2: The spreads of the basis functions of wavelet in the time and frequency domains.

There are three approaches in PMC techniques to estimate the noisy speech parameters: numerical integration [8], a data-driven approach [20], and log-normal approximation [6]. Since practical implementation requires less computation, we chose the log normal approximation approach which demands the least computation. The details of the log-normal approximation approach for the MFCCs can be found in [6]. The PMC technique was originally developed for the MFCCs, but it can easily be adopted for the MFDWCs, since the only difference between the MFCCs and MFDWCs is the linear transformation as shown in Figure 1 (discrete cosine transform for the MFCCs and discrete wavelet transform for the MFDWCs). The PMC technique can be applied to the MFDWCs in a similar way that it is applied to the MFCCs. There is little difference between applying the PMC technique to the MFDWCs and MFCCS. DCT and inverse DCT are used when aplying the PMC to the MFCCs. On the other hand, DWT and inverse DWT are used when aplying the PMC to the MFDWCs.

When the noise is stationary, a single state noise model with one mixture may be sufficient to model the noise. When the noise is not stationary, partially stationary (each stationary part may be represented by a mixture), or may not be represented by one mixture component, it may be necessary to use multiple mixtures [21] for the noise model. For example, in the NOISEX-92 database the single mixture model may be enough for speech noise, F16 noise and the Lynx helicopter noise. However, more mixtures are needed to satisfactorily model STITEL and factory noises since they are not stationary, partly stationary, or periodic. STITEL noise is a periodic noise. which may be better modeled by multi-state noise model [7]. However, a multi-state noise model can be approximated with a single-state multiple mixture noise model. In this paper, the noise was modeled by a single state with one mixture component since only stationary noises were used for the experiments. It is common practice to use delta coefficients to achieve better performance. Therefore, we also estimated the delta coefficients [22, 23] using the PMC technique.

4. Experimental Setup and Results

We used the NIST 1998 speaker recognition [24] and NOISEX-92 [25] databases to evaluate and compare the performance of MFDWCs and MFCCs utilizing the PMC technique on a speaker verification task. The NIST 1998 speaker recognition database contains conversational telephone speech signals of 250 male and 250 female speakers sampled at 8 kHz. Only the training and test data of male speakers were used in the experiments. There are three training conditions: one session, two-session, and two-session-full. Two-session full training data were used in the experiments. For each speaker, there are 5 training files with one minute speech in each taken from two different conversations collected from the same phone number for the two-session-full training condition. There are three different test conditions: test segment duration, same/different phone number, and same/different handset type. Only the test data with 30 seconds durations collected from the same phone number using the same handset type were used in the experiments. There are 1308 such speech files for testing in the database. For each test file, there are one trail for the target speaker and nine trails for non-target speakers. Thus, the total number of trails is 13080.

Noise signals from NOISEX-92 database were downsampled from 16 kHz to 8 kHz to have the same sampling rate with the NIST 1998 speaker recognition database. Then F16, Speech, and Lynx noisese were artificially added to the test speech signals (the NIST 1998 speaker recognition database) at SNR levels of -6, 0, 6, 12 dB to obtain noisy speech data.

The speech signal was analyzed with a 32 ms hamming window every 10 ms. The FFT of each frame was used to calculate the power spectrum of the signal. For the computation of mel-scaled log filterbank energies, 26 triangular mel-scaled band-pass filters were designed. The mel-scaled log filterbank energies were interpolated to have 33 mel-scaled log filterbank energies. The folded DWT requires [18] the input vector size to be $2^N + 1$ where N is an integer. In our case $2^N + 1 = 33$. MFCCs were computed by taking the DCT of mel-scaled log filterbank energies. The first sixteen of the MFCCs as well as the zeroth coefficient were used. Our previous experimental results [1, 2] using the TIMIT database have shown that symmetric wavelets give better results than anti-symmetric wavelets. Therefore, a symmetric wavelet was used in this paper. The wavelet used is shown in Figure 2. MFDWCs were computed using the filters associated with wavelet shown in Figure 2 [26]. Eight coefficients at scale four, four coefficients at scale eight, two coefficients at scale sixteen, and one coefficient at scale thirty two and the zeroth coefficient were used. The total number of static coefficients is therefore seventeen. All feature vectors also include delta coefficients.

Each speaker was modeled with a 64 component GMM. The background model was also modeled with a 64 component GMM and trained with all speaker's training data. The silence model is a one-state continuous density HMM. The HTK toolkit [27] was used for training and testing. We conducted a series of experiments under different noise conditions, different noise levels using MFCCs and MFDWCs utilizing the PMC technique.

Table 1 shows EERs for 3 different stationary noise conditions, 4 noise levels using MFCCs and MFDWCs. MFDWCs and MFCCs yielded approximately the same EER for clean speech while the MFDWCs dramatically improved the performance for all noise types and noise levels. The average EERs of the MFCCs and MFDWCs for each noise level are included for an overview of the improvements. The seventh row of Table 1 shows average improvements of MFDWCs over MFCCs, and the last row shows % reduction in EERs. MFDWCs yielded 2.57, 5.87, 7.92 and 6.29 points improvement in average over MFCCs for -6, 0, 6, and 12 dB noises, respectively. These correspond to 6.84%, 20.16%, 36.87% and 38.33% error reductions, respectively.

5. Conclusions

In this paper we investigated use of PMC technique with MFDWCs and MFCCs for text-independent noise robust speaker verification. The PMC technique was applied to the local features (MFDWCs) to take advantage of both noise compensation and local feature for noise robust speaker verification. It was shown that the MFDWCs give better performance than the MFCCs for speaker verification in adverse environments when they are used in conjuction with the PMC technique. The experimental results suggest that conveying local information could be the reason the MFDWCs yielded better results than MFCCs.

Table 1: Equal eror rates for the MFCCs and MFDWCs that both of them utilizing the PMC technique.

Noise Type	MFCCs					**MFDWCs**				
	-6 dB	0 dB	6 dB	12 dB	Clean	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**Clean**
Speech	19.73	15.06	11.62	9.10	5.89	**15.44**	**11.31**	**8.49**	**6.80**	**5.58**
Lynx	16.28	12.31	9.25	7.95	5.89	**14.22**	**9.79**	**7.19**	**6.35**	**5.58**
STITEL	21.48	13.23	9.63	7.49	5.89	**12.54**	**9.40**	**7.80**	**6.65**	**5.58**
F16	21.79	13.84	9.63	8.41	5.89	**16.13**	**11.01**	**8.41**	**6.73**	**5.58**
Average	19.82	13.61	10.03	8.24	5.89	**14.58**	**10.38**	**7.97**	**6.63**	**5.58**
Improvement						**5.24**	**3.23**	**2.06**	**1.61**	**0.31**
% Reduction in EER						**26.44%**	**23.73%**	**20.54%**	**19.54%**	**5.26%**

6. References

[1] Z. Tufekci and J. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in *Proceedings of SoutheastCon*, 2000.

[2] J. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proceedings of ICASSP*, 2000.

[3] J. Campbell, "Speaker recognition: A tutorial," in *Proceedings of IEEE*, September 1997, vol. 85, pp. 1437–1462.

[4] Douglas A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

[5] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, 1995.

[6] M. J. F. Gales and S. J. Young, "An improved approach to the hidden markov model decomposition of speech and noise," in *Proceedings of ICASSP*, March 1992.

[7] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for hmm recognition in noise," *Speech Communication*, vol. 12, pp. 231–240, 1993.

[8] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.

[9] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model compansation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 4, no. 5, September 1996.

[10] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proceedings of ICASSP*, 2002.

[11] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," in *Proceedings of ICSLP*, 1996.

[12] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards asr on partially corrupted speech," in *Proceedings of ICSLP*, 1996.

[13] Z. Tufekci and J. Gowdy, "Subband feature extraction using lapped orthogonal transform for speech recognition," in *Proceedings of ICASSP*, 2001.

[14] S. Vaseghi, N. Harte, and B. Milner, "Multi resolution phonetic/segmental features and models for hmm based speech recognition," in *Proceedings of ICASSP*, 1997.

[15] H. Fletcher, *Speech and Hearing in Communication*, New York: Krieger, 1953.

[16] J. B. Allen, "How do humans process and recognize speech," *IEEE Transactions on Speech, and Audio Processing*, vol. 2, no. 4, October 1994.

[17] Z. Tufekci, J. N. Gowdy, S. Gurbuz, and E. Patterson, "Applying parallel model compensation with mel-frequency discrete wavelet coefficients for noise-robust speech recognition," in *Proceedings of EUROSPEECH*, 2001.

[18] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[19] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, 1995.

[20] M. J. F. Gales and S. J. Young, "A fast and flexible implementation of parallel model combination," in *Proceedings of ICASSP*, 1995, pp. 133–136.

[21] R. Yang and P. Haavisto, "Noise compensation for speech recognition in car noise environments," in *Proceedings of ICASSP*, 1995, pp. 433–436.

[22] M. J. F. Gales and S. J. Young, "Hmm recognition in noise using parallel model combination," in *Proceedings of EUROSPEECH*, 1993, pp. 837–840.

[23] R. Yang and P. Haavisto, "An improved noise compensation algorithm for speech recognition in noise," in *Proceedings of ICASSP*, 1996, pp. 49–52.

[24] NIST, "The 1998 speaker recognition evaluation plan (www.nist.gov/speech/tests/spk/1998/current_plan.htm)," 1998.

[25] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., DRA Speech Research Unit, 1992.

[26] A. Cohen, I. Daubechies, and J. Feauveau, "Biortogonal bases of compactly supported wavelets," *Comm. on Pure and Applied Math.*, vol. 45, 1992.

[27] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory Ltd., version 2.1, 1997.