The Predictive Differential Amplitude Spectrum for Robust Speaker Recognition in Stationary Noises

Jing Deng, Thomas Fang Zheng, Jian Liu, Wenhu Wu

Center for Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, P.R. China
dengj02@mails.tsinghua.edu.cn, fzheng@cst.cs.tsinghua.edu.cn,
liuj@cst.cs.tsinghua.edu.cn, wuwh@tsinghua.edu.cn

Abstract

The performance of any speaker identification system degrades quite seriously when the acoustic conditions for testing mismatch those for training. In this paper, we propose a method to restore clean speech from noisy speech with two steps: 1) a predictive difference function is employed to estimate the differential amplitude spectrums (DAS) from both the left-side and right-side of the amplitude spectrum of the noisy speech, so as to eliminate the noise as precisely as possible, and 2) an average of the left-side and right-side integral DASs is taken as the estimated amplitude spectrum of the original clean speech. The spectrum in the traditional MFCC calculation is then replaced with this estimated amplitude and the extracted features based on this are referred to as predictive differential amplitude spectrum (PDAS) based cepstral coefficients (PDASCCs). We compare PDASCCs with cepstral mean subtraction (CMS) based, spectral subtraction (SS) based, and differential power spectrum (DPS) based cepstral coefficients at different noise levels. Experimental results show that the PDASCCs are more effective in enhancing the robustness of a speaker recognition system, and used with the CMS method the average error rate can be reduced by 7.5%.

1. Introduction

It is well known that the performance of a speaker identification system will degrade very seriously if the acoustic conditions for testing do not match those for training. To date, many methods have been tried so as to overcome the influence of noise and to achieve good recognition accuracy in speaker recognition applications where various types of noise may exist. These methods are typically either at the feature representation level using robust parameterization or at the model level using compensation.

This paper primarily focuses on finding new noise robust features for speaker recognition. In general there are two kinds of principal approaches for noise robust feature extraction. The first one is feature compensation based, such as cepstrum mean subtraction (CMS) and cepstrum normalization (CN) [1,2]. The second is to perform noise suppression in the short-time spectral amplitude domain, such as spectral subtraction (SS) [3,4], nonlinear spectral subtraction [5] and Weiner filter [6,7]. SS is effective for background noise suppression and hence is being widely used in speech recognition and speech enhancement. However, SS will distort the signal and also tend to introduce a perceptually annoying residual noise, which is often referred to as musical noise.

In [8], cepstral features derived from the differential power spectrum (DPS) were proposed for improving the robustness of a speech recognizer in the presence of background noise. The schematic procedure to extract DPS-based cepstral coefficients, which for simplicity is denoted as DPSCC, is showed in Figure 1.

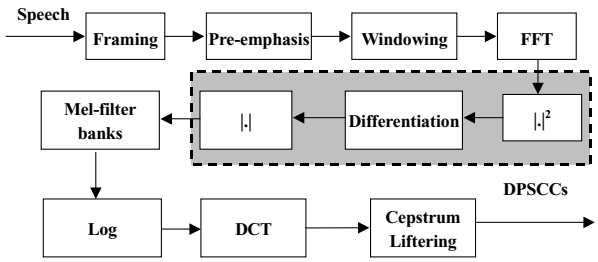


Figure 1: Schematic diagram for DPSCCs extraction

This new cepstrum can be expressed as the superposition of the conventional cepstrum and its nonlinearly liftered counterpart. While a linear liftering transform on cepstral coefficients has no effect on the logarithmic likelihood score, several experiments showed that the nonlinearly liftered counterpart helps the recognizer be more tolerant to noise compared with MFCCs. In our comparison experiments, the differential operation used for DPSCCs is the same as defined in [8] as

$$D(k) = Y(k) - Y(k+1) \tag{1}$$

Because the influence of noise varies in different frequency sub-bands in noisy environments, it is difficult to accurately estimate the differential amplitude spectrum (DAS), especially at low SNRs. Because the differentiation operation will split each peak into two, one positive and one negative, the problem of converting them into cepstral coefficients needs to be solved [8]. In [8], an absolute operation as shown in Figure 1 was proposed to solve this problem after the differentiation operation. But this operation will make the difference between valleys and peaks unclear and may introduce recognition error in noisy environments. Based on this analysis, we propose a predictive difference function with a proposed *sine* filter for a more precise estimation of the DAS from both the left side and the right side. An average of both sides' integral DASs is then taken as the estimated amplitude spectrum of the original clean speech. This operation is proved to be able to make its negative parts positive. By replacing the spectrum in the traditional MFCC calculation with this amplitude spectrum, the newly derived spectrum can then be used to restore clean speech from the

original corrupted speech. Our experimental results show that in stationary noisy environments, the proposed method could effectively improve the robustness of a speaker recognition system, and together with CMS it can reduce the average error rate by 7.5%.

In Section 2, we will describe our proposed method in detail. Experimental results will be given in Section 3. Finally, conclusions are drawn in Section 4.

2. Predictive differential amplitude spectrum (PDAS) based cepstral coefficients

If a speech sequence $s(n)$ is corrupted with an additive noise sequence $b(n)$, the corrupted speech can be expressed as

$$y(n) = s(n) + b(n) \tag{2}$$

If the speech and the noise are assumed to be uncorrelated with each other, we will have in the frequency domain

$$Y(k) = S(k) + B(k), \quad 1 \le k \le N \tag{3}$$

where N is the length of the analyzing frame and is also the FFT size, and $Y(k)$, $S(k)$, and $B(k)$ are the amplitudes of corrupted speech, clean speech, and noise at FFT point k, respectively.

In the calculation of the DAS, we make two assumptions. One is that between two adjacent points of the FFT spectrum, the amplitudes of the noise changes smoothly. The other is that in the frequency domain the amplitude shapes of the clean speech in a short-term region are sinusoid-like. According to the first assumption, the differential amplitudes of the noise at adjacent frequency points are small. According to the second assumption, if the amplitude of a speech signal at FFT point k is lower than that at point k+1 and the amplitude at point k+i is higher than that at point k+i+1, then points k and k+1 are mostly at the left side of a peak, and so on. Here i is a small value between 3 and 6 for the speech signal at 8kHz sample rate and 256-point FFT. Based on these two assumptions, we propose to restore the spectrum of clean speech using the predictive differences of the amplitude spectrum, as follows.

According to the second assumption, if we know the accurate amplitude at points k+i and k+i+1, we can use them to calculate more accurate differential amplitude between points k and k+1 than normal difference functions. A *sine* filter is proposed in this paper to predict the amplitude at FFT point k+i as

$$A'(k) = \max_i \left[Y(k+i) \cdot h(i) \right] \tag{4}$$

In Equation (4) $h(i)$ is the proposed *sine* filter defined as

$$h(i) = \sin\left(\frac{\pi}{2} \cdot \frac{i}{W} \right), \quad 0 \le i \le W \tag{5}$$

where W is the width of the filter. The *sine* filter is shown in Figure 2.

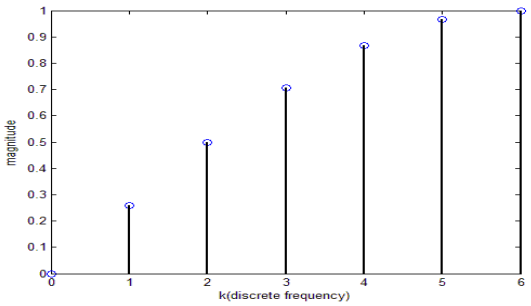


Figure 2: $h(i)$ used in our experiment

Afterwards, the right-side differential amplitude sequence/spectrum is calculated with the following equation

$$D_{right}(k) = \begin{cases} Y(k) - \alpha \cdot Y(k+1), if\ A'(k) > A'(k+1) \\ \qquad\qquad\qquad and\ Y(k) < Y(k+1) \\ Y(k) \cdot \alpha - Y(k+1), if\ A'(k) \le A'(k+1) \\ \qquad\qquad\qquad and\ Y(k) \ge Y(k+1) \\ Y(k) - Y(k+1), otherwise \end{cases} \tag{6}$$

where α is a weight normally ranging inside interval [1.0, 1.10] and was set to 1.05 in our experiment. The first condition in Equation (6) means that FFT points k+i and k+i+1 are on the right side of a peak while points k and k+1 on the left side of the peak. The second condition means that points k+i and k+i+1 are on the right side of a valley while points k and k+1 are on the left side of the valley. The third condition means that points k+i, k+i+1, k and k+1 are all on the same side of a peak or a valley.

Similarly the left-side differential amplitude sequence/spectrum can be calculated with the equation

$$D_{left}(k) = \begin{cases} Y(k) - \alpha \cdot Y(k-1), if\ A'(k) > Y(k-1) \\ \qquad\qquad\qquad and\ Y(k) < Y(k-1) \\ Y(k) \cdot \alpha - Y(k-1), if\ A'(k) \le Y(k-1) \\ \qquad\qquad\qquad and\ Y(k) \ge Y(k-1) \\ Y(k) - Y(k-1), otherwise \end{cases} \tag{7}$$

Likewise, the first condition in Equation (7) means that FFT points k-1 and k are on the left side of a valley while points k+i-1 and k+i are on the right side of the valley; the second condition means that points k-1 and k are on the left side of a peak while points k+i-1 and k+i are on the right side of the peak; and the third condition means that points k-1, k, k+i-1 and k+i are all on the same side of a peak or a valley.

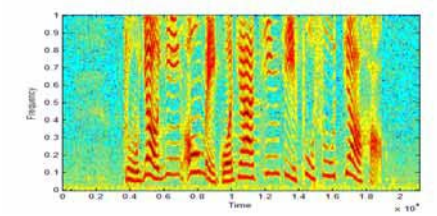
As we know, the spectral peaks convey the most important information in a speech signal. Using Equations (6) and (7) can make the difference at the top of a peak or at the bottom of a valley clearer than using the normal difference functions. With the above defined functions, we propose to restore the amplitude sequence/spectrum of the clean speech by averaging the right-side integral DAS (as defined in Equation (8)) and the left-side integral DAS (as defined in Equation (9)) according to the following equations

$$Y'_{right}(k) = D_{right}(k) + D_{right}(k+1) \tag{8}$$

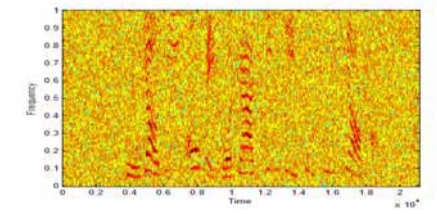
$$Y'_{left}(k) = D_{left}(k) + D_{left}(k-1) \quad (9)$$

$$Y'(k) = \frac{Y'_{left}(k) + Y'_{right}(k)}{2} \quad (10)$$

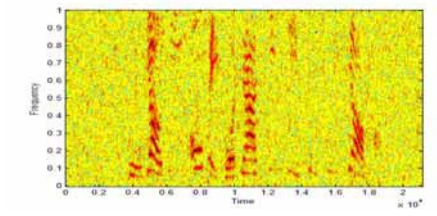
where $D_{right}(N\text{-}1)$, $D_{left}(0)$ and any $Y'(k)$ with a value below 1,000 are set to zero. $\{Y'(k)\colon 1 \le k \le N\}$ is the estimated amplitude sequence/spectrum of the restored speech. The spectrograms of the clean, noisy, and restored speech using the SS method and our method are shown in Figure 3. The SNR of the noisy speech is about 0 dB.



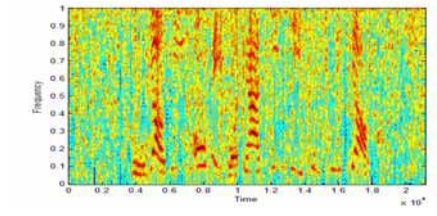
(*a*) Clean speech



(*b*) 0dB noisy speech



(*c*) Restored speech using SS



(*d*) Restored speech using PDAS

Figure 3: Spectrograms of clean, noisy, and restored speech

The amplitude spectrum of the restored speech is then passed into a Mel-frequency filter bank whose outputs are the inputs for the following logarithm operation. Finally, the outputs of the Mel-filter bank are compressed into a feature vector using discrete cosine transform (DCT). We refer to these kinds of extracted features as predictive differential amplitude spectrum (PDAS) based cepstral coefficients, abbreviated as PDASCCs. The schematic procedure to extract PDASCCs is shown in Figure 4. The difference between the proposed PDASCCs and the DPSCCs [8] can be illustrated in the grayed region in Figures 1 and 4.

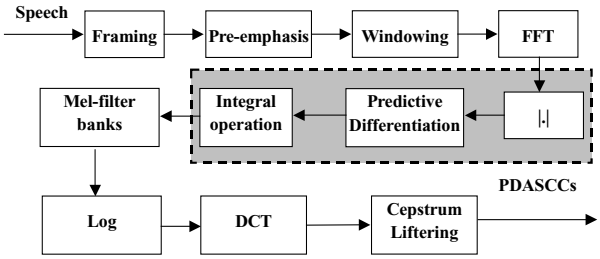


Figure 4: Schematic diagram for PDASCCs extraction

3. Experiments

The database used in our experiments included utterances by 522 speakers (347 males and 175 females), where the speech was recorded in ordinary laboratory environments at 8 kHz sampling rate with 16-bit precision. Three utterances were recorded for each speaker. For each speaker one utterance was used for training while the other two we used for testing. White Gaussian noise was added to each utterance in the testing set at SNRs of 20 to 0 dB every 5 dB. The total length of each utterance used for training was about 32s while for testing the length was about 8s.

The speech was analyzed at a frame size of 24 milliseconds every 12 milliseconds. The pre-emphasis factor was 0.97. Hamming windowing was applied to each pre-emphasized frame. After that, a 256-point FFT was calculated for each frame. A bank of 30 triangular Mel filters was used. The cube root operation and the DCT were performed sequentially, and finally a 16-dimension feature vector was obtained for each frame.

The Gaussian Mixture Model-Universal Background Model (GMM-UBM) system has been proven to be very effective for speaker recognition tasks [9, 10] and thus was used in our experiments. Half an hour of male speech and half an hour of female speech were used to construct a 512 UBM for each of the seven systems.

Table 1: Performance comparisons of different systems on clean and noisy speech (the maximum value in each column is showed in bold font)

System (%)	SNR (dB)						Avg
	Clean	20	15	10	5	0	
MFCCs	94.1	74.7	36.8	14.8	4.8	3.3	38.1
MFCCs+CMS	94.8	86.8	69.2	40.2	15.9	13.6	53.4
MFCCs+SS	93.9	80.8	61.5	31.2	13.6	9.6	48.4
DPSCCs	**96.0**	77.6	43.5	18.2	4.6	2.9	40.5
DPSCCs+CMS	94.6	86.6	70.3	44.4	17.4	**16.5**	55.0
PDASCCs	95.2	84.7	64.6	31.0	13.0	8.2	49.5
PDASCCs+CMS	95.8	**90.0**	**76.0**	**51.2**	**21.3**	15.9	**58.4**
ERR	-4.7	24.7	19.3	12.1	4.6	-6.9	7.5

To evaluate the robustness of different systems, we used 24-seconds of clean, valid speech for deriving each speaker model from the UBM by the Bayesian adaptation method and

3-seconds of valid, noisy speech at different noise levels for testing. The correct percentages of these seven systems and the error rate reductions (ERRs) for the system using PDASCCs+CMS compared with the best one among all other systems at different noise levels are shown in Table 1.

In Table 1, '*Avg*' denotes the average value of recognition accuracies over different noise levels for the corresponding system.

The recognition accuracy comparisons of these seven systems are shown in Figure 5.

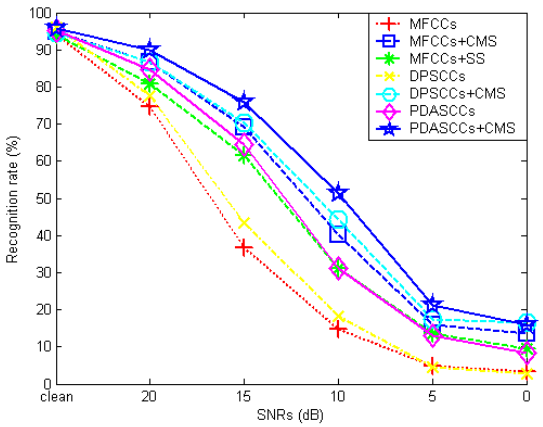


Figure 5: Recognition accuracy comparisons of seven systems

We can see from Table 1 and Figure 5 that the proposed PDASCCs are more robust than DPSCCs and SS based cepstral coefficients at different noise levels. This shows that our method is effective in stationary noise environments.

4. Conclusions

In this paper, predictive differential amplitude spectrum based cepstral coefficients (PDASCCs) for robust speaker recognition in stationary noise environments were proposed. Our experiments showed that the use of PDASCCs could improve the robustness of a speaker recognition system. The average error rate reduction (ERR) for DPASCCs+CMS compared with DPSCCs+CMS is about 7.5%, while for DPASCCs+CMS and MFCCs+CMS at SNRs of 20db, 24.7%.

The PDAS method cannot remove noise completely, but can help improve the SNRs of the noisy speech. The features based on PDAS, i.e. the PDASCCs, outperform the spectral subtraction technique and the cepstral coefficients derived from DPS. We can also see that systems using PDASSCs and CMS can achieve the best performance at different stationary noise levels.

The proposed PDASCCs has been proved effective to enhance the robustness of speaker recognition with stationary noise and so we have reason to assume that it may be effective for non-stationary noisy speech as well, though this needs to be verified by further experiments.

5. References

[1] Rosenberg A.E., Lee C.H., and Soong F.K., "Cepstral Channel Normalization Techniques for HMM-based Speaker Verification", *ICSLP*, 1992.

[2] Viikki O. and Laurila K., "Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature vector Normalization", in ESCA NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France 1997.

[3] Boll S.F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic. Speech, Signal Processing*, ASSP-33, vol.27, pp. 113-120, 1979.

[4] Hirsch H.G. and Ehrlicher C., "Noise Estimation Techniques for Robust Speech Recognition", *ICASSP* 1995, pp. 153-156.

[5] Bcrouti M., Schwartz R., Makhoul J., "Enhancement of speech corrupted by additive noise," P*roceedings of the IEEE Conference on Acoustics. Speech. and Signal Processing*, pp. 208-211, April 1979.

[6] Lim J. S. and Oppenheim A. V., "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 3, pp. 197-210, June 1978

[7] Moon S. Y. and Hwang J. N., "Noisy speech recognition via wavelet coefficient enhancement," in *Proc. IEEE 26th Asilomar Conf. Signals, Syst., Comput*., Monterey, CA, Oct. 1992, pp. 1086-1090.

[8] Chen J. D., Paliwal K. K. and Nakamura S., "Cepstrum derived from differentiated power spectrum for robust speech recognition", *Speech Communication* 41 (2003), pp. 469-484.

[9] Reynolds D. A., Quatieri T. F. and Dunn R. B., "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing* 10 (2000), pp. 19-41.

[10] Reynolds D. A., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication* 17 (1995), pp. 91-108.