Channel Robust MFCCs for Continuous Speech Speaker Recognition

Sharada Vikram Chougule and Mahesh S. Chavan

Abstract. Over the years, MFCC (Mel Frequency Cepstral Coefficients), has been used as a standard acoustic feature set for speech and speaker recognition. The models derived from these features gives optimum performance in terms of recognition of speakers for the same training and testing conditions. But mismatch between training and testing conditions and type of channel used for creating speaker model, drastically drops the performance of speaker recognition system. In this experimental research, the performance of MFCCs for closed-set text independent speaker recognition is studied under different training and testing conditions. M*agnitude spectral subtraction* is used to estimate magnitude spectrum of clean speech from additive noise magnitude. The mel-warped cepstral coefficients are then normalized by taking their mean, referred as *cepstral mean normalization* used to reduce the effect of convolution noise created due to change in channel between training and testing. The performance of this modified MFCCs, have been tested using *Multi-speaker continuous (Hindi) speech database* (By Department of Information Technology, Government of India). Use of *improved MFCC* as compared to conventional MFCC perk up the speaker recognition performance drastically.

Keywords: Text independent speaker recognition, MFCC, magnitude spectral subtraction, cepstral mean normalization.

Sharada Vikram Chougule
Department of Electronics and Telecommunication Engineering,
Finolex Academy of Management and Technology, Ratnagiri
Maharashtra, India

Mahesh S. Chavan
Department of Electronics Engineering,
KIT's College of Engineering, Kolhapur
Maharashtra, India

S.M. Thampi, A. Gelbukh, and J. Mukhopadhyay (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 264,
DOI: 10.1007/ 978-3-319-04960-1_48, © Springer International Publishing Switzerland 2014

1 Introduction

The largest challenge to use speaker recognition technology is the channel variability, which refers to changes in channel effects between enrolment and successive recognition (verification/identification).This mismatch between training and testing, greatly degrades the performance of automatic speaker recognition systems (e.g.[1],[2]). The most widely used speech recognition features are the Mel Frequency Cepstrum Coefficients (MFCCs), which are also used for speaker recognition. The wide-spread use of the MFCCs is due to the low complexity of the estimation algorithm and their good performance for automatic speech and speaker recognition tasks under clean and matched conditions [3],[6]. However, MFCCs are easily affected by common frequency localized random perturbations, to which human perception is largely insensitive. MFCC's lack of robustness in noisy or mismatched conditions have led many researchers to investigate robust variants of MFCCs. Studies in [5] had shown that estimation of both vocal source and vocal track related features were extracted by denoising the speech and using MFCC with wavelet octave coefficients of residues (WOCOR). Furthermore, dynamic cepstral features such as delta and delta-delta cepstral have been shown to play an essential role in capturing the transitional characteristics of the speech signal. So, delta MFCC, delta-delta MFCC, and other related features such as delta cepstral energy (DCE) and delta-delta cepstral energy (DDCE) are also has been introduced into the speaker recognition systems [7]. Also several acoustic features, like MFCC, LPCC, PLP are admired and extensions of these (frequency-constrained LPCC, LFCC) and new features called PYKFEC [6],[8],[9] are evaluated over on the different conditions and measured their respective contribution to feature fusion.

In this work, we design a front-end that is motivated from auditory perception. The speech signal is improved by pre-processing, done with the help of *speech activity detection* for detecting speech portion in continuous speech and also to determine voiced and unvoiced part of the speech. M*agnitude spectral subtraction* is used to estimate magnitude spectrum of clean speech from noisy speech magnitude and *cepstral mean normalization* to reduce convolution distortion in slowly varying channel.

The organization of this paper is as follows: in Section 2, we provide the theoretical background of speech pre-processing. In Section 3, the proposed improved feature extraction algorithm using MFCC is presented. In Section 4, the performance of the improved features is evaluated under different recording conditions in terms of recognition. Conclusions are presented in Section 5.

2 Speech Pre-processing

As features related to speech as well as speaker are present in spectral content, it is desired to have input speech to the recognition system to be as clean as possible. This is especially required for different recording devices and transmission channels. Therefore we have pre-processed the speech using speech activity detection and pre-emphasis filtering. The pre-processing of speech before actual feature extraction will help to eliminate some part of noise as well as raw speech data.

2.1 Speech Activity Detection (SAD)

The fundamental problem in many speech and speaker recognition systems is to separate our speech signal from non-speech part such as silence and various types of noise and disturbances. Speech activity detection is an algorithm used in speech processing wherein, the presence or absence of human speech is detected from the audio samples. The primary function of it is to provide an indication of speech presence in order to facilitate speech processing as well as possibly providing delimiters for the beginning and end of a speech segment. As our database for training and testing consists of continuous speech, there is possibility of voice inactive or silence segments. We have used SAD to acquire a speech segment, eliminating non-speech part such as silence and noise. Thus it is possible to distinguish speech and silence/noise and to get feature vectors better representing true speech characteristics.

The discriminative characteristics of the speech can be extracted in time domain, spectral domain or cepstral domain. As signal energy remains the basic of the feature vector, we have used Energy-based SAD [14]. Here it assumed that speech is louder than silence and background noise. Therefore we can assign high energy frames as speech, whereas low energy frames as silence or noise. Speech is detected when the energy estimation lies over the threshold. Short-time energy is used to distinguish voiced speech and zero-crossing rate is used to distinguish unvoiced part. We set two constant thresholds in SAD. If achieve the higher one, decide voiced speech. If between the higher threshold and the lower one, calculate the zero-crossing rate to decide whether it is unvoiced speech or noise. A problem here is obvious: If the input SNR is low (small speech amplitude), after the normalization of SAD, it is more likely to decide noise as speech since it has a relative large energy. Thus, to choose proper decision thresholds and keep large SNR are very important. We get the best by trying and adjusting.

As shown in figure (1), the combination of energy and zero crossing rate is used to distinguish voiced part and unvoiced part of the speech signal. It is observed that energy is high for voiced part and low for unvoiced part, whereas zero crossing rates are low for voiced part and high for unvoiced part. Thus, a burst of energy in a stipulated time is used to recognize a voiced speech whereas based on assumption that zero crossing rate of speech and noise are different, it is possible to distinguish speech and noise.

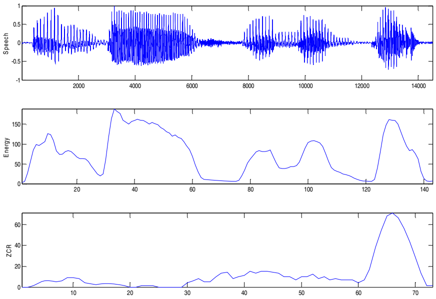

Fig. 1 Energy and zero-crossing rate of speech signal in speech activity detector

2.2 *Pre-emphasis*

In next step, the speech signal is emphasized using a highpass filter. Speech spectrum has more energy at low frequencies compared to high frequencies. This is due to nature of glottal pulse. This is called as *spectral tilt*. Pre-emphasis boosts the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.

The speech signal s(n) is sent to a high-pass filter:

$$y(n) = s(n) - a * s(n-1) \tag{1}$$

where *s(n)* is the output signal and the value of a is usually between 0.95 and 0.97. The z-transform of the filter is:

$$H(z) = 1 - a * z^{-1} \tag{2}$$

which is an FIR highpass filter. Here we choose a=0.97.

3 Improved MFCC

Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction technique used in speech and speaker recognition systems. The technique is so-called *FFT-based,* which means that feature vectors are extracted from the frequency spectra of the windowed speech frames. MFCCs can be considered as: (a) as a filter-bank processing adapted to speech specificities and (b) as a modification of the conventional cepstrum, a well known deconvolution technique based on homomorphic processing.

The following section discusses various steps to get channel robust MFCCs.

3.1 Framing and Windowing

Though the speech signal is constantly changing, it is assumed to have quasi-stationary spectral characteristics over short time interval. Therefore, speech signal is processed in small chunks called frames. Framing divides the speech signal to get piecewise stationarity. The purpose of windowing is to limit the time interval to be analyzed so that the properties of the waveform do not change appreciably. For this, speech is usually segmented in frames of 20 to 30 msec. A typical frame overlap is around 30 to 50 % of the frame size. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered at some frames. Simply cutting out speech signal into frames is equivalent to using rectangular window. But as rectangular window causes discontinuities at the edges of the segments, smooth tapers (like Hamming or Hanning) are usually used. Thus each frame is multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame.

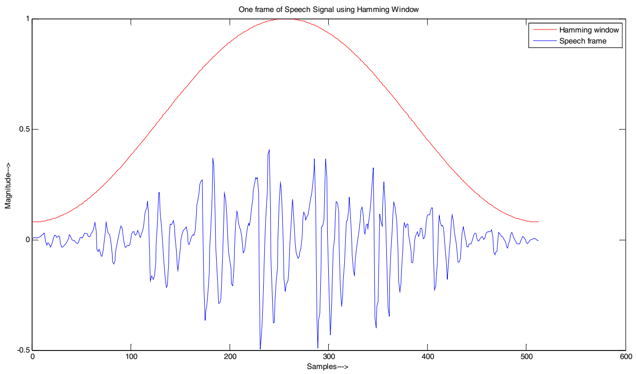

Fig. 2 One frame of voiced speech multiplied by Hamming window

3.2 Spectral Analysis

Spectral analysis is required to determine the spectral contents of the speech signal from a finite time samples obtained through framing and windowing. It gives the knowledge of distribution of power over frequency. Short time fourier transform (STFT) is a tool most widely used for speech signal analysis. Given the time series of speech signal *s(n)*, the STFT at time *n* is given as:

$$S(n,w) = \sum_{m=-\infty}^{\infty} s(m)w(m-n)e^{-jwm} \tag{3}$$

where *w(n)* is the analysis window (of length N), which is assumed to be non-zero only in the interval 0 to N-1. The estimated power spectrum contains details of spectral shape as well as spectral fine components. Here the length of frame/window decides the resolution in time and frequency domain. A short frame width gives high time resolution and low frequency resolution, whereas long frame width gives low time resolution but high frequency resolution. A small window length (results in wider bandwidth) can capture fast time varying components (e.g. in rapid conversational speech), whereas a longer window length (narrow bandwidth) gives better information about sinusoidal components (e.g. harmonics of formants). Thus, we can say that short time window (approximately 5-10 msec) will represent vocal fold details (source information) whereas longer duration window (approximately 20-30 msec) gives vocal track details (filter characteristics) considering source-filter model of human speech production mechanism.

3.3 Spectral Subtraction (SS)

The magnitude or power estimate obtained with STFT is susceptible to various types of additive noise (such as background noise). To compensate for additive noise and to restore the magnitude or power spectrum of speech signal, spectral subtraction is used. Magnitude of the spectrum over short duration (equal to frame length) is obtained eliminating phase information. Here spectrum of noise is subtracted from noisy speech spectrum, therefore the name spectral subtraction. For this, noise spectrum is estimated and updated over the periods when signal is absent and only noise is present [18]. Thus speech signal is enhanced by eliminating noise.

In case of additive noise, we may write the noise contaminated speech signal as :

$$y(i) = s(i) + n(i) \tag{4}$$

where n(i) is some noise signal.

In frequency domain (considering linear operation) we can write,

$$Y(k) = S(k) + N(k) \tag{5}$$

Assuming n(i) with zero mean and is uncorrelated with x(i), the power spectrum of y(i) can be written as:

$$|Y(k)|^2 = |X(k)|^2 + |N(k)|^2 \tag{6}$$

Using (6), it is possible to estimate $|X(k)|^2$ as:

$$|\hat{X}(k)|^2 = |Y(k)|^2 - |\hat{N}(k)|^2 \tag{7}$$

where $|\hat{N}(k)|^2$ is some estimate of noise.

One way to estimate this noise is to average $|Y(k)|^2$ over the sequence of frames known to be non-speech (by using speech activity detection):

$$|\hat{N}(k)|^2 = \frac{1}{M}\sum_{t=0}^{M-1} |Y_t(k)|^2 \tag{8}$$

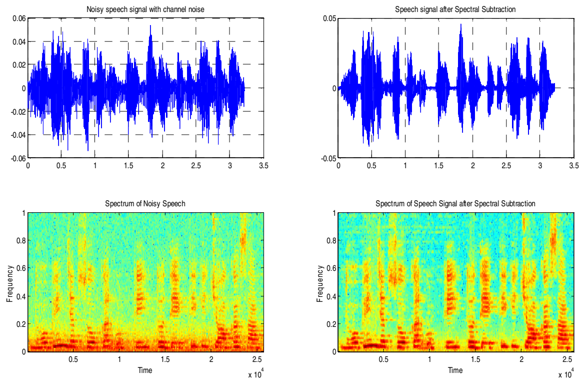

Fig. 3 Effect of Spectral Subtraction on noisy speech signal

As observed in figure (3), the noisy speech signal (contaminated by background noise) processed with magnitude spectral subtraction improves the spectral content of the speech signal by removing noise. This helps to extracts the true features of the speaker from his/her speech.

3.4 Mel-scale Bank and Cepstral Analysis

The mel scale is based on an empirical study of the human perceived pitch or frequency. The scale is divided into the units of "mel" s. The mel scale is generally speaking a linear mapping below 1000 Hz and logarithmically spaced above that frequency[3], [6]. The mel frequency warping is most conveniently done by utilizing filter bank with filters centered according to mel frequencies. The width of the triangular filters vary according to the mel scale, so that the log total energy in a critical band around the centre frequency is included. The result after warping is a number of coefficients Y(k):

$$Y(\boldsymbol{k}) = \sum_{\boldsymbol{j}=1}^{\frac{\boldsymbol{N}}{2}} \boldsymbol{S}(\boldsymbol{j})\, \boldsymbol{H}_{\boldsymbol{k}}(\boldsymbol{j}) \tag{9}$$

Using N point IDFT, the cepstral coefficients are calculated by transforming log of the quefrency domain coefficients to the frequency domain as:

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(k) e^{j\frac{k2\pi}{N}n} \tag{10}$$

which can be simplified, because Y(k) is real and symmetric about N/2, by replacing the exponential by a cosine:

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(k) \cos\left(k \frac{2\pi}{N} n\right) \tag{11}$$

A reliable way of obtaining an estimate of the dominant fundamental frequency for long, clean, stationary speech signals is to use the *cepstrum*. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: i.e. the fundamental frequency.

3.5 Normalization of Cepstral Coefficients

Cepstral mean normalization (CMN) is an alternate way to high-pass filter cepstral coefficients. In cepstral mean normalization the mean of the cepstral vectors *c(n),* is subtracted from the cepstral coefficients of that utterance on a sentence-by-sentence basis:

$$y(n) = c(n) - \frac{1}{N} \sum_{n=1}^{N} c(n) \tag{12}$$

Here we try to enhance the characteristics present in speech signal and reduce the channel effects on speech signal by computing the mean over finite number of frames. To compensate the channel effect, the channel cepstrum can be removed by subtraction of the cepstral mean. This temporal mean is a rough estimate of the channel response.

4 Experiments and Results

The proposed improved MFCC coefficients extracts the features on frame by frame basis. We use the standard MFCCs as the baseline features. Speech pre-processing is performed before framing. Each frame is multiplied with a 30 ms Hamming window, shifted by 20 msec. From the windowed frame, FFT is computed, and the magnitude spectrum is subtracted using MMSE (Maximum Mean Square Error) algorithm. These samples are filtered with a bank of 13 triangular filters spaced linearly on the mel-scale. The log-compressed filter outputs are converted into cepstral coefficients by DCT, and the 0th cepstral coefficient is ignored. The cepstral coefficients thus obtained are normalized by cepstral mean technique discussed earlier. Further speaker models are generated by the LBG/VQ clustering algorithm [16]. The quantization distortion with Euclidean distance is used as the matching function. The number of MFCCs and model sizes were fixed to 12 and 64, respectively. The effect of the number of MFCCs was also studied. Increasing the number of coefficients improved the identification accuracy up to 10–15 coefficients, after which the error rates stabilized. Therefore, we fixed the number of coefficients to 12.

4.1 Database

Speaker recognition (identification) experiments have been conducted to test the performance of the proposed algorithm. For performance evaluation, we have used the *multi-speaker, continuous (Hindi) speech database* generated by TIFR, Mumbai (India) and made available by Department of Information Technology, Government of India. The database contains a total of approximately 1000 Hindi sentences, a set of 10 sentences read by each of 100 speakers. These 100 sets of sentences were designed such that each set is `phonetically rich' [17]. The speech data was simultaneously recorded using two microphones: one good quality, close-talking, directional microphone and another desk-mounted Omni-directional microphone.

4.2 Database Set and Performance

(I) Database Set I- Continuous Speech Hindi Database (100 speakers, 59 Male and 41 female speakers)

***Training Set*- Recorded with close-talking, directional microphone**
***Testing Set* - Recorded with close-talking, directional microphone**

Table 1 Identification Results of Speaker Recognition with baseline MFCC

Feature Extraction Technique	**Training Database**	**Testing Database**	**Result (%) Identification**
Baseline MFCC	Phonetically Rich	Phonetically Rich	100
Baseline MFCC	Phonetically Rich	Broad Acoustic Class of phonemes in different phonetic contexts	99
Baseline MFCC	Broad Acoustic Class of phonemes in different phonetic contexts	Broad Acoustic Class of phonemes in different phonetic contexts	88
Baseline MFCC	Broad Acoustic Class of phonemes in different phonetic contexts	Phonetically Rich	98

(II) Dataset-II

Continuous Speech Hindi Database (97 speakers, 59 Male and 38 female)
***Training Set*- Recorded with close-talking, directional microphone**
***Testing Set* - Recorded with desk-mounted Omni-directional microphone**

Table 2 Identification Results of Speaker Recognition with modified MFCCs under phonetically rich condition

Feature Extraction Technique	Training Database	Testing Database	Result (%) Identification
Baseline MFCC	Phonetically Rich	Phonetically Rich	15.46
Baseline MFCC with CMN	Phonetically Rich	Phonetically Rich	52.57
Baseline MFCC with SS	Phonetically Rich	Phonetically Rich	60.82
Baseline MFCC with CMN & SS	Phonetically Rich	Phonetically Rich	88.65

Table 3 Identification Results of Speaker Recognition with modified MFCCs under phonetically mismatched condition

Feature Extraction Technique	Training Database	Testing Database	Result (%) Identification
Baseline MFCC	Phonetically Rich	Broad Acoustic class of phonemes	9.27
Baseline MFCC with CMN	Phonetically Rich	Broad Acoustic class of phonemes	32.61
Baseline MFCC with SS	Phonetically Rich	Broad Acoustic class of phonemes	44.33
Baseline MFCC with CMN & SS	Phonetically Rich	Broad Acoustic class of phonemes	74.22

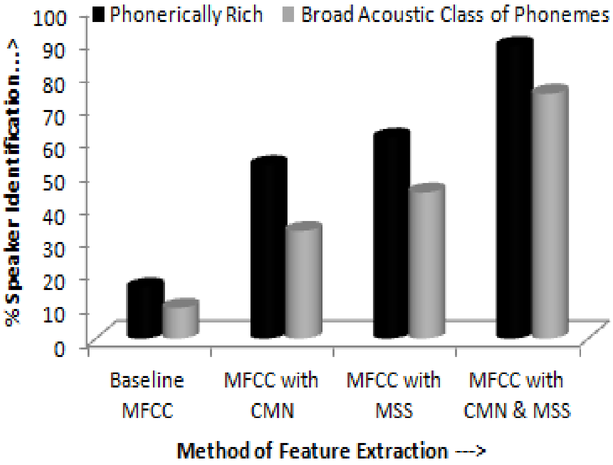

Fig. 4 Comparison of Robustness of MFCCs for Speaker Identification

5 Conclusion

In this work, we have evaluated the performance of baseline MFCCs as well as modified MFCCs under two different recording conditions, one recorded with close-talking, directional microphone and other recorded with desk-mounted Omni-directional microphone. The baseline MFCC gives optimum recognition performance for same training and testing conditions, recorded with close-tracking directional microphone (clean speech). Whereas the speaker recognition performance totally fall down (100 % to 15.46 %) under different training and testing conditions, which proves that MFCC is very susceptible to mismatched conditions.

It is observed that, percentage correct identification rate of the system is improved by modifying MFCCs with magnitude spectral subtraction and cepstral mean normalization. The combination of spectral subtraction and cepstral mean normalization compensates the adverse effect of channel mismatch. Modified MFCCs improves the performance of speaker recognition system from 15.46 % to 79.38 % when both training and testing data is phonetically rich and 9.27% to 74.22 % when under acoustically mismatched training and testing data. Experimental results demonstrate the effectiveness and robustness of the modified MFCCs as compared to regular MFCCs in different recording conditions. From results it is observed that the nature of speech (clean or noisy) and characteristics of speech (phonetical contents) also plays an important role for extracting true speaker related features in individual's speech.

Acknowledgements. The authors would like to thank Department of Information Technology, Government of India and Dr. Samudravijaya K. of TIFR, Mumbai for providing Speech Database. Special thanks to Dr. K.L.Asanare (Director, FAMT Ratnagiri), for his constant support and motivation.

References

1. Shao, Y., Wang, D.: Robust speaker identification using auditory features and computational auditory scene analysis. In: ICASSP. IEEE (2008)
2. Mammone, R.J., Zhang, X., Ramachandran, R.P.: Robust Speaker Recognition: A Feature based approach. In: IEEE Signal Processing Magazine (September 1996)
3. Rabiner, L., Schafer, R.: Digital Processing of Speech Signal. Prentice Hall, Inc., Englewood Cliffs (1978)
4. Hermansky, H.: Perceptual linear predictive (PLP) analysis for speech. J. Acoust. Soc. Am., 1738–1752 (1990)
5. Wang, N., Ching, P.C.: Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features. IEEE Transactions on Audio, Speech, and Language Processing 19(1) (January 2011)
6. Kinnunen, T., Li, H.: An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. In: Speech Communication (2010)
7. Nosratighods, M., Ambikairajah, E., Epps, J.: Speaker Verification Using A Novel Set of Dynamic Features. In: Pattern Recognition, ICPR 2006 (2006)
8. Openshaw, J., Sun, Z., Mason, J.: A comparison of composite features under degraded speech in speaker recognition. In: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 371–374 (1993)
9. Reynolds, D., Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing 3 (January 1995)
10. Campbell Jr., J.P.: Speaker Recognition- A Tutorial. Proceedings of The IEEE 85(9), 1437–1462 (1997)
11. Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., Stauffer, A.: Survey and Evaluation of Acoustic Features for Speaker Recognition. In: ICASSP 2011. IEEE (2011)
12. Reynolds, D.A.: An Overview of Automatic Speaker Recognition Technology. In: ICASSP 2001. IEEE (2001)
13. Glsh, H., Schmidt, M.: Text Independent Speaker Identification. In: IEEE Signal Processing Magazine (1994)
14. Prasad, V., Sangwan, R., et al.: Comparison of voice activity detection algorithms for VoIP. In: Proc. of the Seventh International Symposium on Computers and Communications, Taormina, Italy, pp. 530–532 (2002)
15. Menéndez-Pidal, X., Chan, R., Wu, D., Tanaka, M.: Compensation of channel and noise distortions combining normalization and speech enhancement techniques. Speech Communication 34, 115–126 (2001)
16. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Trans. Commun. 28(1), 84–95 (1980)
17. Samudravijaya, K., Ra0, P.V.S., Agrawal, S.S.: Hindi Speech Database. In: Proceedings of International Conference on Spoken Language Processing, China (2000)
18. Vaseghi, S.V.: Advanced Digital Signal Processing and Noise Reduction, 2nd edn. John Wiley & Sons Ltd. (2000)