Development of a Singer Identification System Using a Spectral Subtraction Technique

Jian-Da Wu, Chen-Wei Chung
Graduate Institute of Vehicle Engineering, National Changhua University of Education
Email:jdwu@cc.ncue.edu.tw

Abstract: This paper presents a study of a singer identification system using the signal spectral subtraction enhancement technique and artificial neural network. Signal spectral subtraction enhancement is a signal processing technique widely used for eliminating the background noise from sound signals. In the present study, the spectral subtraction technique is applied to the singer signal enhancement with background music. To compare the proposed speech enhancement algorithm, spectral subtraction, multi-band spectral subtraction and non-linear spectral subtraction are used and compared in this experimental investigation. In the signal classification stage, the post-processing features of signals are used as input information to a generalized regression neural network (GRNN) classifier for singer identification. The experimental result indicates the proposed singer identification is effective and has satisfactory performance.

Keywords: Spectral subtraction, Singer identification, Generalized regression neural network.

I. INTRODUCTION

With the progress of information techniques, digital signal processing has gradually become an important area of information application. Digital music has also become popular because of the general public's liking for CDs and MP3s that can be downloaded from the internet. However, the amount of music data has increased and the types of songs have become more varied. The sheer range of songs has made it more difficult for listeners to find the music they want. Therefore, music catalog search has become very important. Commonly used classification methods include the melody [1], instruments [2] and genre [3]. In recent years, some luxury or intelligent vehicles have gradually joined the music search function in a vehicle multimedia system. Simplifying the instrument panel and interface to rapidly search for music are important goals. Reducing the need for touch panels improves driving safety. Thus, a singer identification system is proposed in the present study.

The song tracks facilitate fast retrieval and summary. Suppose a certain singer released a new album just a few steps can be completed quickly category song. However, speech recognition is not entirely flawless. The impact of background noise is the biggest factor in speech recognition rates. As the noise level increases, the recognition rate decreases [4]. In the previous study, it is assumed there was a no noise environment and a pure voice signal was analyzed [5]. However, in reality, the effect of background noise is very significant. In communication systems, echo and noise affect communication quality. Echo cancellation using adaptive filters can improve communication quality. When there is echo cancellation, background noise must also be considered. Adaptive filter algorithms, such as the Kalman filter [6], Wiener filter [7], and least-mean-square (LMS) algorithm [8] have been proposed for echo cancellation.

The LMS algorithm is a well-known adaptive algorithm because of the significant feature of its adaptive property, which is important in practical applications. The LMS algorithm is simple, has an easy to follow process, and requires less calculation. It is widely used in many applications. Unfortunately, the LMS algorithm in the time domain has slow convergence. However, spectral subtraction in the frequency domain is an effective method for speech enhancement. The spectral subtraction algorithm is simple, rapid, requires less calculation, and obtains a higher signal noise ratio (SNR). However, the process of spectral subtraction will produce a music background noise problem. To eliminate the music background noise problem, many studies have proposed various algorithms such as multi-band spectral subtraction [9], non-linear spectral subtraction [10]. Spectral subtraction is a simple, effective and quick. For the above reasons, this method is adopted in the present study.

Speech is a time-varying signal where the frequency changes with time. In traditional techniques, the sound features are usually obtained by fast Fourier transform (FFT) or short time Fourier transform (STFT) [11]. However, these methods lose some information in the time domain while the signals are converted into the frequency domain. Research has been done using the wavelet decomposition method to analyze speech signals [12]. The speech feature parameters can be energy, fundamental frequency and resonance peak. Present in speech recognition, often extracting acoustic feature methods are linear prediction cepstrum coefficients (LPCC)[13] and Mel-frequency cepstrum coefficients (MFCC) [14]. LPCC and MFCC are signals from the time domain used in frequency domain analysis. MFCC simulates the human ear's auditory model. Speech output through the filter transforms it into acoustic characteristics. MFCC has two advantages compared with LPCC. First, voice signals are mostly concentrated in the low-frequency part. This is useful because high-frequency voice signals pick up interference too easily. MFCC emphasize low-frequency voice messages and mask noise. Second, MFCC does not require any assumptions, so it can be used in every kind of environment. Therefore, this method is used to extract the speech signal characteristics.

The neural network for recognition has been reported in recent years in many applications. It has been widely applied in data analysis and signal classification. Neural architecture stems from the understanding of the human nervous system. The current neural network is composed of many non-linear operations units. These units are usually based on parallel and distributed design for computing. It can process a huge amount of information at the same time. In the present study, the general regression neural network (GRNN) is used in sound signal classification. GRNN has some advantages such as rapid learning, stability, and a few artificially selected parameters. In the following sections, the proposed methods and performance of the singer identification system will be described.

II. PRINCIPLE OF SPEECH ENHANCEMENT ALGORITHM

A. Spectral Subtraction-SS

Spectral subtraction is a noise reduction technique used in signal processing. Since Boll [15] first proposed this method, many different variations of spectral subtraction have been proposed [16]. Assuming *y(t)* is a noisy input signal which includes the clean speech signal *s(t)* and uncorrelated additive noisy signal *d(t)*, the resulting corrupted speech can be expressed as:

$$y(t)=s(t)+d(t) \tag{1}$$

The power spectrum of the noisy signal can be approximately written as:

$$|y(k)|^2 \approx |s(k)|^2 + |d(k)|^2 \tag{2}$$

Because the additive noisy spectrum *d(k)* cannot be directly obtained, a time average of the power spectrum $\hat{d}(k)$ is calculated during periods of silence. The modified speech spectrum can be written as:

$$|\hat{s}(k)|^2 \approx |y(k)|^2 - |\hat{d}(k)|^2 \tag{3}$$

An important variation of spectral subtraction was proposed by Berouti et.al [17]. To minimize the residual music noise, this proposed could be expressed as:

$$|\hat{s}(k)|^2 = |y(k)|^2 - \alpha|\hat{d}(k)|^2 \tag{4}$$

$$|\hat{s}(k)|^2 = \begin{cases} |\hat{s}(k)|^2 & if \quad |\hat{s}(k)|^2 > \beta|\hat{d}(k)|^2 \\ \beta|\hat{d}(k)|^2 & else \end{cases} \tag{5}$$

where α is an over-subtraction factor function of the noisy signal-to-noise (NSNR) and calculated as:

$$\alpha = \alpha_0 - 3/20\, SNR \qquad -5dB \le SNR \le 20dB \tag{6}$$

In Eq.(5), the spectral floor β prevents the spectral components of the enhanced spectrum from falling below the lower value, $\beta|\hat{d}(k)|^2$.

B. Multi-Band Spectral Subtraction-MBSS

In real environments, the noise spectral is not uniform for all frequencies. This account for the fact colored noise affects the speech spectrum differently at different frequencies. According to Eq.(4), the speech spectrum is divided into *N* non-overlapping bands. The estimate of the speech spectrum in the i^{th} band is obtained by

$$|\hat{s}_i(k)|^2 = |y_i(k)|^2 - \alpha_i|\hat{d}_i(k)|^2 \tag{7}$$

The over-subtraction factor α_i is a function of the segmental $NSNR_i$ of the i th frequency band.

$$SNR_i(dB) = 10\log\left(\frac{\sum_{k=b_i}^{e_i}|y_i(k)|^2}{\sum_{k=b_i}^{e_i}|\hat{d}_i(k)|^2}\right) \tag{8}$$

where b_i and e_i are the beginning and ending frequency bins of the i^{th} frequency band. The over-subtraction factor α_i may be calculated as

$$\alpha_i = \begin{cases} 4.75 & SNR_i < -5 \\ \alpha_0 - \frac{3}{20} NSNR_i & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \tag{9}$$

where $\alpha_0 = 4$ is the desired value at 0 dB $NSNR_i$ and floor parameter was set to β=0.002.

C. Non-linear Spectral Subtraction-NSS

Non-linear spectral subtraction can be expressed in terms of a filter operation,

$$|\hat{s}_i(k)| = H_i(k)|y_i(k)| \tag{10}$$

where $H_i(k)$ depends on a smoothed estimate of the noisy speech magnitude spectrum $|\ddot{y}_i(k)|$, and nonlinear subtraction term, $\Phi_i(k)$,

$$H_i(k) = \frac{|\ddot{y}_i(k)| - |\Phi_i(k)|}{|\ddot{y}_i(k)|} \tag{11}$$

The subtraction term, $\Phi_i(k)$ is given by

$$\Phi_i(k) = \frac{\max_{i-40\le\tau\le i}|d_\tau(k)|}{1+\gamma\rho_i(k)} \tag{12}$$

where $\rho_i(k) = \frac{|\ddot{y}(k)|}{|\ddot{d}_i(k)|}$, γ is a constant scaling factor dependent on the range of $\rho_i(k)$. For practical purposes, the dynamic

range of $\Phi_i(k)$ is limited from 1 to 3 times the smoothed noise magnitude estimate (i.e., $|\ddot{d}_i(k)| \le \Phi_i(k) \le 3|\ddot{d}(k)|$) and a noise-flooring operation is utilized.

III. PRINCIPLE OF VOICE SIGNAL FEATURE EXTRACTION

A. *Mel-frequency cepstrum coefficients and normalization*

In speech recognition, linear prediction cepstrum coefficients and Mel-frequency cepstrum coefficients are both acoustic feature methods in many applications The MFCC is the most popular method for speaker identification and is often used to extract the acoustic characteristics of a person. In recent years, MFCC has also been used for speech recognition [18, 19]. A diagram of the MFCC design is shown in Fig. 1. The Mel filter banks (MFB) that primarily within the Basiler Menbrane in a simulated human ear to listen to learn to stimulate nerve transfer process. This is an important part of MFCC. The present study uses 39 dimensions of MFCC coefficients which contain 12 MFCCs, the 1st log energy, and the 1st and 2nd differential. The results indicated M represents the dimensions and N is the coefficients for each dimension. Details of the parameter settings are shown in Table 1. In the experiment, the recording volume level and recording time affect the identification results. The normalization post-processing is implemented. The calculation of the post-processing procedure is as follows:

$$\|M\| = \sqrt{A_i^2} \qquad i = 1 \ldots 39 \tag{13}$$

$$avg = \frac{\sum_{i=1}^{n} \|M\|}{39} \tag{14}$$

$$std = \left[\frac{1}{n} \sum_{1}^{n} (A_i - \bar{A})^2 \right]^{1/2} \quad where \quad \bar{A} = \frac{1}{n} \sum_{1}^{n} A_i \tag{15}$$

$$s = \frac{\sum_{i=1}^{n} \|M\| - avg}{std} \tag{16}$$

The above post-processing procedure A_i is a vector of each dimension, and *avg* and *std* are the mean and standard deviation, respectively. To understand the differences between different singers, the same condition is assumed to analyze the voice of three singers. Fig.2 shows the major difference between the characteristics of the three singers from 3 to 8 dimensions. Only three dimensional feature vectors are required for the calculation. Therefore, less feature vector space is required and the amount of computation is less.

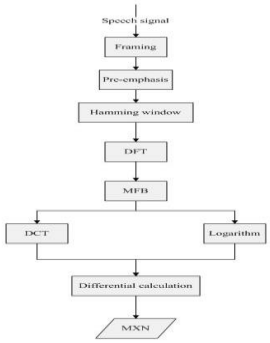
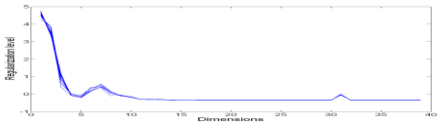
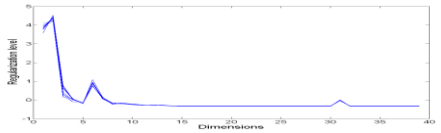


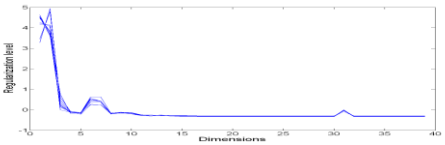
Fig 1. Mel-frequency cepstrum coefficients block diagram



(a)



(b)



(c)

Fig 2. Difference in the MFCC feature of the three singers: (a) Singer 1 (b) Singer 2 (c) Singer 3

Table 1. Mel-frequency cepstrum coefficients settings

Sampling rate	44.1 kHz
Frame	20 ms
Frame shift	10 ms
Pre-emphasis coefficient	0.975
Hamming window coefficient	0.46
Dimensions	39

B. *Zero crossing rates*

The zero-crossing rate (ZCR) is one of the basic acoustic features that can be easily calculated. The waveform of the voice signal can be a zero-line, and the amplitude above the zero-line will be positive, otherwise it is negative. The unit time is the number of times across the zero-line is frequently that presents waveform swing intensely. Calculation of the number of times across the zero-line in a frame leads to the obtaining of the zero crossing rates.

$$ZCR = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{1}{2} |\mathrm{sgn}[x(n)] - \mathrm{sgn}[s(n-1)]| \tag{17}$$

where $\mathrm{sgn}[\bullet]$ is a sign functions. The amplitude of the opposite two adjacent signals has a value of 1, otherwise the value is 0. In Fig. 3, the blue line represents the original speech signal and the red line is a voice signal for the zero rate curves. In general, the ZCR of both the unvoiced sounds and environment noise are larger than the voiced sounds (which

has observable fundamental periods). Some research used ZCR for fundamental frequency estimation, but it is highly unreliable unless the procedure is further refined.

C. *Short time energy*

In the speech features, the sound intensity changes and intensity waveforms swing are two useful features. In Fig. 4, the blue line represents the original speech signal and the red line is a voice signal for the short time energy curves. Usually, the following formula is used to calculate sound intensity.

$$S_{\text{int}}(m)=\frac{1}{N}\sum_{n=m-N+1}^{m}\left|f_x(n,m)\right|^2 \tag{18}$$

m is the frame position and *1/N* is the time averaging. In fact S_{int} represents the power. Since there is a fixed length frame, *N* is a constant that can be eliminated. Eq. (18) can be rewritten as:

$$E_x(m)=\sum_{n=m-N+1}^{m}\left|x(n)\right|^2 \tag{19}$$

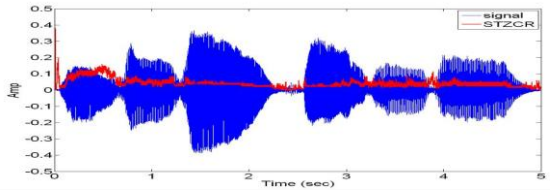


Fig 3. Zero crossing rate curve

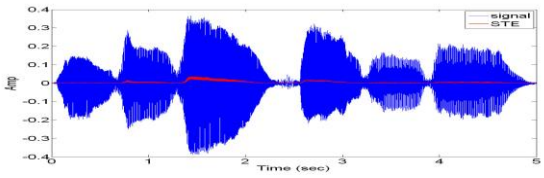


Fig 4. Short time energy curve

IV. Principle of the General Regression Neural Network

The general regression neural network (GRNN) evolved from the probability neural network (PNN) as a method of supervised learning. GRNN was widely applied to many recognition tasks. GRNN is a one-pass learning algorithm with a highly parallel structure. It has the advantages of fast learning and convergence to the optimal regression surface where there is a large number of samples. GRNN can be divided into four layers: input layer, pattern layer, summation layer, and output layer. The input layer is an input unit providing all of the measured values to the pattern layer. Each pattern unit represents a training example. When a new input vector is put into the network, it can be subtracted with training examples. The square difference between the two values will be summed and the non-linear function will be entered. However, the two output values are $\sum_{i=1}^{n} y^i \exp\left(-\frac{P_i^2}{2\nu^2}\right)$ and $\sum_{i=1}^{n}\exp\left(-\frac{P_i^2}{2\nu^2}\right)$ of the summation layer, respectively. The estimates $Y(x)$ can from be obtained from the output layer

$$Y(x)=\sum_{i=1}^{n} y^i \exp\left(-\frac{P_i^2}{2\nu^2}\right)\Big/\sum_{i=1}^{n}\exp\left(-\frac{P_i^2}{2\nu^2}\right) \tag{20}$$

V. Experimental Work and Results

To estimate the proposed singer identification system, experiments are performed to measure the sound signals for various singers and background noise conditions. Fig.5 is the experimental structure of the automatic singer identification system. All signals are recorded in a laboratory environment. Since the volume and length of recording time affect the recognition rate, all the recording signals must carry out pre-processing steps after recording. There are 20 individual singers with 10 males and 10 females singing for five seconds each. Each song was repeated 10 times. Details of the experimental parameters are summarized in Table 2. For simplifying the experimental conditions, not losing the background music of choice was based on two conditions. The first was a single instrument accompaniment of music and the second was a personal solo.

The music data in our experiments were collected from commercial audio CDs at a 44.1 *kHz* sample rate, and 16 bits per sample in stereo. The background noise could be divided into two types. One was musical accompaniment and the other one was Gaussian white noise. The musical accompaniment condition could be divided into three levels: level_1 (80 db), level_2 (90 db) and level_3 (100 db). On the other hand, for the Gaussian white noise conditions, use of the signal to noise ratio (SNR) as the basis divided the noise into three different levels: (10 db), (20 db) and (30 db). According to the above, pre-processed data and input are used for our proposed automated singer identification system. Fig. 6 shows the signal analysis block diagram. The overall analysis framework includes three parts. The pre-processed data undergo speech enhancement and then feature extraction by MFCC. Finally, the neural network is used to categorize the singer. In the experiment for the musical accompaniment noise condition, if only eigenvalues of MFCC were used, the results were not good. Therefore, this study added the zero crossing rate and short-time energy of the two eigenvalues. This is because one segment had a noisy signal, and noise and unvoiced sounds have lower energy and higher ZCR than vowels. Therefore, the recognition rate by the ZCR and STE was increased.

This section also evaluates the proposed speech enhancement algorithm by comparing its performance with that of spectral subtraction (SS), multi-band spectral subtraction (MBSS), and non-linear spectral subtraction (NSS). Fig. 7 shows the same singer and the same song using different analyses of the spectrum subtraction. MBSS could retain the characteristics of the voice signals however, there was relatively less noise reduction. NSS noise reduction clearly lost some characteristics of the speech signal, such as

vibrato sounds. The recognition rates in various levels of background noise (Gaussian white noise) conditions are summarized in Table 3 and Table 4. It is shown non-linear spectrum subtraction has a better recognition rate than the other two. The results of the recognition rate for various background music (musical accompaniment) conditions are summarized in Table 4 and Table 5. It is shown the multi-band spectral subtraction has a better recognition rate than the other two. In the experiment, pure speech can achieve a 92% recognition rate and is summarized by the results in the tables (3) to (6), the noise increased the recognition rate is relatively lower.

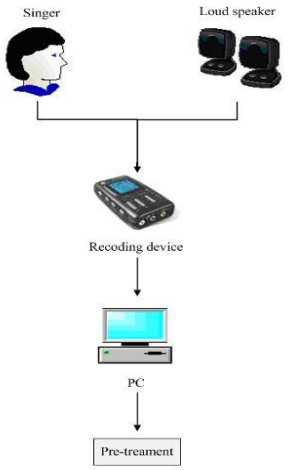


Fig 5. Experimental structure of system.

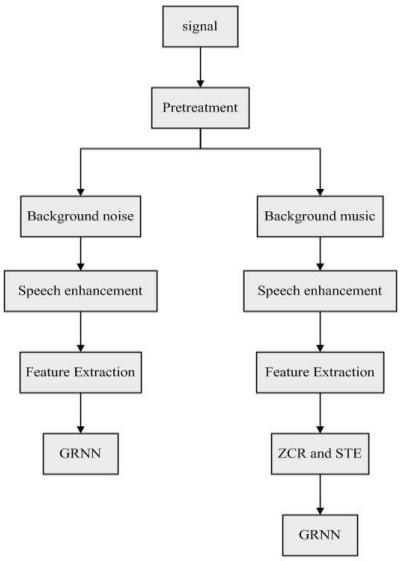


Fig 6. Signal analysis block diagram.

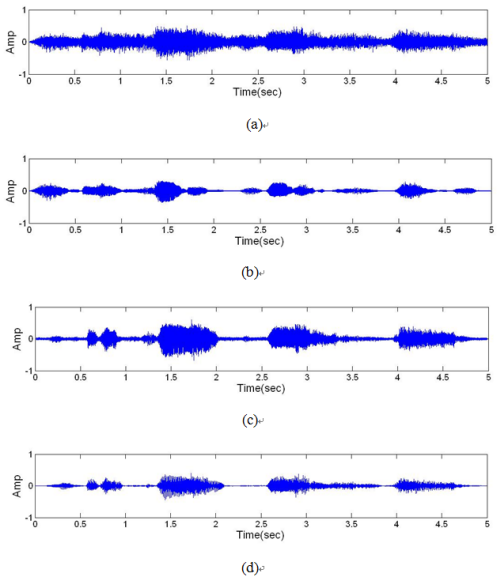


Fig 7. Comparison of three types of speech enhancement: (a) without enhancement (b) SS (c) MBSS (d) NSS

Table 2. Experimental parameters set

Experimental parameters set	
Song	Tian Hei Hei
Singers	10 male and 10 female
Sampling rate	44.1 kHz
Sampling time	5 sec
Background noisy (SNR)	10db, 20db, 30db
Background music (db)	Level_1 (80 db)
	Level_2 (90 db)
	Level_3 (100 db)

Table 3. Recognition rates of background noisy for male

methods / SNR (dB)	Without enhancement	SS	MBSS	NSS
10	51	67	60	66
20	65	71	66	75
30	80	82	75	84

Unit: (%)

Table 4. Recognition rates of background noisy for female

methods / SNR (dB)	Without enhancement	SS	MBSS	NSS
10	32	63	34	55
20	52	74	66	81
30	78	80	77	85

Unit: (%)

Table 5. Recognition rates of background music for male

methods / level (dB)	Without enhancement	SS	MBSS	NSS
80	79	88	88	91
90	79	86	90	85
100	64	80	88	89

Unit: (%)

Table 6. Recognition rates of background music for female

methods / level (dB)	Without enhancement	SS	MBSS	NSS
80	85	86	97	95
90	85	90	95	90
100	74	86	83	81

Unit: (%)

VI. CONCLUSION

This paper presents a singer identification system using the signal spectral subtraction enhancement technique and artificial neural network. The experimental results show the proposed system with various spectrum subtraction algorithms can achieve an acceptable recognition rate in different conditions and different types of background noise. For the background noise conditions, NSS has a better recognition rate than the other two. However, under the background music conditions, MBSS has a better recognition rate than the other two. The proposed system use only 6 to 8 dimensions of the feature coefficients, reducing the feature vector space and a lot of the computation. Overall, the proposed singer identification system is effective and the performance is satisfactory.

ACKNOWLEDGMENT

The study was supported by the National Science Council of Taiwan, Republic of China, under project number NSC-100-2221-E-018 -009.

REFERENCES

[1] F. F. Kuo, M. K. Shan, Looking for new, not known music only: music retrieval by melody style. Processing of the 2004 Joint ACM/IEEE Conference on Digital Libraries. pp. 243-251, 2004.

[2] S. Essid, G. Richard, and B. David, Instrument recognition in polyphonic music based on automatic taxonomies. IEEE Transaction on Audio, Speech, and Language Processing, 2006, 14. pp. 68-80.

[3] G. Tzanetakis, and P. Cook, Musical genre classification of audio signals. IEEE Transaction on Speech and Audio Processing 2002, 10. pp.293-302.

[4] B. Raj, V. N. Parikh, and R. M. Stern, The effects of background music on speech recognition accuracy. Acoustic, IEEE International Conference on Speech, and Signal Processing, ICASSP 1997, pp.851-854.

[5] H. Sheikhzadeh, and L. Deng, Waveform-based speech recognition using hidden filter models: parameter selection and sensitivity to power normalization. IEEE Transaction on Speech and Audio Processing 1994, 2. pp.80-89.

[6] G. Enzner, P. Vary, Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones. Signal Processing 2006, 86, pp. 1140-1156.

[7] J. D. Chen, J. Benesty, and Y. T. Huang, On the optimal linear filtering techniques for noise reduction. Speech Communication 2007, 49. pp. 305-316.

[8] S. Xu, G. Meng, LMS algorithm for active noise control with improved gradient estimate. Mechanical Systems and Signal Processing 2006, 20. pp.920-938.

[9] F. A. Chowdhury, J. Alam, , F. Alam, and D. O'Shaughnessy, Perceptually weighted multi-band spectral subtraction speech enhancement technique. 5th International Conference on Electrical and Computer Engineering ICECE. 2008, pp.395-399.

[10] R. M. Udrea, N. Vizireanu, S. Ciochina, and S. Halunga, Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale. Signal Processing 2008, 88. pp.1299-1303.

[11] W. G. Daniel, S. D. Douglas, and S. L. Jae. Speech synthesis from short-time Fourier transform magnitude and its application to speech processing. Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP 1984, pp.61-64.

[12] E. F. Richard, Compound wavelets: wavelets for speech recognition. Time-Frequency and Time-Scale Analysis. 1994, pp.600-603.

[13] M. Zbancioc, and M. Costin, Using neural networks and LPCC to improve speech recognition. Signals, Circuits and System 2003, 2. 445-448.

[14] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, Classification of general audio data for content-based retrieval. Pattern Recognition Letters 2001, 22. pp.533-544..

[15] S. F. Boll, A spectral subtraction algorithm for suppression of acoustic noise in speech. Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP 1979, 79. 200-203.

[16] B. Y. Xia, Y. Liang, and C. C. Bao, A modified spectral subtraction method for speech enhancement based on masking property of human auditory system. IEEE International Conference on Wireless Communication & Signal Processing .2009, pp. 1-5.

[17] M. Berouti, R. Schwartz, and J. Makhoul, Enhancement of speech corrupted by acoustic noise. Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP 79. 1979, pp.208-211.

[18] D. J. Mashao, M. Skosan, Combining classifier decisions for robust speaker identification. Pattern Recognition 2006, 39. pp.147-155.

[19] J. Bi, S. C. Liu. A speaker identification system for video content analysis. International Conference on Intelligent Information Hiding and Multimedia Signal Processing. 2008, pp.200-203.

First Author

Jian-Da Wu received his MS degree in automotive engine and vehicle design from Institute of Sound and Vibration Research (ISVR), University of Southampton, UK, in 1995, respectively, and PhD degree in mechanical engineering from National Chiao-Tung University of Taiwan in 2001. He is currently a professor in Institute of Vehicle Engineering, National Changhua University of Education. His current research interests are in intelligent vehicle system, vehicle noise and vibration control and digital signal processing in vehicle applications.

Second Author

Cheng-Wei Chung received his BEng degree in mechanical engineering from Tatung University of Taiwan and MS degree in Graduate Institute of Vehicle Engineering from National Changhua University of Education, Taiwan, in 2009. His current research interests are in speech recognition system and digital signal processing.