Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Adaptive wavelet shrinkage for noise robust speaker recognition

Sumithra Manimegalai Govindan [a], Prakash Duraisamy [b], Xiaohui Yuan [b,*]

[a] Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu 638401, India
[b] Department of Computer Science and Engineering, University of North Texas, 3940 N. Elm, Denton, TX 76201, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Speaker recognition
Noise suppression
Wavelet
Feature extraction

ABSTRACT

Speaker recognition faces many practical difficulties, among which signal inconsistency due to environmental and acquisition channel factors is most challenging. The noise imposed to the voice signal varies greatly and a priori noise model is usually unavailable. In this article, we propose a robust speaker recognition method that employs a novel adaptive wavelet shrinkage method for noise suppression. In our method, wavelet subband coefficient thresholds are automatically computed, which are proportional to the noise contamination. In the application of wavelet shrinkage for noise removal, a dual-threshold strategy is developed to suppress noise, preserve signal coefficients and minimize the introduction of artifacts. The recognition is achieved using modification of Mel-frequency cepstral coefficient of overlapped voice signal segments. The efficacy of our method is evaluated with voice signals from two public available speech signal databases and is compared with state-of-the-art methods. It is demonstrated that our proposed method exhibits great robustness in various noise conditions. The improvement is significant especially when noise dominates the underlying speech.

1. Introduction

Speaker recognition extracts features from a piece of voice to deduce the speaker's identity [1]. It is usually treated as two tasks: verification and identification [2]. Speaker verification determines if a person is the claimed identity based on a piece of voice sample; whereas speaker identification determines which one of a group of known voices best matches the input voice sample. This article focuses on speaker identification, but the algorithm can be extended to speaker verification. Speaker recognition is an important tool for countless applications such as access control and user privacy protection. Many such systems have been developed that achieve very good results given clean, high-quality voice signals with similar training and testing acoustic conditions. However, under noisy environments, which is often expected in a large number of real-world applications (e.g., voice-based user verification using cell phones), system performance degrades dramatically, far from a satisfactory level [3,4]. The feature vectors generated from corrupted speech are no longer similar to the class distributions represented by the training data sets. Because of the channel effects, there is inherently more variability in the training data, and as a result, the variance of speaker classes distributions increases. This leads to increased errors over the cases where the training and test speech are both clean. To make it practical and feasible for consumer devices, it is highly desired that robust methods are developed to withstand noise interference from various sources of unknown types.

Existing methods to tackle noise contamination in speaker recognition mostly focus on creating acoustic model of the background [5] and removing noise from input voice signals [6]. The challenge remains that well-trained models result in inferior performance when training and testing signals are contaminated with different types of noise. Motivated by the success of the noise-aware image fusion method in [7], we propose a robust speaker recognition method that employs an adaptive Bionic wavelet shrinkage (ABWS). Without the knowledge of the noise characteristics or assuming a model of the underlying clean speech signal, our method leverages the sparsity of the wavelet coefficients and suppresses noise to improve signal quality and hence recognition accuracy. In our ABWS method, subband coefficient thresholds are automatically computed that are proportional to the amount of noise contamination. In addition, our dual-threshold strategy (DuTS) preserves signal coefficients and introduces little artifact with gradual coefficient amplitude suppression. The effectiveness of our method is evaluated with samples from two public available speech databases (TIMIT and KING) and a comparison study with state-of-the-art methods is conducted.

The rest of this paper is organized as follows: Section 2 presents the architecture of our speaker recognition system and related work. Section 3 describes our noise robust speaker recognition

[*] Corresponding author.
E-mail address: xiaohui.yuan@unt.edu (X. Yuan).

http://dx.doi.org/10.1016/j.dsp.2014.06.007
1051-2004/© 2014 Elsevier Inc. All rights reserved.

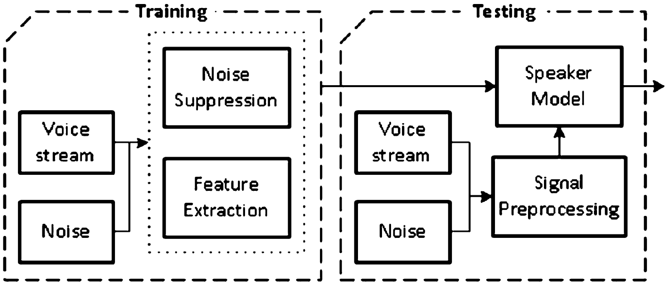

Fig. 1. An overview of an automatic speaker recognition system.

method in detail. We discuss the noise driven adaptive subband coefficient threshold estimation method followed by our dual-threshold wavelet shrinkage. Section 4 presents the experimental results and discussions. Finally, Section 5 concludes the paper with a summary of our method.

2. Background and related work

Given exemplar voice streams, it usually requires some processing steps to construct an automatic speaker recognition system. These steps include noise suppression, feature extraction, and a learning component to create multi-speaker models (see Fig. 1 for a system overview). In real-world applications noise is inevitable and its characteristics vary, which makes noise removal a key step in automatic speaker recognition.

Among the steps shown in Fig. 1, feature extraction aims at extracting acoustic signatures that are critical for differentiating speakers. It usually transforms the speech signal into a compact representation. The performance of a speaker recognition system is highly dependent on the quality of the selected speech features.

Speaker recognition has been extensively investigated and many algorithms have been developed to address noisy voice problem. Hermansky and Morgan [8] developed a band-pass filtering method to suppress spectral components that vary more slowly or quickly than the change of the speech. Ming et al. [9] combined a multi-condition model and the missing features to compensate signal noise. A similar strategy was developed in [5] where a universal background model [10] was used, which is based on Gaussian Mixture Model (GMM) using acoustic features to represent the general, speaker-independent distribution of features. Zao and Coelho [11] proposed a multi-condition training technique that employs GMM for speaker modeling. Kim and Gales [12] extended linear transform for model-based adaptation that uses a modified version of generative model between clean and noisy signals. Deng et al. [13] proposed a stereo-based piecewise linear compensation method and demonstrated its effectiveness with white noise, babble noise and office noise. Liao et al. [14] used latent prosody analysis to extract spectral features for speaker identification. It combines prosodic feature-based system with maximum-likelihood a priori knowledge. Wang and Gales [6] extended the acoustic factorization method that assigns separate transforms to represent the speaker and noise. Nemala et al. [4] used a multistream feature processing to address additive noise as well as slow-varying channel conditions. Padilla et al. [15] proposed a soft spectral subtraction method that handles missing features in speaker verification, which improved verification performance as long as a minimum number of features were obtained. Brajevic et al. [16] proposed a method that uses short time Fourier transform and Ephraim–Malah estimation to deal with signals with stationary noise. The method reduced spectral coefficients and hence suppresses noise. Abd El-Fattah et al. [17] described an adaptive Wiener filter in time domain to accommodate varying nature of speech signals with noise. Penda and Srikanthan [18] presented a compensation scheme to adapt model parameters to reduce the disparity between training and testing data sets. Despite the great effort devoted to accommodate noisy voice signals, mismatched noise type in training and testing (i.e., recognition) is still an open problem.

Alternatively, methods using wavelet transformation have become increasingly popular in many signal processing applications [7,19] because noise components in wavelet subbands are usually characterized by coefficients with small magnitude. Mallat and Hwang [20] have shown that effective noise suppression may be achieved by transforming the noisy signal into the wavelet domain, and preserving only the coefficients of local maxima. A reconstruction that used only the large-magnitude coefficients was shown to approximate the uncorrupted signal well. In other words, noise suppression is achieved by suppressing the magnitude of the wavelet coefficients of the contaminated signal. In [21], Donoho employed the thresholding in the wavelet domain and demonstrated the denoised result to have near optimal properties for a wide class of signals that were corrupted by additive white Gaussian noise. Johnstone and Silverman [22] proposed a level-dependent threshold to remove colored noise. Johnson et al. [19] extended the Bionic Wavelet Transform (BWT) [23] in combination with the existing wavelet denoising techniques to construct a wavelet thresholding method for speech enhancement. Bahoura and Rouat [24] proposed the method of level dependent wavelet thresholding, using the Teager energy operator to improve the discrimination for determining whether a voice stream was dominated by speech or noise. Khaled Daqrouq [25] investigated the utilization of wavelet filters via multistage convolution by reverse biorthogonal wavelets in high-pass and low-pass frequency bands of a speech signal. Speech signal was decomposed into two frequency bands and the noise was removed in each band individually at different stages via wavelet filters. Ghanbari and Karami-Mollaei [26] developed a node dependent wavelet thresholding and modified thresholding functions were introduced to improve accuracy. Despite the advancement of wavelet shrinkage based noise removal, the aforementioned methods have not been applied to speaker recognition. The threshold used in shrinkage needs to be automatically decided according to the noise type and noise strength.

When applying an automatic speaker recognition system, the acquired speaker voice stream is usually processed with the same pre-processing steps used in the training phase with a goal of removing the inner speaker disparity. However, the voice acquired for recognition often contains different noise characteristics, e.g., magnitude and frequency characteristics. Such mismatch could degrade the recognition accuracy. The different noise characteristics in the training and testing voice signals are not fully addressed in the aforementioned methods. In this article, we propose a method to adaptively suppress noise component in voice signals to improve the robustness with respect to noise and hence facilitate speaker recognition in real-world applications.

Since the research in speaker recognition and the related fields have yielded a large number of methods, the acronyms and symbols used in this article could be confusing. For the ease of understanding our description, Table 1 summarizes the acronyms and symbols used in this article, which are ordered alphabetically for the ease of search. Further description of variables and functions is presented in the text when they are first used.

3. Adaptive discrete Bionic wavelet shrinkage

Inspired by the success of noise-aware image fusion method in [7], we propose an ABWS method to account for excessive noise in the speech signals. In this section, we first briefly review the Bionic wavelet transform and present our extension to include capability of adapting to time. We then describe our automatic

Table 1
Acronyms and symbols used in this article.

Symbol	Description	Symbol	Description
ABWS	Adaptive Bionic Wavelet Shrinkage	c_j	Wavelet coefficients in the jth scale
BFCC	Bark Scale Filter bank Cepstrum Coefficients	d_j	Number of samples in scale j
BWT	Bionic Wavelet Transform	$E(\cdot)$	Expectation
DuTS	Dual Threshold Shrinkage	f	Input voice signal
EMF	Ephraim–Malah Filtering	$\hat{f}_s$	Noise suppressed signal
IWF	Iterative Wiener Filtering	p, q	Gains factor in function λ
IS	Itakuro–Saito distance	s	Saturation constant in function λ
LBG	Linde–Buzo–Gray algorithm	$\mathcal{T}_d$	Dual thresholding function
LPCC	Linear Prediction Cepstral Coefficients	$\mathcal{T}_h$	Hard thresholding function
MFCC	Mel-Frequency Cepstral Coefficient	$\mathcal{T}_s$	Soft thresholding function
MMFCC	Modification of MFCC	T_j	Threshold for the jth scale
MOS	Mean Opinion Score	η	Cut-off amplitude
MSE	Mean Square Error	λ	Time adaptive function
RPLP	Revised Perceptual Linear Prediction	$\hat{\sigma}$	Estimated noise standard deviation
SS	Spectral Subtraction	$\hat{\sigma_j}$	Median coefficient magnitude
SNR	Signal Noise Ratio	σ_n	Noise standard deviation
a	Dilation factor in BWT	τ	Time shift in BWT
$\mathcal{B}$	BWT coefficients	φ	Mother wavelet function
$\widetilde{B}$	Time adaptive BWT coefficients	φ^*	Complex conjugate of φ

method for estimating subband noise magnitude to determine a scale dependent threshold. Lastly, we introduce our ABWS with dual-threshold shrinkage function.

3.1. Discrete Bionic wavelet transform

The Bionic wavelet transform (BWT) [23] is a time adaptive wavelet transform based on the Morlet wavelet designed specifically to model the human vocal signals. This signal transformation is based on the Giguere–Woodland non-linear transmission line model of the auditory system [27].

The BWT of an input signal $f(t)$, denoted by $\mathcal{B}(\omega)$, is formulated as follows:

$$\mathcal{B}_{a,\tau}(\omega) = \frac{1}{\lambda\sqrt{a}} \int f(t)\varphi^*\left(\frac{t-\tau}{a\lambda}\right) e^{(-j\omega(\frac{t-\tau}{a}))}\,dt, \tag{1}$$

where φ^* is the complex conjugate of the mother function $\varphi(t)$ and $\varphi(t) = \frac{1}{\sqrt{a}}\tilde{\varphi}(t)e^{j\omega t}$. Factors a and τ are the dilation and time shift parameters, respectively. $\mathcal{B}_{(a,\tau)}$ is the coefficient of the BWT at time τ and scale a.

The time adaptive function $\lambda(\tau + \delta\tau)$ is derived from the active auditory model [28] and is expressed as follows:

$$\lambda(\tau + \delta\tau) = \frac{1}{(1 - p\frac{s}{s+|\mathcal{B}_{a,\tau}|})} \frac{1}{(1 + q|\frac{\partial}{\partial t}\mathcal{B}_{a,\tau}|)}, \tag{2}$$

where s is a saturation constant, and p and q are the gains. By changing the values of p and q, we can adjust the resolution in frequency and time domains, respectively. Factor λ in Eq. (1) must be a constant in the period $\delta\tau$. This approximation is plausible if the signal and its first derivative are continuous and the signal amplitude is small enough [23].

Based on the λ function, the coefficients of time adaptive BWT, denoted by $\tilde{\mathcal{B}}$, are expressed as the weighted coefficient of BWT as follows:

$$\tilde{\mathcal{B}}(\omega) = \frac{\sqrt{\pi}}{2\sqrt{\lambda^2 + 1}}\mathcal{B}(\omega). \tag{3}$$

3.2. Noise-driven adaptive threshold selection

In the wavelet domain, noise is characterized by coefficients with small amplitude, while the underlying clean signal dominates the coefficients with large amplitude [21]. Taking advantage of this property, noise can be removed via thresholding the wavelet coefficients, namely wavelet shrinkage.

The main issue in wavelet shrinkage is to determine an appropriate threshold. Donoho and Johnstone [21] proposed the SUREshrink method based on the Stein's Unbiased Risk Estimator. The threshold for wavelet shrinkage is chosen as the value that results in a minimized estimation error, i.e., $\sqrt{2\log(d)}$, where d is the dimension of the signal and is independent of the data. This threshold is attractively simple and works well for the uncorrelated noise.

If the noise is colored and non-stationary, the variance of the noise wavelet coefficients will be different for each scale in the wavelet decomposition. In this case, scale dependent thresholding proposed by Johnstone and Silverman [22] accounts for the different variances of the noise wavelet coefficients in each scale. For scale dependent thresholding, the noise variance for scale j can be estimated using the median coefficient magnitude of the high-pass subband as follows [29]:

$$\hat{\sigma_j} = \frac{Median(|c_j|)}{0.6745}, \tag{4}$$

where c_j represents the wavelet coefficients in the jth scale high-pass subband. This method is insensitive to the outliers with large magnitude but gives a rough estimation. The set of standard deviation values can now be used as the “noise profile” for selecting thresholds.

In our prior work [30] the authors developed a method to overcome the bias caused by sample size. This method takes advantage of the sparseness property of the wavelet subband. The idea is to identify the coefficients that were distorted and calculate the variance, which closely approximates the noise distribution if the following two conditions are satisfied: 1) The noise variance is much greater than that of the underlying clean coefficients used for variance computation; 2) The number of such coefficients is statistically large enough.

The noise component in each subband is modeled with the zero-mean super Gaussian function and the coefficients that constitute noise are small in amplitude. Hence, we can construct a subband variance function with respect to the cutoff amplitude η and the optimal η can be identified by finding the inflection point of the variance function's first derivative [7] as follows

$$\eta = \arg\max_{\eta} \frac{\partial\sigma^2(\eta)}{\partial\eta}. \tag{5}$$

Using the estimated cutoff amplitude η, the noise variance is computed using the coefficients that are below the cutoff value:

Please cite this article in press as: S.M. Govindan et al., Adaptive wavelet shrinkage for noise robust speaker recognition, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.06.007

$$\sigma_n^2(\eta) = \frac{1}{N-1}\sum_{|f|\le\eta}(f-\bar{f})^2, \tag{6}$$

where N is the number of coefficients that satisfy the condition $|f| \le \eta$.

With the noise variance function in Eq. (6), the optimal noise coefficient range can be determined by minimizing the following error function:

$$\eta^* = \arg\min_{\lambda} E\big(\hat{f}_s(\lambda) - f_s\big)^2, \tag{7}$$

where $\hat{f}_s(\lambda)$ is the noise suppressed signal and $E(\cdot)$ is the expectation function. The closed form solution is derived as follows:

$$\eta^* = \frac{\sigma_n^2(\nu)}{\sqrt{\sigma_f^2 - \sigma_n^2(\nu)}}. \tag{8}$$

The scale dependent threshold depends on the noise coefficient magnitude in each subband. The threshold for the jth scale is computed as follows:

$$T_j = \hat{\sigma}_n^{(j)}\sqrt{2\,\log(d_j)}, \tag{9}$$

where d_j is the number of samples in scale j and $\hat{\sigma}_n^{(j)}$ is the estimated noise variance.

3.3. Dual-threshold shrinkage function

In wavelet shrinkage, wavelet coefficients are suppressed according to their amplitude. Two widely used shrinkage functions are hard thresholding and soft thresholding. In hard thresholding method, wavelet coefficients are suppressed to zero if their amplitude is below the threshold; otherwise, they retain the original value. The hard thresholding function $\mathcal{T}_h$ is expressed as follows:

$$\mathcal{T}_h(T) = \begin{cases} \tilde{\mathcal{B}} & |\tilde{\mathcal{B}}| > T \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The discontinuity at threshold T could introduce artifact after signal reconstruction. Hence, soft thresholding is developed. Instead of suppressing only the coefficients with small amplitude to zero, the amplitude of all coefficients is universally reduced by the magnitude of the threshold. The thresholding function $\mathcal{T}_s$ is formulated as follows:

$$\mathcal{T}_s(T) = \begin{cases} sgn(\tilde{\mathcal{B}})(|\tilde{\mathcal{B}}| - T) & |\tilde{\mathcal{B}}| > T \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where function $sgn(\tilde{\mathcal{B}})$ gets the sign of the wavelet coefficient $\tilde{\mathcal{B}}$.

Using soft thresholding method, coefficients with large amplitude are "punished" in the same scale as those noise coefficients. The large coefficients, however, constitute the fundamental part of the signal and suppressing them changes the high frequency components in the reconstructed signal, which in turn alters the speaker's acoustic property.

To circumvent the artifacts introduced by coefficient shrinkage functions, we propose a dual-threshold shrinkage (DuTS) function. The idea is to retain the large coefficients without any change and truncate the small ones. For the coefficients fall between the two thresholds, the suppression follows a piece-wise linear function that imposes stronger suppression to the smaller coefficients in the range. The DuTS function $\mathcal{T}_d$ is as follows:

$$\mathcal{T}_d(T_1, T_2) = \begin{cases} \tilde{\mathcal{B}} & |\tilde{\mathcal{B}}| > T_2 \\ sgn(\tilde{\mathcal{B}})W(\tilde{\mathcal{B}}) & T_1 < |\tilde{\mathcal{B}}| \le T_2 \\ 0 & |\tilde{\mathcal{B}}| \le T_1 \end{cases} \tag{12}$$

and

$$W(\tilde{\mathcal{B}}) = \frac{T_1 T_2(|\tilde{\mathcal{B}}| - T_1)}{T_2 - T_1}. \tag{13}$$

Following the DuTS function, if the amplitude of a coefficient is less than the lower threshold T_1, its value is suppressed to zero; if it is greater than upper threshold T_2, its value is retained. If its amplitude is between the two thresholds, its value is rescaled according to Eq. (13).

In this method, threshold T_2 is always greater than T_1 and can be expressed as

$$T_2 = kT_1, \tag{14}$$

where $k > 1$. Hence the function in Eq. (13) is rewritten as

$$W(\tilde{\mathcal{B}}) = \frac{k}{k-1}\big(|\tilde{\mathcal{B}}|T_1 - T_1^2\big). \tag{15}$$

The suppressed coefficients between T_1 and T_2 must be less than or equal to $\tilde{\mathcal{B}}$, i.e.,

$$T_1 T_2 \frac{|\tilde{\mathcal{B}}| - T_1}{T_2 - T_1} \le |\tilde{\mathcal{B}}|. \tag{16}$$

Because the shrinkage function is symmetric about the origin, we derive the following relations with $\tilde{\mathcal{B}} \ge 0$ without loss of generality. We can rewrite Eq. (16) as

$$kT_1 T_1 \frac{\tilde{\mathcal{B}} - T_1}{kT_1 - T_1} \le \tilde{\mathcal{B}}. \tag{17}$$

Since $kT_1 - T_1 > 0$, we have

$$kT_1^2(\tilde{\mathcal{B}} - T_1) \le (kT_1 - T_1)\tilde{\mathcal{B}}.$$

We simplify both sides and get an explicit expression for k:

$$k \ge \frac{\tilde{\mathcal{B}}}{T_1^2 - T_1 + \tilde{\mathcal{B}}}.$$

Applying the boundary condition on both sides in Eq. (17), the equality is reached when $\tilde{\mathcal{B}} = T_2$, and, hence, we have

$$k = \frac{T_2}{T_1^2 - T_1 + T_2} = \frac{T_2/T_1}{T_1 - 1 + T_2/T_1} = \frac{k}{T_1 - 1 + k}.$$

The value of k is therefore bounded by $2 - T_1$, i.e.,

$$k \ge 2 - T_1 > 1. \tag{18}$$

That is,

$$T_1 < 1. \tag{19}$$

The range for k and T_1 ensures that the shrinkage function always suppresses the noise coefficient amplitude.

Fig. 2 depicts the thresholding functions. Since all these functions are symmetric about the origin, only the positive side of each function is illustrated. Our proposed shrinkage function retains the original coefficients for the ones with large amplitude and allows gradual suppression between the two cutoff thresholds to avoid artifacts.

Our ABWS method is described in Algorithm 1. In the speaker recognition process, speech signals are processed with ABWS method before they are used for training a speaker model or performing classification.

Please cite this article in press as: S.M. Govindan et al., Adaptive wavelet shrinkage for noise robust speaker recognition, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.06.007

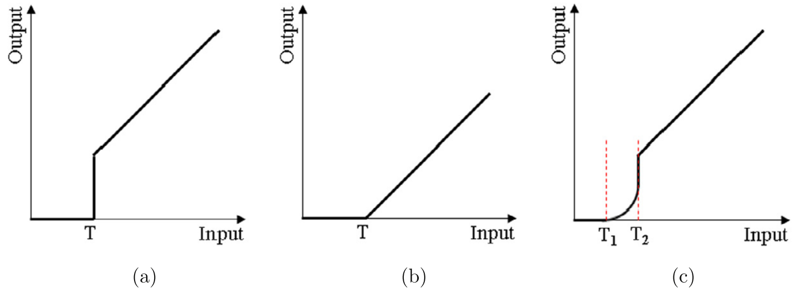

Fig. 2. Three shrinkage functions. Since the function is symmetric about the origin, only the positive side is illustrated. (a) Hard thresholding function. (b) Soft thresholding function. (c) Dual-threshold quadratic shrinkage function.

Table 2
Experimental data sets and their properties.

Properties	TIMIT	KING
No. of speakers	438 male and 192 female	51 male
No. of sessions/speaker	1	10
Type of speech	10 phonetically rich sentences	Extemporaneous descriptions of photograph to interlocutor
Microphones	Fixed wide-band headset	Dual: Wide-band and telephone handsets
Channels	Wide-band/clean	Dual clean: clean and PSTN
Sampling rate	16 kHz	8 kHz
Digital quantization	16 bit	16 bit

Algorithm 1 Adaptive Bionic wavelet shrinkage.

```
Input: speech signal $f$.
Decompose signal $f$ following Eq. (1) and Eq. (3) to get time adaptive Bionic wavelet coefficients.
for each highpass subband do
    Estimate noise variance following Eq. (6).
    Compute the thresholds $T_1$ and $T_2$ following Eq. (9) and Eq. (14), respectively.
    Perform dual-threshold wavelet shrinkage following Eq. (12).
end for
Recover Bionic wavelet coefficients
    $\mathcal{B}(\omega) = \frac{2\sqrt{\lambda^2+1}}{\sqrt{\pi}}\tilde{\mathcal{B}}(\omega).$    (20)
Apply the inverse wavelet transform to the modified coefficients $\mathcal{B}(\omega)$ to get clean speech signal $\hat{f}$:
    $\hat{f} = \mathcal{B}^{-1}(\omega).$    (21)
```

Table 3
MOS 3 rating scale and description.

Rating	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Perceptible and annoying
1	Bad	Perceptible and very annoying

4. Experimental results and discussion

4.1. Experiment settings and evaluation metrics

Experiments are conducted to evaluate the performance of the proposed method in both controlled and uncontrolled environments. Data from TIMIT Acoustic-Phonetic Continuous Speech Corpus [31] and the King database [32] were used in our experiments, properties of which are presented in Table 2.

In our experiments, ten noise types from the NOISEX-92 database [33] were used including Babble (Multiple talkers) noise, HF channel noise, Train noise, Airport noise, Car noise, Street noise, Factory noise, Exhibition noise, Station noise and Restaurant noise. The clean speech utterance was corrupted by these noises and the energy level of the noise was scaled such that the SNR of the noise distorted voice signals was at scales of −5 dB, 0 dB, 5 dB, 10 dB and 15 dB.

When applying Bionic wavelet transform, the saturation constant s in Eq. (2) was set to 0.8. Also the gain parameters p and q were decided empirically to be 0.87 and 0.45, respectively. These parameters were used in both our method and the BWT based wavelet shrinkage method.

The quality of speech denoising was evaluated using both subjective and objective distortion metrics. The objective metrics include Signal to Noise Ratio (SNR), Mean Square Error (MSE) and Itakuro–Saito (IS) distance [34] as follows:

$$\mathrm{SNR} = \frac{f^2}{n^2}, \tag{22}$$

$$\mathrm{MSE} = \frac{1}{N}\sum (f - \hat{f})^2, \tag{23}$$

$$\mathrm{IS} = \frac{1}{2\pi}\sum \left(\frac{f}{\hat{f}} - \log\frac{f}{\hat{f}} - 1\right), \tag{24}$$

where f is the clean underlying signal, $\hat{f}$ is the denoised signal, and n is the noise signal. The IS distance was proposed by Itakura and Saito [35] from the maximum likelihood estimation of short-time speech spectra under autoregressive modeling. We use f and $\hat{f}$ in Eq. (24) to be consistent with the notation in other two metrics, but they denote the spectra of the reference and to-be-recognized speech signals. IS distance is a measure of the perceptual difference between the two spectra. Ideally, if the spectra match, the IS distance reaches zero; otherwise, IS distance is positive.

The average of human evaluations provides another means of assessing audio quality. A popular method is the absolute category rating test, in which volunteers are asked to rate audios using the discrete scale as described in Table 3. This test is commonly referred to as the mean opinion score (MOS) test and it provides a numerical measure of the audio quality [36]. In our experiments, ten human subjects were recruited to evaluate the processed voice signals based on MOS ratings.

Four noise suppression methods are used in our comparison including spectral subtraction (SS) [15], Iterative Wiener filtering (IWF) [17], Ephraim–Malah Filtering (EMF) [16] and wavelet shrinkage using BWT [23].

Please cite this article in press as: S.M. Govindan et al., Adaptive wavelet shrinkage for noise robust speaker recognition, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.06.007

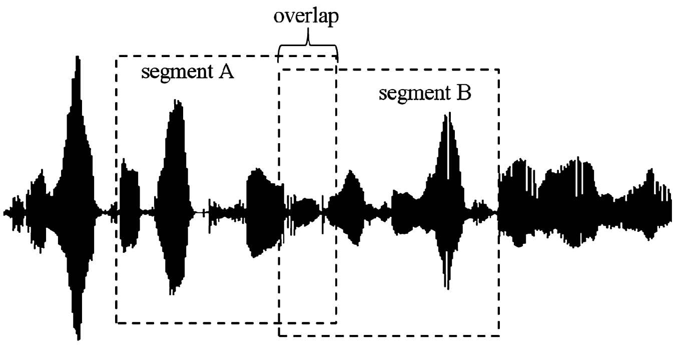

Fig. 3. A speech signal is divided into segments for feature extraction. Overlap is allowed between two adjacent segments.

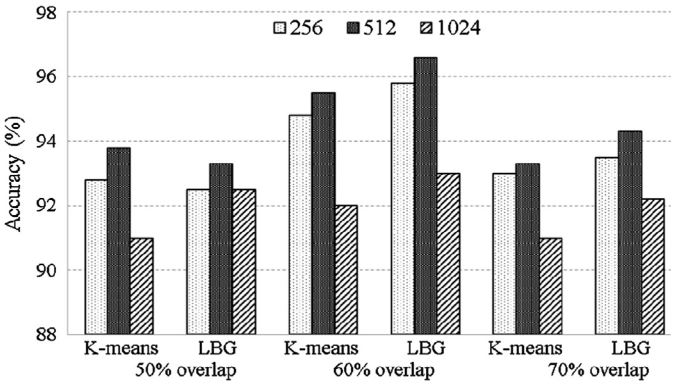

Fig. 4. Recognition accuracy with respect to sample segment size and overlap percentage. The overlap is represented as the percentage of a speech segment size.

4.2. Speech feature extraction

There have been many features developed such as MFCC [37], MMFCC [38], LPCC [39], BFCC [40], and RPLP [41]. In our method, a speech signal was divided into fixed size segments (in terms of number of samples), and two adjacent speech segments could overlap (as depicted in Fig. 3), which makes a speech sample quasi-stationary. The acoustic features were extracted from each segment.

Fig. 4 illustrates the average recognition accuracy using K-means and Linde–Buzo–Gray [42] (LBG) methods with different segment sizes and overlap percentages. In this experiment, we used MFCC features. As shown in this figure, when the segment size becomes larger, the overall performance drops (such a trend can also be observed in Table 4). Hence, based on our empirical results we used LBG method with the segment size of 512 and the adjacent segment overlap of 60%.

Table 4 lists the average recognition accuracy using the six feature extraction techniques and five speech segment sizes. Each element in this table is an average accuracy of all corrupted voice signals using the specified segment size and feature. In this comparison, we fixed the segment overlap percentage to 60%. The bottom row in each table lists the overall average accuracy with respect to a feature. It is evident that the MFCC + MMFCC method consistently yielded the best performance for all cases. Also worth of noting is that the recognition accuracy dropped as the speech segment size increased regardless of the feature used. The best performance is usually achieved with the segment size of 256 or 512. This implies that features extracted from small speech segments provide better discrimination among speakers. Note that no matter what segment size and overlap size were used the entire speech signal was processed for recognition.

Table 4
Accuracy (%) comparison of feature extraction techniques and speech segment size using K-means and LBG methods.

K-means method

Segment size	MFCC	MMFCC	LPCC	RPLP	BFCC	MFCC + MMFCC
256	89.02	73.59	43.09	64.77	39.67	94.82
512	90.79	68.97	39.77	65.77	35.82	95.60
1024	83.52	72.28	30.68	46.17	23.73	86.48
2048	79.06	65.97	23.47	45.87	19.69	81.31
4096	74.15	53.98	20.78	29.78	16.26	76.22
Mean	83.31	66.96	31.56	50.47	27.03	86.89

LBG method

Segment size	MFCC	MMFCC	LPCC	RPLP	BFCC	MFCC + MMFCC
256	89.67	75.37	47.82	67.57	40.64	95.76
512	92.78	74.27	40.87	65.49	37.82	**96.67**
1024	86.42	72.08	30.98	58.23	25.27	88.49
2048	80.86	66.89	28.97	47.39	23.34	84.06
4096	75.85	38.98	25.65	38.86	20.07	78.32
Mean	85.12	65.52	34.86	55.51	29.43	88.66

Support vector machine

Segment size	MFCC	MMFCC	LPCC	RPLP	BFCC	MFCC + MMFCC
256	88.01	72.41	47.1	65.82	38.26	91.61
512	91.4	75.82	44.72	68.31	34.2	92.73
1024	85.31	70.03	39.5	63.27	22.84	85.83
2048	79.63	65.12	31.9	58.26	19.03	83.23
4096	72.4	54.52	27.8	42.08	17.94	74.7
Mean	83.35	67.58	38.2	59.55	26.45	85.62

With K-means, LBG method, and SVM, the combination of MFCC and MMFCC gives the highest performance consistently. The improvement rates with respect to the best MFCC and MMFCC performances (underlined in the table) are 4.2% and 27.5%, respectively. It is evident that the improvement using the combination of MFCC and MMFCC was significant, which achieved the average recognition accuracy in the range of upper 90 s. Based on the results of these two experiments, we used LBG method with segment size of 512 and overlap of 60% in the rest of our experiments.

4.3. Evaluation of noise suppression

In our experiments, factor k in our shrinkage function was chosen empirically. We had $k = \sqrt{2}$, that is, $T_2 = \sqrt{2}T_1$. The adaptive subband threshold given by Eq. (9) decides the T_1 threshold as follows

$$T_1 = \frac{\sqrt{2}}{2} T_j, \tag{25}$$

where j is the wavelet subband scale. Depending on the noise strength, T_1 threshold usually lies in the range of [0 0.8].

Fig. 5 illustrates the temporal view and the power spectra of the voice signals using different noise-suppression methods. In this example, “factory noise” was added to the clean voice signal, which resulted in SNR = 0 dB. The left column shows the temporal view of the signals, and the right column shows the corresponding power spectrum. With the presence of noise, signal energy was spread across the entire spectrum (as shown in Fig. 5(b)). Using SS, IWF, EMF, BWT based shrinkage methods and our ABWS method, noise was suppressed to different degrees. After noise suppression with SS and IWF methods, there still existed fairly high noise energy across all frequencies as shown in the spectrograms. It is evident that the enhanced speech signals obtained with EMF,

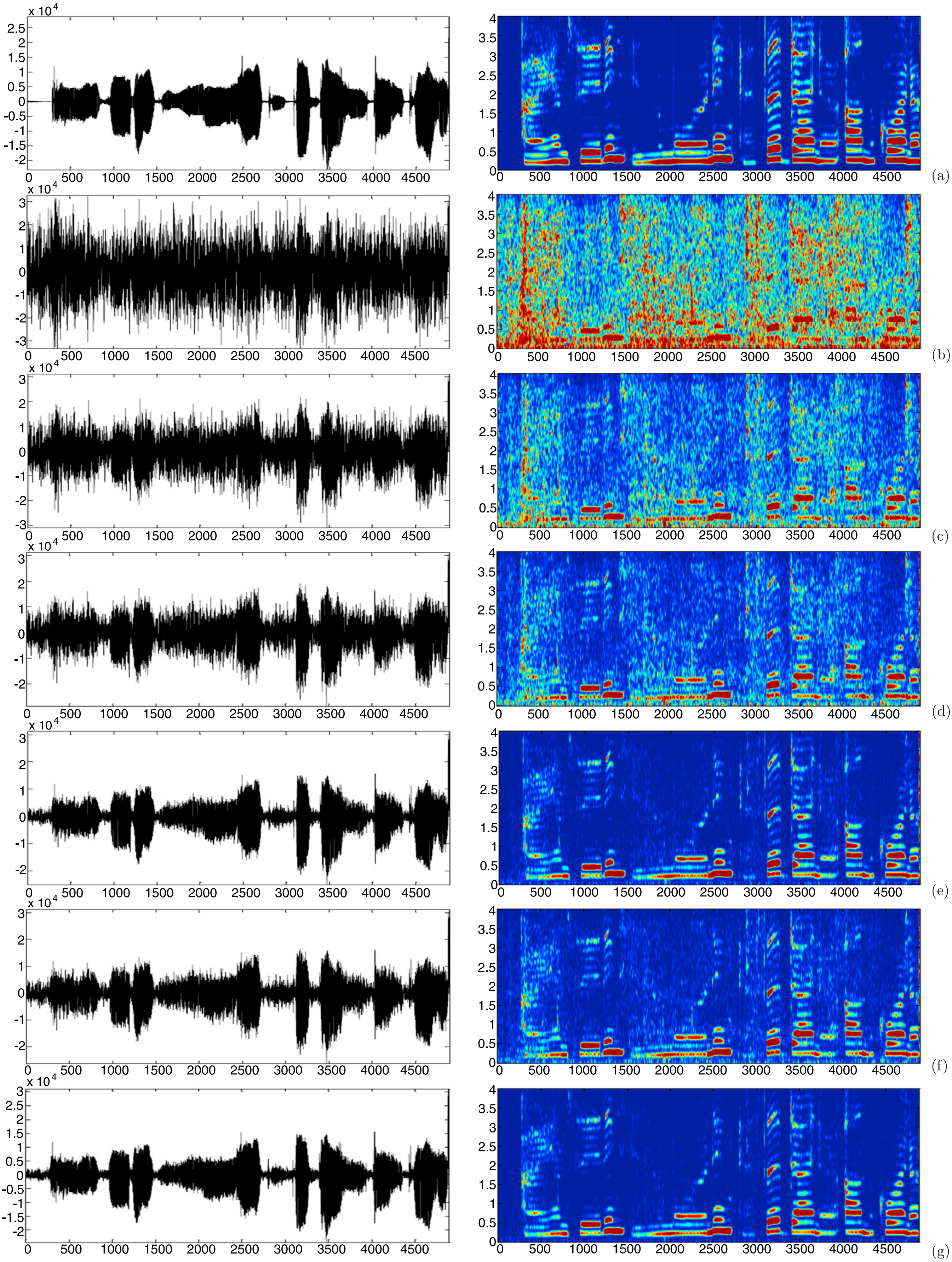

Fig. 5. Temporal and spectral views of the signals and the denoised results. (a) and (b) are clean and noisy signals. (c)–(g) are the denoised signal using SS method, IWF method, EMF method, BWT method, and our method, respectively.

Please cite this article in press as: S.M. Govindan et al., Adaptive wavelet shrinkage for noise robust speaker recognition, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.06.007

Table 5
Objective voice quality evaluation.

Metric	Noise	SS	IWF	EMF	BWT	ABWS
SNR	−5 dB	−0.43	−0.51	1.39	1.44	3.09
	0 dB	4.97	8.04	9.74	9.71	11.45
	5 dB	11.50	14.59	16.31	16.66	19.24
	10 dB	16.24	19.41	21.49	22.87	26.15
	15 dB	19.18	22.05	27.85	28.54	32.15
	Mean	10.29	12.82	15.36	15.84	18.42
MSE (10^{-3})	−5 dB	3.72	3.01	2.35	2.50	1.49
	0 dB	5.14	3.94	3.32	3.10	2.35
	5 dB	7.13	5.79	4.95	4.53	3.17
	10 dB	13.74	9.83	7.73	6.77	4.97
	15 dB	25.37	19.94	15.57	13.64	11.21
	Mean	11.02	8.50	6.78	6.11	4.64
IS	−5 dB	0.86	0.75	0.67	0.78	0.43
	0 dB	2.00	1.70	1.31	1.51	1.25
	5 dB	2.91	2.98	2.83	2.65	2.16
	10 dB	3.19	3.16	2.86	2.67	2.02
	15 dB	3.37	3.08	2.80	2.49	2.09
	Mean	2.47	2.33	2.09	2.02	1.59

Table 6
Subjective voice quality evaluation using MOS test.

Noise	SS	IWF	EMF	BWT	ABWS
−5 dB	3.48	3.60	3.82	3.92	4.04
0 dB	3.19	3.36	3.52	3.66	3.87
5 dB	2.65	2.92	3.07	3.35	3.62
10 dB	2.00	2.15	2.17	2.36	2.61
15 dB	1.20	1.30	1.39	1.52	1.65

Table 7
The average time cost to perform shrinkage-based denoising methods including hard thresholding, soft thresholding, fast adaptive shrinkage [43], and our method. The times are in seconds.

Hard thresholding	Soft thresholding	Fast adaptive shrinkage	ABWS
31.36	43.67	63.71	72.47

BWT based shrinkage and our method were comparatively superior while EMF and BWT had some residual noise components in their middle to upper frequency range, as shown in Fig. 5(e) and (f). Compared to the clean voice signal (as shown in Fig. 5(a)), the result of our method (as shown in Fig. 5(g)) depicts the highest fidelity to the clean signal, which can also be observed in its spectrogram.

Table 5 lists the quantitative evaluation results using the SNR, MSE, and IS. Note that large SNR indicates superior signals; whereas small MSE and IS imply superior results. The SNRs are reported in dB. The results represent the average across signals and noise types. All methods suppress noise to different degrees. As noise level increases, EMF, BWT, and our method (ABWS) demonstrated much better performance. As shown in SNR, the difference can be as much as 8 dB between SS and ABWS. Our method consistently yielded better results than the others, which agrees with the observation from the temporal/spectral plots depicted in Fig. 5. It is interesting to note that using the average IS measures we can clearly see the superior performance of EMF, BWT, and ABWS.

Alternatively, we conducted subjective evaluation based on MOS test. Ten volunteers were recruited and each person was asked to rate the voice according to Table 3. Each voice piece was repeated twice and randomly played back to the volunteer. If the two ratings of the same piece were similar (i.e., less than or equal to one grade value difference), the average was recorded. If the two ratings differ significantly, the voice was replayed in random order. Each voice was repeated up to three rounds if disparity in rating existed. Table 6 presents the average MOS rating of various noise types and across all subjects. Although the difference among rating of five methods was limited, it is not difficult to see that our proposed method was rated the best (the highest MOS rating).

We also examined the efficiency of noise suppression of our method. The methods were implemented with MATLAB and the computer used in our experiments includes 4 GB memory, Intel Core i5 2.8 GHz CPU and runs a Windows 7 operating system. The average time used to complete signal denoising is reported in Table 7. Provided with the additional steps to achieve a more accurate description of noise components in wavelet subbands, extra time was required and hence our method was less efficient than the others.

4.4. Speaker recognition performance analysis

Table 8 summarizes the average recognition accuracy across all noise types using ten methods in comparison to ours. The column DEG shows the results of speaker recognition without any signal preprocessing. When noise dominates, the recognition accuracy was extremely poor. The average accuracy reached 0.8% with a standard deviation of 0.37 in the case of SNR = −5 dB. With noise suppression, the recognition rate was improved. The best performance and the second best are highlighted with bold face font and underline in the table, respectively. The right most column lists the improvement rate with respect to the second best case for each SNR. It is evident that our proposed method ABWS exhibits superior recognition performance under various noise cases. The standard deviation (STD) is shown in parenthesis. Except when SNR is −5 dB, our method yielded the smallest STD, which indicates that our method consistently achieved better performance across all noise types and speakers. RASTA [8] method exhibited competitive accuracy. However, the standard deviation is greater than ABWS, which implies that our proposed method is more consistent with greater accuracy.

Fig. 6 illustrates the recognition accuracy under ten different noise types. The height of each bar indicates the correct recognition rate, which is the average among all subjects. In almost all

Table 8
Average recognition accuracy (%) across all noise types and speakers. The value in parentheses gives the standard deviation. DEG refers to degraded speech signal and the results show the recognition accuracy without any preprocessing. IMP denotes the improve rate.

SNR	DEG	SS	IWF	EMF	BWT	HMM	GMM	GMM-EM	GMM-CMN	RASTA	ABWS
−5 dB	0.8	6.6	6.0	11.7	11.4	17.8	18.8	20.0	20.0	21.5	**22.9**
	(0.37)	(1.18)	(1.95)	(2.62)	(2.44)	(1.12)	(1.22)	(1.57)	(2.68)	(2.64)	(2.09)
0 dB	4.5	22.7	19.8	29.6	29.0	37.4	37.1	39.8	40.1	41.2	**42.5**
	(2.35)	(4.52)	(4.29)	(5.24)	(5.17)	(2.29)	(3.38)	(2.43)	(3.07)	(3.08)	(1.57)
5 dB	12.8	44.7	41.0	54.0	53.1	46.6	47.7	50.0	59.5	61.1	**62.0**
	(3.89)	(8.14)	(6.8)	(6.88)	(6.04)	(11.54)	(11.37)	(11.50)	(2.87)	(3.27)	(2.39)
10 dB	18.6	58.2	54.6	65.3	65.4	62.8	66.1	67.1	72.5	74.2	**75.8**
	(3.24)	(7.94)	(9.69)	(10.01)	(8.88)	(12.57)	(8.22)	(8.21)	(2.71)	(3.01)	(2.46)
15 dB	27.1	68.8	65.5	76.5	75.9	74.8	76.5	78.9	81.4	83.0	**83.6**
	(6.79)	(4.55)	(4.92)	(3.66)	(3.31)	(4.42)	(3.59)	(4.47)	(3.07)	(2.05)	(0.53)

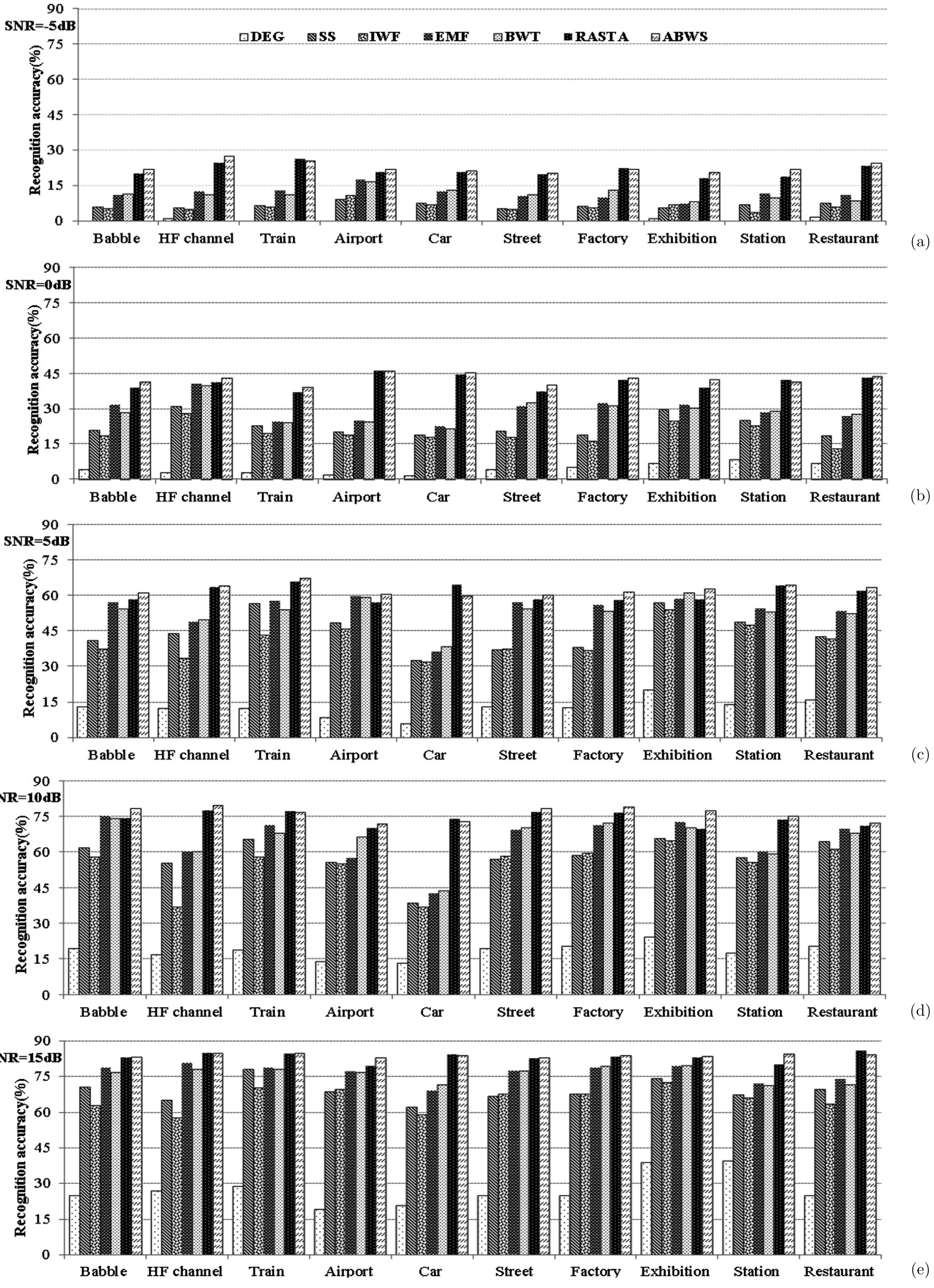

Fig. 6. Recognition accuracy under ten different noise types. (a)–(e) depict the accuracies achieved by processing signals with different SNRs. Each group of bars consists of results for DEG, SS, IWF, EMF, BWT, RASTA and our method (ABWS), from left to right, respectively. The legend is shown on the top of (a).

cases, our method exhibits the best performance. It is interesting to note that our proposed method handles car, station, and restaurant noises most effectively on average. Compared to the second best results in each SNR scenario, our method improved the accuracy by about 100% when SNR $= -5$ dB, 0 dB, and by about 50% when SNR $= 5$ dB, 10 dB, and about 20% when SNR $= 15$ dB. It is evident that our method improved the recognition performance greatly, especially when noise was significant in the recorded voice signals. In some types of noise, e.g., car noise, our method exhibited superior performance consistently across all SNR ratios, which is very important in real-world applications as mobile computing becomes pervasive.

Please cite this article in press as: S.M. Govindan et al., Adaptive wavelet shrinkage for noise robust speaker recognition, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.06.007

5. Conclusions

In this paper we describe a noise robust speaker recognition method using ABWS. Without knowing the noise characteristics or assuming a model of the underlying clean speech signal, our method leverages the sparsity of wavelet coefficients and automatically suppresses noise to improve recognition accuracy. Compared to the wavelet shrinkage based methods, the proposed method automatically decides the subband coefficient thresholds that are proportional to the amount of noise contamination. The DuTS in our proposed shrinkage method ensures intact signal coefficients for the ones with large amplitude as well as minimum artifact through gradual amplitude suppression in the range of the two thresholds.

The evaluation is conducted using speech signals from two public available speech databases: TIMIT database [44] and King database [45]. Synthetic noisy signals are created by blending clean voice signal with artificial noise. Ten different types of noise were used, which represent a group of common noise sources. It was demonstrated that MFCC and MMFCC features from speech signal of small segment size (512 samples per segment) effectively capture the speech characteristics.

Both objective and subjective metrics were employed in our evaluation. It is evident that our proposed method exhibited great robustness in various noise conditions, especially when noise was significant in the recorded voice signals. The comparison study with state-of-the-art methods also demonstrated the superior performance of the new method. Compared to the second best results in the SNR scenarios, our method yielded improved results consistently.

References

[1] H. Beigi, Fundamentals of Speaker Recognition, Springer, New York, 2011.
[2] A.K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (1) (2004) 4–20.
[3] T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, H. Li, Low-variance multitaper MFCC feature: a case study in robust speaker verification, IEEE Trans. Audio Speech Lang. Process. 20 (1) (2012) 1990–2001.
[4] S.K. Nemala, K. Patil, M. Elhilali, A multistream feature framework based on bandpass modulation filtering for robust speech recognition, IEEE Trans. Audio Speech Lang. Process. 21 (2) (2013) 416–426.
[5] T. May, S. van de Par, A. Kohlrausch, Noise-robust speaker recognition combining missing data techniques and universal background modeling, IEEE Trans. Audio Speech Lang. Process. 20 (1) (2012) 108–121.
[6] Y. Wang, M.J.F. Gales, Speaker and noise factorization for robust speech recognition, IEEE Trans. Audio Speech Lang. Process. 20 (7) (2012) 2149–2158.
[7] X. Yuan, B. Buckle, A wavelet-based noise-aware method for fusing noisy imagery, in: Proceedings of the International Conference on Image Processing, San Antonio, TX, September 2007.
[8] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Process. 2 (4) (1994) 578–589.
[9] J. Ming, T.J. Hazen, J.R. Glass, D.A. Reynolds, Robust speaker recognition in noisy conditions, IEEE Trans. Audio Speech Lang. Process. 15 (5) (2007) 1711–1723.
[10] T. Hasan, J.H.L. Hansen, A study on universal background model training in speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (7) (2011) 1890–1899.
[11] L. Zao, R. Coelho, Colored noise based multicondition training technique for robust speaker identification, IEEE Signal Process. Lett. 18 (11) (2011) 675–678.
[12] D.K. Kim, M.J.F. Gales, Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition, IEEE Trans. Audio Speech Lang. Process. 19 (2) (June 2010) 315–325.
[13] L. Deng, A. Acero, M. Plumpe, X. Huang, Large vocabulary speech recognition under adverse acoustic environments, in: Proceedings of the Sixth International Conference on Spoken Language Processing, Beijing, China, October 2000, pp. 806–809.
[14] Y.F. Liao, Z.H. Chen, Y.T. Juang, Latent prosody analysis for robust speaker identification, IEEE Trans. Audio Speech Lang. Process. 15 (6) (2007) 1870–1883.
[15] M.T. Padilla, T.F. Quatieri, D.A. Reynolds, Missing feature theory with soft spectral subtraction for speaker verification, in: Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, Pennsylvania, September 2006, pp. 913–916.
[16] Z. Brajevic, A. Petosic, Signal denoising using STFT with Bayes prediction and Ephraim–Malah estimation, in: Proceedings of the 54th International Symposium ELMAR, Zadar, Croatia, September 2012, pp. 183–186.
[17] M.A. Abd El-Fattah, M.I. Dessouky, A.M. Abbas, S.M. Diab, E.M. El-Rabaie, W. Al-Nuaimy, Speech enhancement with an adaptive Wiener filter, Int. J. Speech Technol. 17 (1) (2013) 53–64.
[18] A. Panda, E. Srikanthan, Psychoacoustic model compensation for robust speaker verification in environmental noise, IEEE Trans. Audio Speech Lang. Process. 20 (3) (2012) 945–953.
[19] M.T. Johnson, Y. Xiaolong, R. Yao, Speech signal enhancement through adaptive wavelet thresholding, Speech Commun. 49 (2) (2007) 123–133.
[20] S. Mallat, W.L. Hwang, Singularity detection and processing with wavelets, IEEE Trans. Inf. Theory 38 (2) (1992) 617–643.
[21] D. Donoho, I. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, J. Am. Stat. Assoc. 90 (432) (1995) 1200–1224.
[22] I.M. Johnstone, B.W. Silverman, Wavelet threshold estimators for data with correlated noise, J. R. Stat. Soc. 59 (2) (1997) 319–351.
[23] J. Yao, Y.T. Zhang, Bionic wavelet transform: a new time–frequency method based on an auditory model, IEEE Trans. Biomed. Eng. 48 (8) (2001) 856–863.
[24] M. Bahoura, J. Rouat, Wavelet speech enhancement based on the Teager energy operator, IEEE Signal Process. Lett. 8 (1) (2001) 10–12.
[25] D. Khaled, S. Ibrahim Abu, D. Omer, K. Emad, An investigation of speech enhancement using wavelet filtering method, Int. J. Speech Technol. 12 (2) (2013) 101–115.
[26] Y. Ghanbari, M.R. Karami-Mollaei, A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets, Speech Commun. 48 (8) (2006) 927–940.
[27] C. Giguere, P.C. Woodland, A computational model of the auditory periphery for speech and hearing research, J. Acoust. Soc. Am. 95 (1) (1994) 331–342.
[28] J. Yao, Y.T. Zhang, The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations, IEEE Trans. Biomed. Eng. 49 (11) (2002) 1299–1309.
[29] S. Mallat, A Wavelet Tour of Signal Processing, second edition, Academic Press, 1999.
[30] X. Yuan, B. Buckle, Subband noise estimation for adaptive wavelet shrinkage, in: Proceedings of the International Conference of Pattern Recognition, Cambridge, UK, August 2004.
[31] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus, http://catalog.ldc.upenn.edu/ldc93s1, 1993.
[32] A. Higgins, D. Vermilyea, KING speaker verification, http://catalog.ldc.upenn.edu/ldc95s22, 1995.
[33] Joseph P. Campbell Jr., D.A. Reynolds, Corpora for the evaluation of speaker recognition systems, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Phoenix, AZ, March 1999, pp. 829–832.
[34] L. Kaur, S. Gupta, R.C. Chauhan, Image denoising using wavelet thresholding, in: Proceedings of the Third Conference on Computer Vision, Graphics and Image Processing, vol. 1, 2002, pp. 100–105.
[35] F. Itakura, S. Saito, Analysis synthesis telephony based on the maximum likelihood method, in: Proceedings of the 6th International Congress on Acoustics, Tokyo, Japan, August 1968, pp. C 17–20.
[36] F. Ribeiro, D. Florncio, C. Zhang, M. Seltzer, CrowdMOS: an approach for crowdsourcing mean score studies, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, 22–27 May 2011, pp. 2416–2419.
[37] B. Yegnanarayana, S.R.M. Prasanna, J.M. Zachariah, C.S. Gupta, Combining evidence from source, supra segmental and spectral features for a fixed-text speaker verification system, IEEE Trans. Speech Audio Process. 13 (4) (2005) 575–582.
[38] W. Hong, J. Pan, Modified MFCCs for robust speaker recognition, in: Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems, vol. 1, Xiamen, China, October 2010, pp. 276–279.
[39] G. Doddington, M. Przybycki, A. Martin, D. Reynolds, The NIST speaker recognition evaluation – overview, methodology, systems, results, perspective, Speech Commun. 31 (2000) 225–254.
[40] J. Rajnoha, P. Pollak, Modified feature extraction methods in robust speech recognition, in: Proceedings of the 17th IEEE International Conference on Radioelektronika, 2007, pp. 1–4.
[41] P. Kumar, A. Biswas, A.N. Mishra, M. Chandra, Spoken language identification using hybrid feature extraction methods, J. Telecommun. 1 (2) (March 2010) 11–15.
[42] Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantizer design, IEEE Trans. Commun. 28 (1) (1980) 84–95.
[43] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 693–696.
[44] Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Acoust. Speech Signal Process. 16 (1) (2008) 229–238.
[45] V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond, Speech Commun. 9 (4) (1990) 351–356.

Please cite this article in press as: S.M. Govindan et al., Adaptive wavelet shrinkage for noise robust speaker recognition, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.06.007

Dr. Sumithra Manimegalai Govindan received B.E. degree in Electronics and Communication Engineering from Government College of Engineering, Salem, India, M.E. degree in Medical Electronics from College of Engineering, Anna University Chennai and Ph.D. in Information and Technology from Anna University, Chennai. She is currently Professor in Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu. She is a member of IEEE, ISTE, IACSIT and IAENG and a reviewer for referred journals. Her research interests lie in the area of signal processing and image processing, biomedical and wireless communications. She has published 42 technical papers in national and international journals and 95 technical papers in national and international conferences.

Dr. Prakash Duraisamy received his B.E. degree in Electronics and Communication Engineering from the Bharathiar University, Coimbatore, India, in 2002 and his master's degree in electrical engineering from the University of South Alabama, Mobile, Alabama, in 2008. He completed his Ph.D. degree in computer science and engineering at the University of North Texas, Denton, Texas in 2012. Since 2013, he has been working as a visiting scientist at Massachusetts Institute of Technology, and he also worked as a postdoctoral fellow at the Old Dominion University, Virginia. His areas of interest are image processing, computer vision, pattern recognition, bio-medical, LiDAR, remote sensing, and multiple-view geometry.

Dr. Xiaohui Yuan received B.S. degree in Electrical Engineering from Hefei University of Technology, China in 1996 and Ph.D. degree in Computer Science from Tulane University in 2004. After graduation, he worked at the National Institutes of Health on medical imaging and analysis till 2006. He joined the University of North Texas and is currently an Associate Professor. His research interests include computer vision, data mining, machine learning, and artificial intelligence. His research findings are reported in over 80 peer-reviewed papers. Dr. Yuan is a recipient of Ralph E. Powe Junior Faculty Enhancement award in 2008 and the Air Force Summer Faculty Fellowship in 2011, 2012, and 2013. He also received two research awards and a teaching award from UNT in 2007, 2008, and 2012, respectively. He served in the editorial board of several international journals and as session chairs in many conferences.