

# Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition

Jia-Ching Wang<sup>1</sup> · Chien-Yao Wang<sup>1</sup> · Yu-Hao Chin<sup>1</sup> ·  
Yu-Ting Liu<sup>2</sup> · En-Ting Chen<sup>2</sup> · Pao-Chi Chang<sup>2</sup>

Received: 3 April 2015 / Revised: 5 October 2015 / Accepted: 4 February 2016 /  
Published online: 17 February 2016  
© Springer Science+Business Media New York 2016

**Abstract** This paper proposes a speaker recognition system using acoustic features that are based on spectral-temporal receptive fields (STRFs). The STRF is derived from physiological models of the mammalian auditory system in the spectral-temporal domain. With the STRF, a signal is expressed by rate (in Hz) and scale (in cycles/octaves). The rate and scale are used to specify the temporal response and spectral response, respectively. This paper uses the proposed STRF based feature to perform speaker recognition. First, the energy of each scale is calculated using the STRF representation. A logarithmic operation is then applied to the scale energies. Finally, a discrete cosine transform is utilized to the generation of the proposed STRF feature. This paper also presents a feature set that combines the proposed STRF feature with conventional Mel frequency cepstral coefficients (MFCCs). The support vector machines (SVMs) are adopted to be the speaker classifiers. To evaluate the performance of the proposed speaker recognition system, experiments on 36-speaker recognition were conducted. Comparing with the MFCC baseline, the proposed feature set increases the speaker recognition rates by 3.85 % and 18.49 % on clean and noisy speeches, respectively. The experiments results demonstrate the effectiveness of adopting STRF based feature in speaker recognition.

**Keywords** STRF · Speaker recognition · Feature extraction · Speaker authentication

## 1 Introduction

With the rapid growth of the usages of portable devices, internet of things (IOT) has seen considerable development. Users may use these devices everywhere, and the problem

---

✉ Pao-Chi Chang  
pcchang@ce.ncu.edu.tw

<sup>1</sup> Department of Computer Science and Information Engineering, National Central University, Zhongli, Taiwan

<sup>2</sup> Department of Communication Engineering, National Central University, Zhongli, Taiwan

associated with user authentication of these devices is an important issue. Since many devices can only be accessed after authentication, the repeated input of account information and passwords is inconvenient for accessing these devices. Therefore, a natural user interface (NUI) is required to solve this problem. Multi-touch, gesture-based, and speech-based NUIs are popular nowadays. They depend on devices such as a touch screen, Kinect, and a microphone to receive and process a signal. Among these NUIs, speech can also be used for identification because it is a unique biometric. In this work, the biometric characteristics of speech are used to recognize the user identity.

The two types of the speaker authentication system are - text-dependent [2, 3, 20] and text-independent [12, 15, 17, 28]. A text-dependent speaker authentication system is similar to traditional authentication systems of speakers. The speaker passes the system upon speaking the pre-specified password. A text-independent speaker authentication system models the acoustic characteristics of a speaker and utilizes the acoustic models to identify speakers. Since a text-independent speaker authentication system is not limited by text, its use is more convenient. This work proposes a text-independent speaker recognition system to authenticate a user.

In a speaker recognition system, audio feature extraction and classifier modeling are the two main components. Audio feature extraction comes behind the pre-processing stage. Previous researches have reported various types of audio features include amplitude and power in the temporal domain, chroma and harmonicity in the frequency domain, and cepstral coefficients in the cepstral domain. The most frequently used audio features are cepstral coefficients [14]. Cepstral coefficients can be extracted by two different approaches [21]. One is the parametric approach, which is developed to match closely the resonant structure of the human vocal tract that produces speech sounds. This approach is mainly based on linear predictive analysis. The linear predictive coefficients (LPCs) obtained can be converted to LPC cepstral coefficients (LPCCs). The other approach is non-parametric and models the human auditory perception system. Mel frequency cepstral coefficient (MFCC) [27] is categorized as the second approach.

Model classifier also plays an important role in a speaker recognition system. The Gaussian mixture model (GMM) [10, 18, 31], the support vector machine (SVM) [1, 19, 25, 28], the neural network (NN) [15, 30], and hybrid of these models [4, 5, 11] are commonly used for this purpose. A GMM-based recognition system uses a GMM of the probability density function of a speech signal with Gaussian components. The mean, variance and weights of the GMM speaker model can be used to recognize a speaker by maximizing the log likelihood. The SVM is a binary classifier, which makes decisions by finding the optimal hyper-plane that separates positive from negative data. An SVM can also map the features of a speaker to high-dimensional space and perform classification. In many works, GMM is combined with the SVM method to recognize speakers. In the audio signal processing method, a NN usually replaces GMM, and the NN can train a model for multiple speakers. In this paper, an SVM is adopted as the classifier.

To make a speaker recognition system more useful, the following two issues are essential to be considered. The first issue is that natural speech is easy to be copied and a speaker recognition system can be fooled by using voice conversion software [9]. However, people seldom whisper, so recording whispered speeches is much more difficult than recording normal speeches. A whisper based speaker recognition system is therefore safer than a normal speech-based speaker recognition system [23]. The second issue is whether a speaker recognition system is robust against background noise. In practical situations, speech utterances

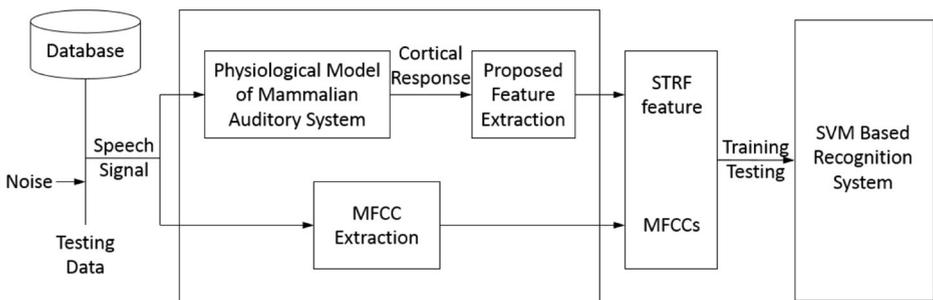
rarely occur in isolation. The captured speech signal is usually accompanied by environmental noises which makes impact on speaker recognition performance. In this paper, we address the second issue and present an audio feature based method to enhance the speaker recognition performance under noisy environments. The used audio feature is derived based on the theory of spectral-temporal receptive fields (STRFs).

The STRF multi-resolution analysis in the spectral and temporal domain is a computational auditory analysis model inspired by psychoacoustical and neurophysiological studies in early and central stages of the mammalian auditory system [8]. This auditory analysis model consists of two primary stages. The first stage is an early auditory system that simulates processing at the auditory periphery and produces an auditory spectrum. The second stage is the primary auditory cortex (A1) model in the central auditory system, which is based on the assumption that a response neuron area tuned to a specific range of frequencies and intensities. In recent years, this theory has been applied to speech recognition research by the work of Woojay et al. [29]. Woojay et al. [29] used STRF theory to propose a new feature selection method that paralleled the computation of the MFCC. The authors' previous work [26] discusses the conception of STRF and reveals that the proposed STRF-based feature can be used in speech recognition. In this paper, we further exploit the possibility to use STRF-based feature in speaker recognition under noisy environments. Our work is not the first attempt to use STRF in speaker recognition. Chi et al. presented a making based speaker identification algorithm [7]. This algorithm distinguishes speech from non-speech based on spectro-temporal modulation energies and can enhance the speaker recognition in noisy environments.

The remaining parts of the paper are organized as follows. Section 2 provides an overview of the proposed speaker recognition system. Section 3 clearly describes the STRF theory and its audio feature extraction method. Section 4 shows the experimental results under various conditions, and Section 5 draws conclusions.

## 2 System overview

Figure 1 shows an overview of the proposed speaker recognition system. The proposed system contains three modules, which are implemented in the following order: the pre-processing module, the feature extraction module, and the recognition module.



**Fig. 1** Block diagram of proposed speaker recognition system

In the first module, we perform framing and voice activity detection as the pre-processing to input speech. In the second module, the MFCC and the STRF-based features are extracted. Conventional speaker recognition usually merely adopts MFCCs as acoustical features. The MFCC takes into account the nonlinear frequency resolution which can simulate the hearing characteristics of human ears. However, these considerations are only crude approximations of the auditory periphery. Therefore, the STRF-based feature is fused with conventional MFCCs to become an effective acoustical feature set for speaker recognition. The STRF theory comes from a physiological model of the mammalian auditory system and generates the cortical response. We convert the cortical response into the useful STRF-based feature. In the third module, SVM models are trained or tested for the speaker recognition task.

### 3 Speaker feature extraction

The STRF has two stages [8, 26]. The first stage is an early auditory system for simulating human hearing system, which generates auditory spectrograms. The second stage is a model of the primary auditory cortex (A1) in the central auditory system.

#### 3.1 Early auditory system

Firstly, an affine wavelet transformation of the signal  $s(t)$  is performed by passing audio signal  $s(t)$  through a cochlear filter bank. The cochlear output  $y_C$  is obtained by

$$y_C(t, f) = s(t) *_t h(t, f) \quad (1)$$

where  $h(t, f)$  denotes the impulse response of each filter, and  $*_t$  represents the convolution operation in the time domain.

Next, the cochlear output  $y_C$  is fed into the hair cell stage. This stage consists of a high-pass filter to change the pressure into a moving speed; a nonlinear compression function  $g(u)$  to protect; a low-pass filter  $w(t)$  to reduce phase-locking on the auditory-nerve.

$$y_A(t, f) = g(\partial_t y_C(t, f)) *_t w(t) \quad (2)$$

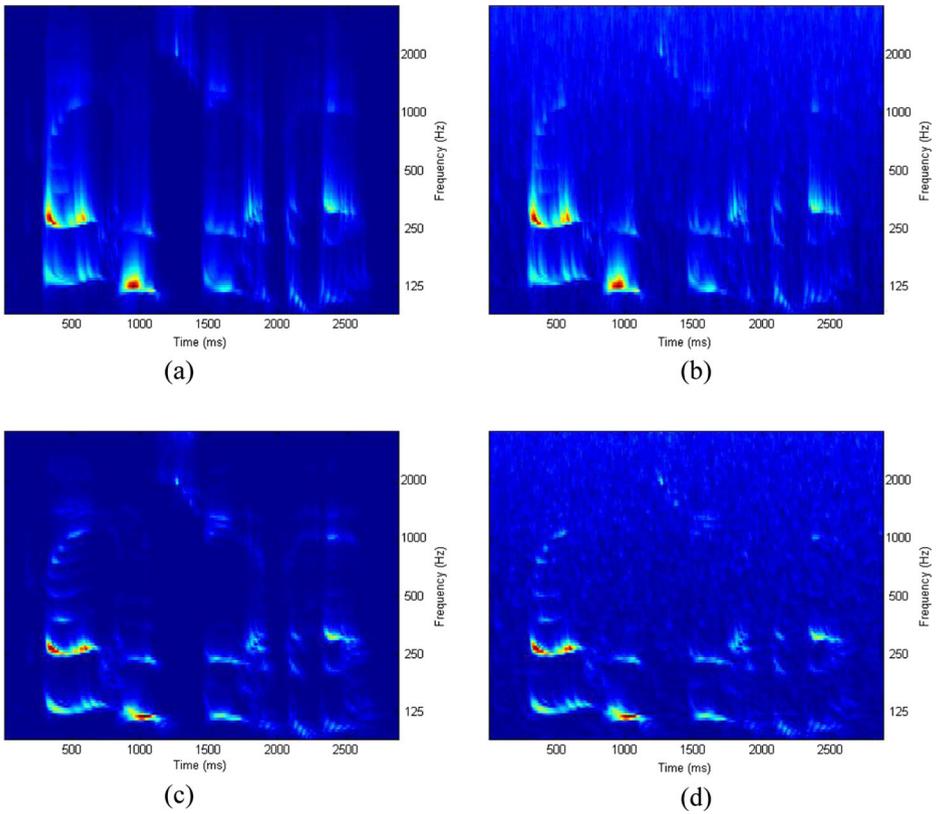
The lateral inhibitory network (LIN) in the cochlear nucleus performs frequency selectivity. As in Eq. (3), taking the first-order derivative with respect to the tonotopic axis first and then passing it through a half-wave rectifier can approximate the function of LIN.

$$y_{LIN}(t, f) = \max(\partial_f y_A(t, f), 0) \quad (3)$$

Finally, the output of the first stage is obtained by convoluting  $y_{LIN}(t, f)$  with a short window.

$$y(t, f) = y_{LIN}(t, f) *_t \mu(t, \tau) \quad (4)$$

where  $y(t, f)$  is the obtained auditory spectrogram,  $\mu(t, \tau) = e^{-\frac{t}{\tau}} u(t)$ , and  $\tau$  is the time constant in msec. Figure 2 gives examples to compare conventional Fourier spectrograms and auditory spectrograms.



**Fig. 2** Examples comparing Fourier spectrograms and auditory spectrograms: **a** Fourier spectrogram of a clean speech; **b** Fourier spectrogram of a noisy speech; **c** auditory spectrogram of a clean speech; **d** auditory spectrogram of a noisy speech

### 3.2 Model of primary auditory cortex (A1)

The so called STRF is the product of a spatial impulse response  $h_S$  and the temporal impulse responses  $h_T$ :

$$STRF = h_S \cdot h_T \tag{5}$$

Equations (6) and (7) give the spatial impulse response  $h_s$  and the temporal impulse response  $h_T$ , respectively.

$$h_S(f, \omega, \theta) = h_{scale}(f, \omega) \cos \theta + \hat{h}_{scale}(f, \omega) \sin \theta \tag{6}$$

$$h_T(t, \Omega, \varphi) = h_{rate}(t, \Omega) \cos \varphi + \hat{h}_{rate}(t, \Omega) \sin \varphi \tag{7}$$

where  $\omega$  and  $\Omega$  denote the spatial density and velocity parameters of the spatial impulse response and temporal impulse response filters, respectively. Besides,  $\hat{h}$  denotes the following Hilbert transform, and  $\theta$  and  $\varphi$  represent the characteristic phases.

For an input auditory spectrogram  $y(t, f)$ , the spectral-temporal response  $STRF(t, f, \Omega, \omega, \varphi, \theta)$ , i.e. cortical representation, is calculated by

$$STRF(t, f, \Omega, \omega, \varphi, \theta) = y(t, f) *_{tf} [h_S(f, \omega, \theta) \cdot h_T(t, \Omega, \varphi)] \tag{8}$$

where  $*_{tf}$  represents the convolution operation in both the time and frequency domains.

Figure 3 gives a scale-rate representation of the 36th frame of the speech signal in Fig. 2. The horizontal axis represents the rate parameter in both downward and upward directions, while the vertical axis presents the scale response. Figure 4 shows the MFCC of the speech signal in Fig. 2. By comparing Fig. 4 with 3, the noise robustness of STRF is better than MFCC.

### 3.3 STRF based feature extraction

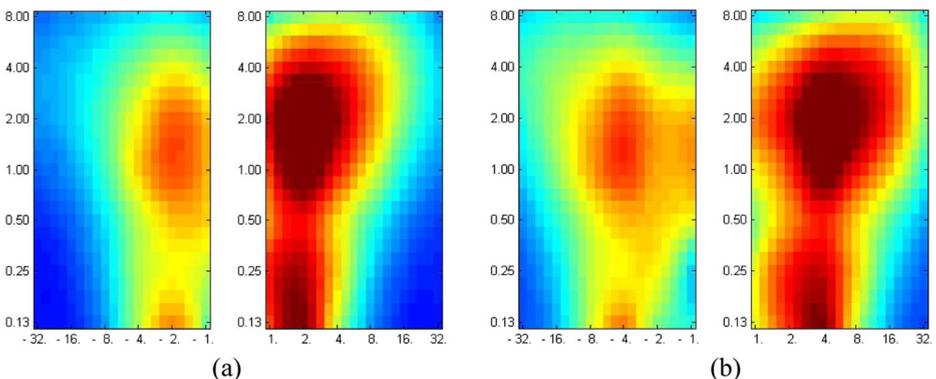
The STRF representation reveals joint spectral-temporal modulations of the auditory spectrogram. In STRF, the low and high scales capture the formants and harmonics, respectively. This scale information is utilized to derive the STRF based features. The scale values from  $2^{-3}$  to  $2^3$  with the intervals of 0.5 cycle/octave are taken to span the frequency range of the auditory spectrum. In this paper, three different STRF based features are derived from the STRF representation [26]. The first STRF based feature  $S(t, \omega)$  is obtained by summing the all the rates and frequency in the STRF magnitude representation.

$$S(t, \omega) = \sum_f \sum_{\Omega} |STRF(t, f, \Omega, \omega, 0, 0)|, \quad \omega = 1, 2, \dots, N_{\omega}. \tag{9}$$

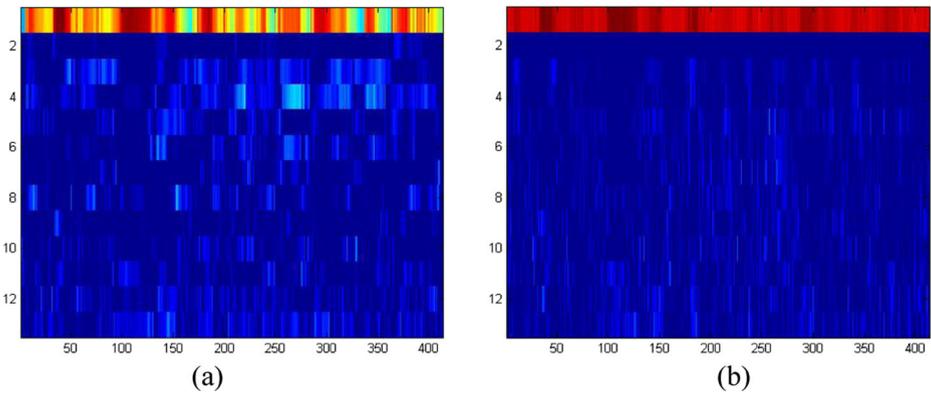
where  $N_{\omega}$  is the scale number. Noted that the characteristic phases  $\varphi$  and  $\theta$  in Eq. (9) are both set to zero [29].

Next, a non-linear function is applied to  $S(t, \omega)$  to generate the second STRF based feature  $S_L(t, \omega)$ . Here, a logarithmic function is utilized for this purpose.

$$S_L(t, \omega) = \log(S(t, \omega)), \quad \omega = 1, 2, \dots, N_{\omega}. \tag{10}$$



**Fig. 3** Scale-rate representation: **a** scale-rate representation of a clean speech; **b** scale-rate representation of a noisy speech



**Fig. 4** MFCC representation: **a** MFCC representation of a clean speech; **b** MFCC representation of a noisy speech

A discrete cosine transform (DCT) [14] is applied to  $S_L(t, \omega)$  to create the third STRF based feature  $S_{DL}(t, k)$ .

$$S_{DL}(t, k) = \sum_{\omega=1}^{N_\omega} S_L(t, \omega) \cos\left(\frac{2\pi\omega k}{N_\omega}\right), \quad k = 1, 2, \dots, N_k \tag{11}$$

where  $N_k$  is the feature dimension of  $S_{DL}(t, k)$ , and  $N_k$  is smaller or equal to  $N_\omega$ .

### 4 Speaker classifier

Conventional speaker recognition uses GMM to model the acoustics characteristics of speakers. The parameters of GMM are estimated through maximizing the likelihood of the speaker data. Alternatively, the SVM has already proven its performance in speaker recognition [1, 17, 28]. An SVM is capable of discriminating 2-class of speaker data by training a nonlinear decision boundary.

Considering speech features from two different speakers, an SVM is able to find out the maximum margin hyperplane [22]. Assume  $\{(\mathbf{x}_i, s_i), i = 1, 2, \dots, N\}$  are the training speaker data. A pair  $(\mathbf{x}_i, s_i)$  denotes training speech feature  $\mathbf{x}_i$  was from speaker  $s_i$ , where  $s_i \in \{+1, -1\}$ . After learning the maximum margin hyperplane by an algorithm such as sequential minimal optimization [16], a testing speech feature  $\mathbf{x}$  is classified as  $s \in \{+1, -1\}$  based on the following speaker decision function:

$$s(\mathbf{x}) = \text{sign}((\mathbf{w} \cdot \mathbf{x}) + b). \tag{12}$$

where  $\mathbf{w} \cdot \mathbf{x} + b = 0$  is the maximum margin hyperplane. A kernel version of the speaker decision function can be derived as

$$s(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \mu_i k(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{13}$$

where  $k$  is a kernel function, and  $\mu_i$  is a Lagrange multiplier.

**Table 1** Performance of each feature

Speech feature	Dimensionality	Recognition rate (%)
MFCC13	13	93.07
$S$	13	54.39
$S_L$	13	52.54
$S_{DL}$	13	70.26

Multiple speaker recognition can be solved by a multi-class SVM, which is built based on a required number of 2-class SVMs. In this paper, the multi-class SVM is achieved by a one-against-one method [13, 24]. A voting strategy is required to determine the final speaker recognition result. For recognizing  $m$  speakers, the required number of 2-class SVM is  $m(m-1)/2$ .

## 5 Experimental results

### 5.1 Experimental setting

Speech samples are taken from a dataset that contains 36 speakers. For each speaker, 37 speech clips were recorded. The duration of each clip for training and testing ranges roughly from 1 s to 2 min. In the experiments, all speech clips were divided into frames of 16 ms with an overlapping between consecutive frames of 8 ms. The Hamming-windowing was applied to each frame. Finally, the MFCC feature and proposed feature were extracted. The two-fold cross-validation was used to estimate the performance of the proposed speaker recognition system.

The proposed STRF feature parameter was set using a scale of  $2^{-3,-2,\dots,3}$ . The settings of SVM and the parameters are the default settings of the LIBSVM tool [6]. The multi-class SVM with radial basis function kernel was used to train the models, and the gamma value of the radial basis function kernel was set to 2.

### 5.2 Comparison of the recognition systems

The first experiment was conducted to compare the speaker recognition performance among 13 dimensional MFCCs (MFCC13),  $S$ ,  $S_L$  and  $S_{DL}$ . The experimental result is listed in Table 1. MFCC13 yields a recognition rate of 93.07 %, which is the highest among all the features. The proposed STRF based feature  $S_{DL}$  yields a recognition rate value that is 17.72 % and 15.87 % higher than that of  $S_L$  and  $S$ , respectively. In the next experiment, we tried to combine MFCCs (MFCC13) with STRF based features. The speaker recognition rate generated by each

**Table 2** Performance of combined features

speech feature	Dimensionality	Recognition rate (%)
MFCC13+ $S$	26	96.76
MFCC13+ $S_L$	26	96.30
MFCC13+ $S_{DL}$	26	96.92

**Table 3** Performance for noisy speech

Speech feature	Clean	20	15	10	5	0
MFCC13	93.07	83.05	81.05	78.89	72.73	67.33
$s$	54.39	55.32	55.16	55.16	53.95	51.31
$s_L$	52.54	52.23	51.77	52.85	51.46	50.69
$s_{DL}$	70.26	69.95	68.72	68.41	66.72	68.57
MFCC13 + $s_{DL}$	96.92	93.22	93.53	91.22	90.45	85.82

combined feature set is described in Table 2. The experimental result reveals that all of the combined feature sets increase the recognition rate in comparison with individual MFCCs. The best result of 96.92 % is obtained when  $s_{DL}$  and MFCCs are adopted as the combined feature set.

Since the access control system may be used in everywhere, it is important to consider the problem of speech in noisy environments. Table 3 shows the speaker recognition rate of the noisy speech. In this experiment, noisy speech signals were generated by adding white noise to the clean speech using five different signal-to-noise ratio (SNR) levels: 20db, 15db, 10db, 5db and 0db. In Table 3, the average recognition rate of  $s_{DL}$  is 14.56 and 16.85 % higher than that of  $s$  and  $s_L$ . In other words, feature  $s_{DL}$  yields the most favorable results among the three STRF based features. Considering the proposed  $s_{DL}$  feature, the difference of speaker recognition rates is 1.69 % between clean speech and 0db noisy speech. However, the speaker recognition rate of MFCC feature decreases 25.74 % from clean speech to 0db noisy speech, and  $s_{DL}$  performed a better result than MFCC when using 0db speech. The results show that the proposed  $s_{DL}$  feature is highly robust to noise condition. In addition, the best recognition rate is produced by the fusion of the  $s_{DL}$  and the MFCC13. The speaker recognition rates can achieve over 90 % when the power of speech signal is higher than the power of noise. Conventional MFCC can be interpreted as the information derived from energies of band-pass filter. Combining captured formant and harmonics information from STRF based feature with conventional MFCCs further improves the speaker recognition performance.

In the speaker recognition field, GMM with MFCCs is a widespread used standard method. A GMM based speaker recognition system with 13 dimensional MFCCs is taken as the baseline and its speaker recognition rates are listed in Table 4. Table 4 reveals that the proposed system outperforms GMM baseline under clear and various noisy conditions.

To compare the computational complexity among different features, the feature extraction time of one second of audio signal is obtained. The 13 dimensional MFCCs (MFCC13) and the proposed STRF based feature require 0.18 and 0.75 s, respectively. Although the required computational time of the proposed STRF scale-based feature is about four times of that of MFCCs, the increased computational time is acceptable. In addition, the SVM classification

**Table 4** Performance comparison with baseline

System	Clean	20	15	10	5	0
Baseline	90.74	69.44	70.37	64.81	63.89	56.48
Proposed system	96.92	93.22	93.53	91.22	90.45	85.82

requires roughly 0.05 s in 1 s of audio signal. Therefore, the proposed speaker recognition system meets the real-time requirement.

Our experiments were conducted to recognize 36 speakers. In case the speaker recognition system is used to deal with thousands of speakers, proximity indexes may be required to use. Several percentages increase in recognition rate may vanish when using proximity indexes because of the increase in dimensionality from the combination with STRF based features and MFCCs. However, the main contribution of the paper is its robust speaker recognition performance under noisy conditions. Take audio with SNR being 0 dB or 5 dB as example, the recognition rate of the proposed combined feature set is about 18 % higher than that of conventional MFCCs. Therefore, the improvement of about 18 % recognition rate is significant even when a drop of several percentages in recognition rate does occur after using proximity indexes.

## 6 Conclusions

This work proposes a robust speaker recognition system using a combination of the STRF based features and MFCCs. The STRF model concerns the spectral and temporal variations of an analyzed auditory spectrogram. The scale features reveal many characteristics in the spatial domain, such as formants and harmonics. Therefore, the following new acoustic features that are based on scale features are proposed; the energy of each scale, the logarithmic scale energy, and the DCT coefficients of the logarithmic scale energy. The proposed feature set that consists of STRF based features and MFCCs significantly improve the recognition rates in robust speaker recognition tasks over those achieved using conventional MFCCs.

## References

1. Andrew O. Hatch, Sachin K, Andreas S (2006) Within-class covariance normalization for SVM-based speaker recognition. In: 2006 ICSLP
2. Anthony L, Kong AL, Bin M, Haizhou L (2013) Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances. Human Language Technology Department, Institute for Infocomm Research, A\*STAR, Singapore
3. Anthony L, Kong AL, Bin M, Haizhou L (2015) The RSR2015: database for text-dependent speaker verification using multiple pass-phrases. Institute for Infocomm Research (I2R), A\*STAR, Singapore
4. Campbell WM, Campbell JP, Reynolds DA, Singer E, Torres-Carrasquillo PA (2006) Support vector machines for speaker and language recognition. In: Computer Speech and Language
5. Campbell WM, Sturim DE, Reynolds DA (2006) Support vector machines using GMM supervectors for speaker verification. In: IEEE Signal Processing Letters
6. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. In: ACM Transactions on Intelligent Systems and Technology
7. Chi TS, Lin TH, Hsu CC (2012) Spectro-temporal modulation energy based mask for robust speaker identification. *J Acoust Soc Am* 131(5):368–374
8. Chi TS, Ru P, Shamma S (2005) Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* 118:887–906
9. Desai S, Black AW, Prahallad K (2010) Spectral mapping using artificial neural networks for voice conversion. *IEEE Trans Audio Speech Lang Process* 18(5):954–964
10. Didier M, Andrzej D (2001) Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM). In: ODYSSEY-2001, Crete, Greece.
11. Ding IR, Yen CT (2013) Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications. In: Multimedia Tools and Applications
12. Douglas AR, Richard CR (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. In: IEEE Transactions on Speech and Audio Processing

13. Hsu W, Lin CJ (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw* 13(2):415–425
14. Juang BH, Chen TH (1998) The past, present, and future of speech processing. *IEEE Signal Process Mag* 15(3):24–48
15. Khan SA, Anil ST, Jagannath HN, Vinay SP (2015) A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), Kolkata
16. Kuan TW, Wang JF, Wang JC, Lin PC, Gu GH (2012) VLSI design of an SVM learning core on sequential minimal optimization algorithm. *IEEE Trans Very Large Scale Integr VLSI Syst* 20(4):673–683
17. Kuruvachan KG, Arunraj K, Sreekumar KT, Santhosh KC, Ramachandran KI (2014) Towards improving the performance of text/language independent speaker recognition systems. In: International Conference on Power, Signals, Controls and Computation (EPSCICON)
18. Lukáš B, Pavel M, Petr S, Ondřej G, Jan Č (2007) Analysis of feature extraction and channel compensation in a GMM speaker recognition system. In: *IEEE Transactions on Audio, Speech, and Language Processing*
19. Srinivas V, Santhi rani C, Madhu T (2013) Investigation of decision tree induction, probabilistic technique and SVM for speaker identification. *Int J Signal Process Image Process Pattern Recog* 6(6):193–204
20. Stafylakis T, Kenny P, Ouellet P, Perez J, Kockmann M, Dumouchel P (2013) Text-dependent speaker recognition using PLDA with uncertainty propagation. Centre de Recherche Informatique de Montreal (CRIM), Canada
21. Tuzun OB, Demirekler M, Nakiboglu KB, (1994) Comparison of parametric and non-parametric representations of speech for recognition. In: *Proc. 7th Mediterranean Electrotechnical Conference*, 1994, pp 65–68
22. Vapnik V (1998) *Statistical learning theory*. Wiley, New York
23. Wang JC, Chin YH, Hsieh WC, Lin CH, Chen YR, Siahaan E (2015) Speaker identification with whispered speech for the access control system. *IEEE Trans Autom Sci Eng* 12(4):1191–1199
24. Wang JC, Lee YS, Lin CH, Siahaan E, Yang CH (2015) Robust environmental sound recognition with fast noise suppression for home automation. *IEEE Trans Autom Sci Eng* 12(4):1235–1242
25. Wang JC, Lian LX, Lin YY, Zhao JH (2015) VLSI design for SVM-based speaker verification system. *IEEE Trans Very Large Scale Integr VLSI Syst* 23(7):1355–1359
26. Wang JC, Lin CH, Chen ET, Chang PC (2014) Spectral-temporal receptive fields and mfcc balanced feature extraction for noisy speech recognition. In: *Asia-Pacific Signal and Information Processing Association (APSIPA)*
27. Wang JC, Wang JF, Weng YS (2002) Chip design of MFCC extraction for speech recognition. *Integr VLSI J* 32(1–3):111–131
28. Wang JC, Yang CH, Wang JF, Lee HP (2007) Robust speaker identification and verification. *IEEE Comput Intell Mag* 2(2):52–59
29. Woojay J, Juang BH (2008) Speech analysis in a model of the central auditory system. *IEEE Trans Audio Speech Lang Process* 15(6):1802–1817
30. Yun L, Nicolas S, Luciana F, Mitchell M (2014) A novel scheme for speaker recognition using a phonetically-aware deep neural network. *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Florence
31. Zhe J, Wei H, Xin J (2013) Duration weighted Gaussian mixture model supervector modeling for robust speaker recognition. In: 2013 Ninth International Conference on Natural Computation (ICNC), Shenyang, China



**Jia-Ching Wang** received the M.S. and Ph.D. degrees in Electrical Engineering from National Cheng Kung University, Tainan, Taiwan in 1997 and 2002, respectively. He was an Honorary Fellow of the Department of Electrical and Computer Engineering, University of Wisconsin–Madison in 2008 and 2009. Currently, he is an Associate Professor in the Department of Computer Science and Information Engineering, National Central University. His research interests include signal processing, machine learning, and VLSI architecture design. Dr. Wang is an honorary member of Phi Tau Phi Scholastic Honor Society, a senior member of IEEE, a member of ACM and IEICE.



**Chien-Yao Wang** received the B.S. degree in applied computer science and information engineering from National Central University, Zhongli, Taiwan, in 2013, and he is currently pursuing the PH. D. degree in computer science and information engineering in NCU. His research interests include signal processing, deep learning, and machine learning.



**Yu-Hao Chin** received the B.S. degree in applied information management from National Central University, Zhongli, Taiwan, in 2013, and he is currently pursuing the PH. D. degree in computer science and information engineering in NCU. His research interests include signal processing, pattern recognition, and machine learning.



**Yu-Ting Liu** received the M.S. degree in the Video-Audio Processing Laboratory (VAPLab) in communication engineering in NCU. Her research interests include speech recognition, audio signal processing and deep learning.



**En-Ting Chen** received the M.S. degree in communication engineering in NCU. Her research interests include audio signal processing and speech processing.



**Pao-Chi Chang** received the B.S. and M.S. degrees from National Chiao Tung University, Taiwan, in 1977 and 1979, respectively, and the Ph. D. degree from Stanford University, California, 1986, all in electrical engineering. From 1986 to 1993, he was a research staff member of the department of communications at IBM T. J. Watson Research Center, Hawthorne, New York. At Watson, his work centered on high speed switching systems, efficient network design algorithms, and multimedia conferencing. In 1993, he joined the faculty of National Central University, Taiwan, where he is presently a Professor in the Department of Communication Engineering. In 1994, Dr. Chang established and has headed the Video-Audio Processing Laboratory (VAPLab) in the Electrical Engineering Department and Communication Department of National Central University since. Dr. Chang is the principle investigator for many joint projects with National Science Council (NSC), Institute of Information Industry (III), Chung Hwa Telecommunication Laboratories (TL), and many other companies. His research interests include speech/audio coding, video/image compression, scalable coding, error resilient coding, digital watermarking and data hiding, and multimedia delivery over packet and wireless networks.