Spectral-Temporal Receptive Fields and MFCC Balanced Feature Extraction for Noisy Speech Recognition

Jia-Ching Wang[1], Chang-Hong Lin[1], En-Ting Chen[2], and Pao-Chi Chang[2]
[1] Department of Computer Science and Information Engineering, National Central University, Jhongli, Taiwan
[2] Department of Communication Engineering, National Central University, Jhongli, Taiwan

***Abstract*—This paper aims to propose a new set of acoustic features based on spectral-temporal receptive fields (STRFs). The STRF is an analysis method for studying physiological model of the mammalian auditory system in spectral-temporal domain. It has two different parts: one is the rate (in Hz) which represents the temporal response and the other is the scale (in cycle/octave) which represents the spectral response. With the obtained STRF, we propose an effective acoustic feature. First, the energy of each scale is calculated from the STRF. The logarithmic operation is then imposed on the scale energies. Finally, the discrete Cosine transform is applied to generate the proposed STRF feature. In our experiments, we combine the proposed STRF feature with conventional Mel frequency cepstral coefficients (MFCCs) to verify its effectiveness. In a noise-free environment, the proposed feature can increase the recognition rate by 17.48%. Moreover, the increase in the recognition rate ranges from 5% to 12% in noisy environments.**

Keywords: speech recognition, spectral-temporal receptive fields, Mel frequency cepstral coefficients

I. INTRODUCTION

Speech recognition is a nature human machine interface which has been extensively studied for the last decades [2].Classifier and speech feature extraction are the two main issues in a speech recognition system. Considering the classifier, hidden Markov model (HMM) is the main stream approach and has been successfully employed in numerous speech recognition systems. An HMM describes the sequential characteristics of the non-stationary signals using state transitions and statistical modeling, which are useful in dealing with dynamic parameters. An HMM is usually based on a maximum likelihood estimation (MLE) with the Baum-Welch algorithm. Given the model of the data, the goal of MLE is to maximize the probability of observing the training data [13]. For an HMM based recognizer, the network processing is selected including the number of states, symbols per state, or left to right topology structure. After training, the recognizer can recognize the result for testing data [14].

Considering the feature extraction, the most widely used speech features are cepsral coefficients [2], [4]. The cepsral coefficients can be extracted by the parametric approach, which is mainly based on linear predictive analysis. The linear predictive coefficients can be transformed to LPC cepsral coefficients. The other way to extract cepsral coefficients is the non-parametric method, which models the auditory perception. Mel frequency cepstral coefficients (MFCCs) belong to this type [3].

However, the effectiveness of conventional speech recognition system remains challenging for real applications. Performance for conventional speech recognition system degrades in noisy environments [16] due to the mismatch between the training data and the testing data. Several methods were proposed to handle this problem. Cepstral mean normalization (CMN) is a popular feature compensation method for dealing with convolutional noise [17]. With the prior information of the noise, noise can be suppressed in different transformed domain, e.g. Fourier domain [18] and wavelet domain [19]. Model adaption can change or compensate parameters of speech model to get better performance in noisy environments [20]. Some methods embed the model of the human auditory system to reach the same performance as human ears [21]. Among these approaches, this paper presents a new acoustic feature based on the modeling of human auditory system. The performance of the recognition system would not be limited by the performance of noise estimation.

Auditory neurons have exclusive ability to sense and track the variations in the stimulus spectrum. Neurons in central auditory stations are quickly activated by the dynamic variations in temporal, spectral, and intensity [22]. Many neurophysiological researches on animals [5] reveal that restricted response in the spectro-temporal domain can reply cells in the auditory cortex [23]. In recent years, psychoacoustic studies provide multi-resolution analysis in the spectral and temporal domain, which is called spectral-temporal receptive fields (STRFs). The STRF proposes new insights into a physiological model of the mammalian auditory system that was originally developed by Chi, Ru, and Shamma [5] and [6]. Recently, this theory has been applied by Woojay *et al.* to speech recognition research [7]. They used STRF theory to propose a new feature selection method that paralleled the computation of the MFCC. In this paper, we discuss the conception of STRF, and proposed a method of extracting the features.

This paper is organized as follows. Section II introduces the system overview. Section III describes the details of STRF representation and the proposed feature extraction method. Section IV shows the experimental results. Finally, conclusions are shown in Section V.

978-616-361-823-8

II. System Overview

The block diagram of the proposed speech recognition system using MFCCs and STRF scale based features is illustrated in Fig. 1. First, the clean speech or the noisy speech is processed by pre-processing, such as framing and voice activity detection (VAD). Second, two types of feature including MFCC and STRF scale based features are extracted. Finally, hidden Markov model (HMM) is used for the recognition task. The main contribution in this system is the use of the proposed STRF based feature. Conventional speech recognition usually adopts MFCCs as acoustical features. The MFCC feature takes into account of 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, STRF feature is proposed herein to be fused with conventional MFCCs [6] to become an effective acoustical feature set.

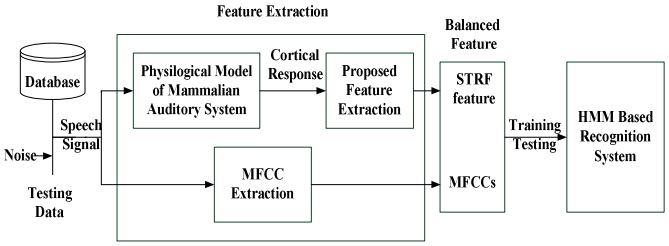

Fig. 1. Block diagram of the speech recognition system using MFCCs and STRF based features.

III. Proposed Feature Extraction

The STRF consists of two stages [5]. The first stage is a central model simulating human hearing system which can generate an auditory spectrum, and the second stage is a model of primary auditory cortex (A1) in the central auditory system.

Firstly, an affine wavelet transform of the signal $s(t)$ is implemented by passing $s(t)$ through 128 band-pass filters with center frequencies that are uniformly distributed along a logarithmic frequency axis. The cochlear output y_C is obtained as

$$y_C(t,f) = s(t) *_t h(t,f) \tag{1}$$

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

Next, the cochlear output y_C is sent into the hair cell stage. This stage comprises a high-pass filter for changing pressure into moving speed, a nonlinear compression function $g(u)$ for protection purpose, and a low-pass filter $w(t)$ for decreasing the phase-locking on the auditory-nerve.

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

The final transformation simulates the response of a lateral inhibitory network. The approximation can be simply achieved by first order derivative with respect to the tonotopic axis.

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

The output of the early stage is obtained by integrating $y_{LIN}(t,x)$ over a short window

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

where $\mu(t,\tau) = e^{\frac{-t}{\tau}} u(t)$, with the time constant τ msec. Figure 2 shows an example of auditory spectrogram.

The STRF is composed of the product of a spatial impulse responses h_S and temporal impulse responses h_T as in (5).

$$STRF = h_s * h_T \tag{5}$$

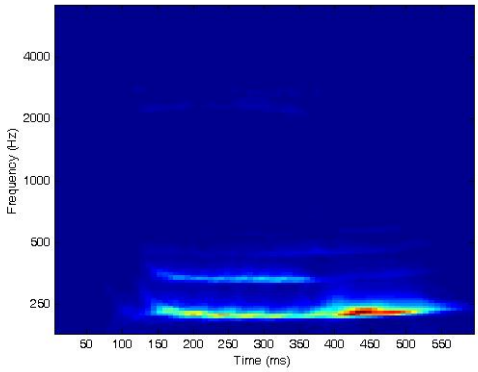

Fig. 2. Auditory spectrogram.

Spatial impulse response h_s (in cycle/octave) and temporal impulse response h_T (in Hz) are represented in (6) and (7), 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 Ω and ω are spatial density and velocity parameters of the filters; φ and θ are characteristic phases. Besides, $\hat{h}$ in Eqs. (6) and (7) denotes the Hilbert transform

$$\hat{h}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{h(z)}{z-x} dz \tag{8}$$

The h_{scale} and h_{rate} are approximated by Eqs. (8) and (9), which are Gaussian and Gamma functions, respectively.

$$h_{scale}(f) = (1-f^2)e^{\frac{-f^2}{2}} \tag{9}$$

$$h_{rate}(t) = t^3 e^{-4t}\cos(2\pi t) \tag{10}$$

For a response area of a cell $STRF(t, f, \Omega, \omega)$, input spectrum $y(t,f)$ can be represented by:

$$STRF(t,f,\Omega,\omega,\varphi,\theta) = y(t,f) * [h_S(f,\omega,\theta)\cdot h_T(t,\Omega,\varphi)] \tag{11}$$

Rate parameter can be dilated into two moving directions: one is downward rate and the other is upward rate. Examples of upward moving rate and downward moving rate are given in

Fig. 3. The rate is 8 Hz and the scale is 1 cyc/oct in Fig. 3. Besides, scale represents how broad the local envelop distributes along the frequency axis. Figure 4 gives the scale-rate representation of the 60th frame of the audio signal in Fig. 2. The horizontal axis represents the rate parameter in both downward and upward directions, while the vertical axis shows the scale response. Take main energy responses at 450 msec as an example, it reflects the downward shift in the pitch or the fundamental frequency.

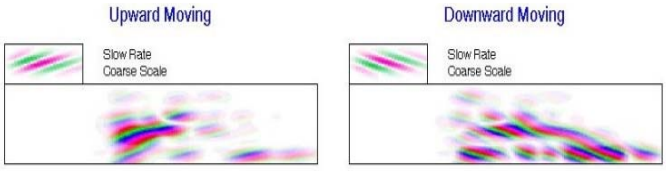

Fig. 3. Upward moving rate and downward moving rate.

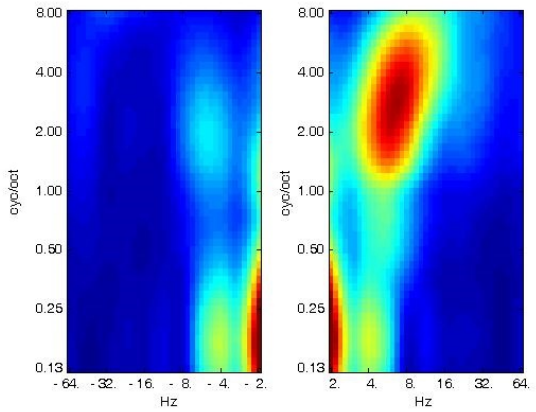

Fig. 4. Scale-rate representation.

The STRF representation reveals joint spectral-temporal modulations of the spectrogram. In this paper, a range of scale $2^{-3,-2,\ldots,3}$ with the interval 0.5 cycle/octave is designed to cover the whole spectrogram. The proposed method is based on the scale information. The characteristic phases φ and θ are set to zero [7]. By summing the magnitude of STRF representation which the rate parameters are the same, the first proposed feature $S(t, \omega)$ is shown in (12).

$$S(t,\omega) = \sum_{f} \sum_{\Omega} \left| STRF(t, x, \Omega, \omega, 0, 0) \right| \tag{12}$$

Next, a logarithmic function is used to calculate the second proposed feature S_L.

$$S_L(t,\omega) = \log(S(t,\omega)) \tag{13}$$

Furthermore, discrete cosine transform (DCT) [8] is applied to S_L to create the third proposed feature S_{DL}.

$$S_{DL}(t,k) = \sum_{\omega=1}^{N} S_{L\omega}(t,\omega) \cos\left(\frac{2\pi\omega k}{N}\right) \tag{14}$$

where N is the length of frame.

TABLE I
RECOGNITION RATE COMPARISON FOR CLEAN SPEECH

Speech Feature	Dimensionality	Recognition Rate (%)
MFCC13	13	68.62
MFCC26	26	70.41
MFCC39	39	75.06
MFCC13 + S_{DL}	26	83.10
MFCC26 + S_{DL}	39	69.65
MFCC39 + S_{DL}	52	66.89

IV. EXPERIENTIAL RESULTS

The speech data used in this study is Chunghwa digits database [15]. Each speaker uttered digits in a quiet environment. Ninety nine speakers are chosen in this corpus. Each speaker speaks " LING", "I" , "ER" , "SAN" , "SU" , "WU" , "LIOU" , "CHI" , "BA" , "JIOU", i.e., digits 0 to 9 in Mandarin. In our implementation, our system cut speech data into frames of 16 ms and the overlapping length between frames is 8ms. The Hamming-windowing is applied to each frame. The left-to right HMMs was used to perform speech recognition.

In the first experiment, three different MFCC based features were calculated as the baselines: 13 dimensional MFCC (MFCC13), MFCC13 concatenated with its first derivative (MFCC26), and MFCC13 concatenated with its first and second derivatives (MFCC36). The three MFCC based features were then respectively concatenated with the proposed STRF feature S_{DL} to generate three new feature sets. The experimental results are listed in Table I. Among the three MFCC based features, MFCC39 provides a recognition rate of 75.06%, which is 6.44% and 4.65% higher than MFCC13 and MFCC26, respectively. After concatenating S_{DL} to MFCC13, the recognition rate reaches as high as 83.10%. However, this concatenation has negative effect in recognition rate for MFCC26 and MFCC39. In the concatenation of S_{DL} and MFCC based features, the more MFCC dimensionality results in the lower recognition rate.

It is important to consider the problem of speech recognition in noisy environments [1]. In the second experiment, noisy speech signals were generated by adding white noise to the clean speech in five different signal-to-noise (SNR) levels: 20dB, 15dB, 10dB, 5dB, and 0dB. Table II gives the experimental results in noisy environments. Considering the scale feature S, the recognition rates are merely about 13-14% in all different SNR cases. After imposing logarithm operation to S, the generated feature S_L provides much better recognition performance. Compared with feature S, the average recognition rate of feature S_L increases by 18.17%. The feature S_{DL} is obtained by applying discrete Cosine transform to feature S_L. In our experimental results, feature S_{DL} yields the most favorable results among the three STRF based features. For clean speech, feature S_{DL} provides a recognition rate of 58.27%. Considering all the SNR cases, the average

recognition rate of feature S_{DL} is 49.17%, which is higher than that yielded by the conventional MFCC features (MFCC13). In addition, the best recognition rate (83.1%) for clean speech condition is also produced by the feature fusion of feature S_{DL} and MFCC13. Conceptually, the MFCC can be interpreted as the information derived from energies of band-pass filter. The proposed feature can give more information about the way frequency changes along the frequency axis. Therefore, feature fusion of the proposed feature and conventional MFCC can improve the recognition rate of conventional MFCC based system.

TABLE II
RECOGNITION RATE (%) COMPARISON FOR NOISY SPEECH

Feature \ SNR	Clean	20	15	10	5	0
MFCC13	68.62	63.79	61.72	58.27	47.93	35.51
S	14.82	14.48	13.79	13.79	13.79	13.10
S_L	38.62	38.62	36.55	35.86	31.03	22.75
S_{DL}	58.27	56.89	53.10	47.93	46.89	41.03
MFCC13+ S_{DL}	80.68	68.96	66.55	62.75	61.37	47.58

V. CONCLUSIONS

In this paper, a new acoustical feature is proposed to combine STRF model and MFCC for speech recognition. The STRF model concerns the spectral and temporal variations of analyzed auditory spectrogram. The scale features reveal many informative characteristics for spatial domain such as formants and harmonics. Hence, we propose new acoustic features based on scale features: the energy of each scale, logarithmic version of scale energy, and the DCT coefficients of the logarithmic scale energy. Comparing to the conventional MFCCs, the proposed STRF features can significantly improve the recognition rates in both noise-free and noisy environments.

REFERENCES

[1] X. Q. Zhao and J. Wang, "A new noisy speech recognition method," in *Proc. IEEE International Symposium on Communications and Information Technology*, Oct. 2005, pp.292-296.

[2] B. H. Juang and T. H. Chen, "The past, present, and future of speech processing,'' *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24–48, May 1998.

[3] J. C. Wang, J. F. Wang, and Y. S. Weng, "Chip design of MFCC extraction for speech recognition," *Integration, the VLSI Journal*, vol. 32, no. 1–3, pp. 111–131, Nov. 2002.

[4] O. B. Tuzun, M. Demirekler, and K. B. Nakiboglu, "Comparison of parametric and non-parametric representations of speech for recognition," *Proc. 7th Mediterranean Electrotechnical Conference*, 1994, pp. 65–68.

[5] T. S. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp 887-906, 2005.

[6] Neural Systems Research. [Online] Available: http://neural.cs.washington.edu/

[7] J. Woojay and B. H. Juang, "Speech analysis in a model of the central auditory system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp.1802-1817, Aug. 2008.

[8] Z. Weng, L. Li, and D. Guo, "Speaker recognition using weighted dynamic MFCC based on GMM," in *proc. International Conference Anti-Counterfeiting Security and Identification in Communication*, July 2010, pp. 285-288.

[9] T. Petersen and S. Boll, "Critical band analysis-synthesis," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 31, no. 3, pp. 656-663, June 1983.

[10] A. Haddad, S. A. Samad, A. Hussain, K. A. Ishak, and H. Mirvaziri, "Decision fusion for isolated Malay digit recognition using dynamic time warping (DTW) and hidden Markov model (HMM)," in *Proc. Student Conference on Research and Development,* pp.1,6, Dec. 2007.

[11] W. C. Lin, H. T. Fan and J. W. Hung, "DCT-based processing of dynamic features for robust speech recognition," in *Proc. IEEE International Symposium on Chinese Spoken Language Processing*, Nov. 2010, pp.12-17.

[12] V. F. S. Alencar and A. Alcaim, "LSF and LPC - derived features for large vocabulary distributed continuous speech recognition in Brazilian Portuguese," in *Proc. Asilomar Conference on Signals, Systems and Computers,* Oct. 2008, pp. 1237-1241.

[13] M. A. Ferrer, I. G. Alonso, and C. M. Travieso, "Influence of initialisation and stop criteria on HMM based recognisers," *Electronics Letters*, vol.36, no.13, pp.1165-1166, Jun. 2000.

[14] D. Sarkar, "Randomness in generalization ability: a source to improve it," *IEEE Transactions on Neural Networks*, vol.7, no.3, pp.676-685, May 1996.

[15] Chunghwa Digits Database. [Online] Available: http://www.aclclp.org.tw/use_mat_c.php

[16] Q. Wu, L. Zhang, and G. Shi, "Robust multifactor speech feature extraction based on Gabor analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 927-936, May 2011.

[17] A. Rosenberg, C. H. Lee, and F. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. ICSLP*, 1994, vol. 4, pp. 1835–1838.

[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[19] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech, on Audio Processing*, vol. 12, no. 1, pp. 59–67, Jan. 2004.

[20] Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. Signal Process.*, vol. 40, no. 6, pp. 1303–1316, Jun. 1992.

[21] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 43–49, Jan. 2006.

[22] A. Qiu, C. E. Schreiner, and M. A. Escabi, "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," *J. Neurophysiol.*, vol. 90, no. 1, pp. 456–476, 2003.

[23] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proc. Interspeech'07*, 2007.