Research Journal of Applied Sciences, Engineering and Technology 12(9): 964-967, 2016
DOI:10.19026/rjaset.12.2814
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2016 Maxwell Scientific Publication Corp.
Submitted: November 30, 2015 Accepted: February 10, 2016 Published: May 05, 2016

Research Article

An Extraction Method of Acoustic Features for Speech Recognition

[1]Ibrahim Missaoui and [1, 2]Zied Lachiri
[1]Signal, Image and Pattern Recognition Laboratory, National Engineering School of Tunis (ENIT),
University of Tunis El Manar, BP. 37 Belvédère, 1002,
[2]Physics and Instrumentation Department, National Institute of Applied Science and Technology,
University of Carthage, BP 676, 1080 Tunis, Tunisia

Abstract: This study presents a novel method that deals with extracting acoustic features for recognition of isolated speech words. This extraction method is based on the use of a bank of 41 Gabor filers, which aim to select the specific modulation frequencies and bring a limitation of information redundancy on feature level. The robustness and performance of proposed features, named as Gabor Mel Spectrum features (GMS features) are validated on isolated speech words in both clean and noisy environment case and compared to those of two classic methods such as PLP-features and MFCC-features. The recognition results obtained using HMM, show that our extraction method is more robust and achieve better recognition rates than the two latter methods.

Keywords: 2-D Gabor filters, acoustic features, clean and noisy environment, speech recognition

INTRODUCTION

The extraction of acoustic features for speech recognition system has been the area of extensive research, with the aim of improving the robustness of this system in clean and noisy environment. Several developed approaches are inspired from the principles of the human auditory mechanism. These principles are integrated into these approaches in terms of various auditory model such as Mel scale filterbank (Davis and Mermelstein, 1980), Bark filtrebank (Hermansky, 1990), Gammatone filterbank (Qi *et al*., 2013) and Gammachirp filterbank (Zouhir and Ouni, 2015) to improve its robustness. However, they still far less robust to background noise compared to auditory mechanism of human.

Some recent study revealed the existence of auditory cortex neurons which are explicitly tuned and sensitive to various patterns in the signal's spectro-temporal representation. The estimation of the activity of these auditory neurons is Spectro-temporal receptive fields referred to as STRFs (Mesgarani *et al*., 2007). These findings motivated many speech processing researchers to model the STRFs and to employ it in many various applications (Mesgarani and Shamma, 2011). For example, 2 dimensional (2-D) Gabor filters is used as a model in Qiu *et al*. (2003) and these filters have been applied to treat the speech recognition problem in many studies (Kovács *et al*., 2015; Kovács and Tóth, 2015; Ravuri and Morgan, 2010; Schädler *et al*., 2012; Schädler and Kollmeier, 2015). These proposed feature extraction methods has achieved a good improvement in term of recognition rate by applying the 2-D Gabor filters to various representation of vocal signal such as log Mel-spectrogram (Schädler *et al*., 2012; Schädler and Kollmeier, 2015) obtained from MFCC (Davis and Mermelstein, 1980) and PNCC spectrogram (Meyer *et al*., 2012) generated by Power-Normalized Cepstral Coefficients (Kim and Stern, 2009) and spectro-temporal representation generated from Bark filterbank output (Missaoui and Lachiri, 2014).

In this study, a new extraction method of acoustic features for recognition of isolated speech words is presented. This method incorporates a set of 41 two-Dimensional (2-D) Gabor filters to improve recognition performance and robustness. This method is tested and evaluated on recognition of isolated speech words of TIMIT database in both clean environments and noisy environments. For this, the Hidden Markov Models are used to obtain the recognition rate of the proposed features and PLP-features (Hermansky, 1990) and MFCC-features (Davis and Mermelstein, 1980).

MATERIALS AND METHODS

An extraction method of acoustic features for recognition of isolated speech words in both clean and

Corresponding Author: Ibrahim Missaoui, Signal, Image and Pattern Recognition Laboratory, National Engineering School of Tunis (ENIT), University of Tunis El Manar, BP. 37 Belvédère, 1002 Tunis, Tunisia

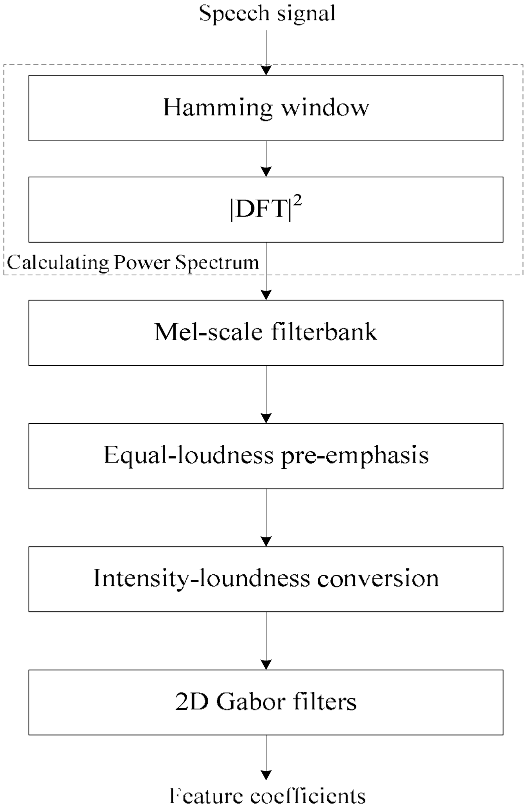

Fig. 1: The different steps of the proposed extraction method of acoustic features

noisy environment case is presented in this section. The different steps of our extraction method are shown in Fig. 1.

The power spectrum of speech is obtained by using windowing operation with 20 ms Hamming window and 10 ms overlap in the first step, followed by a square calculation of Discrete Fourier Transform in the second step.

The third step consists to apply a Mel scale filterbank to the obtained power spectrum. This filterbank consist of a set of filters evenly spaced along a warped scale resolution which is a perceptually motivated scale known as Mel frequency scale (Stevens and Volkmann, 1940). The used filters are triangular bandpass filters and the mel-scale can be approximate by the following equation:

$$mel(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (1)$$

Then, the obtained spectrum is processed by an equal loudness pre-emphasis operation and intensity loudness conversion operation. These two operations aim to reproduce and simulate firstly the non-equal sensitivity and secondly the power law of hearing (Hermansky, 1990).

The last step is to perform a processing of the obtained spectro-temporal representation by 41 2D Gabor filters in order to generate the proposed Gabor features, named as Gabor Mel Spectrum features (GMS features). The 2D Gabor filter have been widely and successively used as front-end in many speech recognition systems (Schädler *et al*., 2012; Meyer *et al*., 2012, 2011; Missaoui and Lachiri, 2014). The used 41 Gabor filters are selected to offer the ability to cover many spectro-temporal modulation directions, to approximate the orthogonal filters and to limit the redundancy between the outputs of these filters.

The 2-D Gabor filter which can be represented as 2-D convolution, is the product of complex sinusoid $\mathrm{s(n,k)}$ defined in equation 2 and Hanning envelope $\mathrm{h(n,k)}$ defined in equation 3 (Schädler *et al*., 2012). The values of the standard deviation of envelope are denoted by $\mathrm{W_n}$ and $\mathrm{W_k}$, while the radian frequencies that definite the periodicity are denoted by ω_n and ω_k.

$$s(n,k) = \exp\left(i\omega_n(n - n_0) + i\omega_k(k - k_0)\right) \quad (2)$$

$$h(n,k) = 0.5 - \cos\left(\frac{2\pi(\mathrm{n}-\mathrm{n_0})}{\mathrm{W_n}+1}\right)\cos\left(\frac{2\pi(\mathrm{k}-\mathrm{k_0})}{\mathrm{W_k}+1}\right) \quad (3)$$

RESULTS AND DISCUSSION

In this section, we present the experimental results conducted to evaluate robustness and performance of the proposed GMS features in the clean and noisy environment case.

This experiment is conducted with 12534 isolated words speech taken from TIMIT database (Garofolo *et al*., 1993). 9240 and 3294 of these extracted words are used respectively for learning phase and the recognition phase. In the noisy environment case, the noisy isolated words are generated by adding to the clean words three different noises « babble noise », « restaurant noise » and « station noise » at different SNR levels. These noises are drawn from the AURORA Corpus (Hirsch and Pearce, 2000).

The recognition results of the GMS features are compared to those of PLP-features and MFCC-features. All results are obtained using speech recognition system which employs the Markov Model Toolkit (Young *et al*., 2009) to build Hidden Markov Models (HMM).

The used HMM topology is left-to-right HMM with 5 states and 8 diagonal Gaussian mixtures per state for each isolated speech word (HMM_8_GM).

Evaluation in a clean environment: Table 1 reports the experimental results of the GMS features and those obtained using PLP and MFCC coefficients in the clean environment case. The used features in this table are: N is the total number of isolated speech words and H, S and D are the number of correct speech words, substitutions speech words and deletions speech words respectively. As shown, we can see that the GMS features provide a higher recognition rate than the two

Table 1: The recognition rates (%) of the proposed features (GMS features), PLP-features and MFCC-features in a clean environment

	Recognition rate using HMM_8_GM				
Features	%	N	H	S	D
The GMS features	97.06	3294	3144	97	0
PLP	89.62	3294	2952	342	0
MFCC	90.26	3294	2973	321	0

Table 2: The recognition results of the proposed features (GMS features), PLP-features and MFCC-features using HMM_8_GM in Babble noise case

	SNR level				
Features	0 db	5 db	10 db	15 db	20 db
The GMS features	16.36	42.17	72.62	89.56	94.81
PLP	19.06	33.36	57.56	78.05	86.61
MFCC	18.09	34.34	58.14	78.42	87.55

Table 3: The recognition results of the proposed features (GMS features), PLP-features and MFCC-features using HMM_8_GM in restaurant noise case

	SNR level				
Features	0 db	5 db	10 db	15 db	20 db
The GMS features	25.23	57.62	81.57	92.59	95.78
PLP	9.84	27.69	55.77	76.72	86.82
MFCC	11.84	30.63	57.35	78.11	87.43

Table 4: The recognition results of the proposed features (GMS features), PLP-features and MFCC-features using HMM_8_GM in station noise case

	SNR level				
Features	0 db	5 db	10 db	15 db	20 db
The GMS features	38.22	71.04	90.86	94.99	96.45
PLP	18.61	36.55	60.26	78.90	87.22
MFCC	18.03	38.71	60.50	78.90	87.95

conventional features. It achieved 97.06 of recognition rate, while the PLP and MFCC coefficients respectively had 89.62 and 90.26.

Evaluation in a noisy environment: In noisy Environment case, the results of recognition experiments with the GMS features and two classic features PLP and MFCC for « babble noise », « restaurant noise » and « station noise » are summarized in the Table 2 to 4. These three noises are taken from the AURORA database. Four SNR levels: 0, 5, 10 and 20 dB, respectively are considered.

The reported results showed that the GMS features provide the highest recognition rate compared to those of PLP and MFCC coefficients in the presence of the three noises. It demonstrates significantly better performance for all SNR levels. For example, the recognition rate obtained using the GMS features is 90.86 while those using PLP and MFCC are 60.50 and 60.50, respectively in the presence of « station noise » with SNR level equal to 10db.

CONCLUSION

An extraction method of acoustic features for speech recognition in the clean and noisy environment case was presented in this study. It is based on a set of 41 2-D Gabor filters which has been applied to a spectro-temporal representation to extract the proposed features. The performance of our features was tested on isolated speech words using HMM with 5 states and 8 diagonal Gaussian mixtures per state and was compared to those of PLP coefficients and MFCC coefficients. It was shown that our Gabor features outperform the two classic features in terms of recognition results.

REFERENCES

Davis, S.B. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE T. Acoust. Speech, 28(4): 357-366.

Garofolo, J.S., L.F. Lamel, W.M. Fisher, J.G. Fiscus and D.S. Pallett, 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Technical Report No. 93, 27403.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87(4): 1738-1752.

Hirsch, H. and D. Pearce, 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proceeding of the ISCA Tutorial and Research Workshop on Automatic Speech Recognition: Challenges for the New Millennium (ASR, 2000). Sep. 18-20, pp: 181-188.

Kim, C. and R.M. Stern, 2009. Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. Proceeding of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH, 2009), pp: 28-31.

Kovács, G. and L. Tóth, 2015. Joint optimization of spectro-temporal features and deep neural nets for robust automatic speech recognition. Acta Cybernet., 22(1): 117-134.

Kovács, G., L. Tóth and D.V. Compernolle, 2015. Selection and enhancement of Gabor filters for automatic speech recognition. Int. J. Speech Technol., 18(1): 1-16.

Mesgarani, N., S. David and S. Shamma, 2007. Representation of phonemes in primary auditory cortex: How the brain analyzes speech. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2007). Honolulu, HI, April 15-20, pp: IV-765-IV-768.

Mesgarani, N. and S. Shamma, 2011. Speech processing with a cortical representation of audio. Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2011). Prague, May 22-27, pp: 5872-5875.

Meyer, B.T., S.V. Ravuri, M.R. Schädler and N. Morgan, 2011. Comparing different flavors of spectro-temporal features for ASR. Proceeding of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH, 2011). August 27-31, pp: 1269-1272.

Meyer, B.T., C. Spille, B. Kollmeier and N. Morgan, 2012. Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition. Proceeding of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Sep. 9-13, pp: 1259-1262.

Missaoui, I. and Z. Lachiri, 2014. Gabor filterbank features for robust speech recognition. Proceeding of the International Conference on Image and Signal Processing (ICISP, 2014). Lecture Notes in Computer Science, Springer International Publishing, Switzerland, 8509: 665-671.

Qi, J., D. Wang, Y. Jiang and R. Liu, 2013. Auditory features based on Gammatone filters for robust speech recognition. Proceeding of the IEEE International Symposium on Circuits and Systems (ISCAS, 2013). Beijing, May 19-23, pp: 305-308.

Qiu, A., C.E. Schreiner and M.A. Escabí, 2003. Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition. J. Neurophysiol., 90(1): 456-476.

Ravuri, S. and N. Morgan, 2010. Using spectro-temporal features to improve AFE feature extraction for ASR. Proceeding of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH, 2010). Makuhari, Shiba, Japan, September 26-30, pp: 1181-1184.

Schädler, M., B.T. Meyer and B. Kollmeier, 2012. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. J. Acoust. Soc. Am., 131(5): 4134-4151.

Schädler, M.R. and B. Kollmeier, 2015. Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. J. Acoust. Soc. Am., 137(4): 2047-2059.

Stevens, S.S. and J. Volkmann, 1940. The relation of pitch to frequency: A revised scale. Am. J. Psychol., 53(3): 329-353.

Young, S.J., G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P.C. Woodland, 2009. The HTK Book Version 3.4.1. Department of Engineering, Cambridge University, Cambridge.

Zouhir, Y. and K. Ouni, 2015. Noise robust speech parameterization using relative spectra and auditory filterbank. Res. J. Appl. Sci. Eng. Technol., 9(9): 755-759.