Physiologically Motivated Feature Extraction for Robust Automatic Speech Recognition

Ibrahim Missaoui and Zied Lachiri
Signal, Image and Information Technology Laboratory
National Engineering School of Tunis, University of Tunis El Manar, BP. 37 Belvédère,1002 Tunis, Tunisia

***Abstract*—In this paper, a new method is presented to extract robust speech features in the presence of the external noise. The proposed method based on two-dimensional Gabor filters takes in account the spectro-temporal modulation frequencies and also limits the redundancy on the feature level. The performance of the proposed feature extraction method was evaluated on isolated speech words which are extracted from TIMIT corpus and corrupted by background noise. The evaluation results demonstrate that the proposed feature extraction method outperforms the classic methods such as Perceptual Linear Prediction, Linear Predictive Coding, Linear Prediction Cepstral coefficients and Mel Frequency Cepstral Coefficients.**

Keywords—Feature extraction; Two-dimensional Gabor filters; Noisy speech recognition

I. Introduction

Over the last years, numerous feature extraction methods have been developed for noise robust Automatic Speech Recognition (ASR) to improve performance and robustness of the recognition task. Several of these methods exploit the principles of speech processing of human speech perception to overcome the lack of robustness against the variability of speech signals. The traditional feature extraction methods such as Mel-frequency cepstral coefficients (MFCC) [1], Linear Prediction coding (LPC) [2] and Perceptual Linear Prediction (PLP) [3] were been based on the use of auditory filter modeling. Further improvements were made by using various auditory modeling in other methods [4][5][6].

Recent physiological and psychoacoustic studies have additionally shown that the primary auditory cortex neurons responsive to spectro-temporal modulations which referred as the Spectro-Temporal Receptive Fields (STRFs) have an important role in speech perception. Two-dimensional spectro-temporal Gabor filters have successfully used for modeling STRFs [7][8]. This has led to various extraction approaches of spectro-temporal features that achieve good performance in ASR noise robustness compared to traditional features [9][10][11]. In [12], Gabor features was obtained by processing a log Mel-spectrogram by a number 2D Gabor filters which were organized in a filterbank while these features were calculated from time-frequency representation derived from Power-Normalized Cepstral Coefficients (PNCCs) [15] in [16].

In this study, a physiologically motivated extraction method of Gabor features for noisy speech recognition is presented. The proposed method was based on the use of a set of 41 two-dimensional Gabor filters organized in a filter bank. It was applied to recognition of the TIMIT isolated words in the noisy environments. The recognition task is performed using Hidden Markov Models, which have been built using HTK toolkit [15].

This paper was organized as follows: Section 2 describes the proposed Gabor features extraction method. The experimental framework and results were detailed in section 3. Section 4 provides conclusions of this paper.

II. The Proposed Feature Extraction Based on Two-Dimensional Gabor Filters

A novel method based on two-dimensional Gabor filters is proposed to extract robust speech features for recognition of isolated speech words. The various steps were illustrated in Figure 2.

After pre-emphasizing the input speech signal, the power spectrum of signal is calculated by performing a windowing operation using a Hamming window (20 ms length with 10 ms overlap) and the square of Discrete Fourier Transform. It is then passed into a Bark-scale filter bank which aims to simulate the critical-band-masking curves, in order to obtain a critical-band power spectrum [3].

Subsequently, the equal loudness pre-emphasis and the intensity loudness conversion (third root amplitude compression) are performed to reproduce the two psychoacoustic properties of human hearing system; the non-equal sensitivity increase across frequency and the power law of hearing, which represents the simulation of the relation between the speech signal intensity and the perceived loudness of speech [3]. These two steps allow the reduction of spectral amplitude variation of the obtained spectrum.

Finally, the proposed features named as Gabor Bark Power Spectrum features or GBPS features were extracted by applying a set of two-dimensional Gabor filters organized in a filter bank to the representation of the obtained spectrum. This filterbank is composed of 41 two-dimensional Gabor filters [12]. These filters represent one of the most recent states of the art methods that were been successfully applied as front-end to noise robust speech recognition [12][16][18]. The Gabor features were obtained by calculating the 2D convolution of the filter and a time-frequency representation of speech to capture spectro-temporal modulations. Each two-dimensional Gabor filter is the product of two function terms: a complex sinusoid term denoted as $s(n,k)$ and a Hanning envelope $h(n,k)$ (with the time and frequency window lengths are W_n and W_k) [12][13][14].

$$s(n,k) = \exp\left(i\omega_n(n-n_0) + i\omega_k(k-k_0)\right) \quad (1)$$

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

The two terms ω_n and ω_k are time modulation frequency and the spectral modulation frequency. These terms determine the periodicity of the Gabor function and allow it to will be being tuned to a wide range of directions of spectro-temporal modulation.

The used bank of 41 Gabor filters were selected to get transfer functions of these filters having a constant overlap in the modulation frequency domain and covering a broad interval, which aimed to offer an approximated orthogonal filter and a limitation of redundancy of the filter output signal. The temporal and spectral modulation frequencies of the used bank of 41 Gabor filters were illustrated in Figure 1.

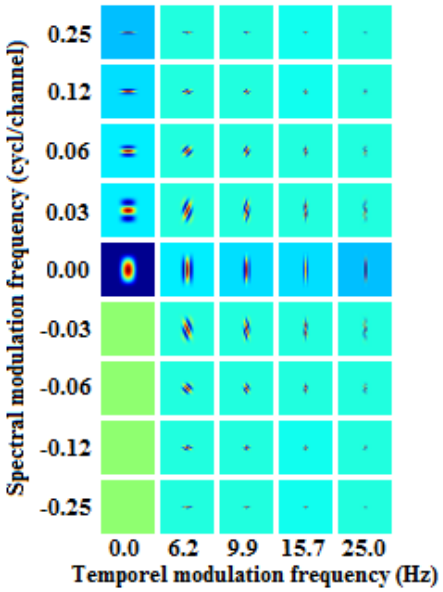

Fig. 1. The real components of a set of 41 Gabor filters employed in the proposed method

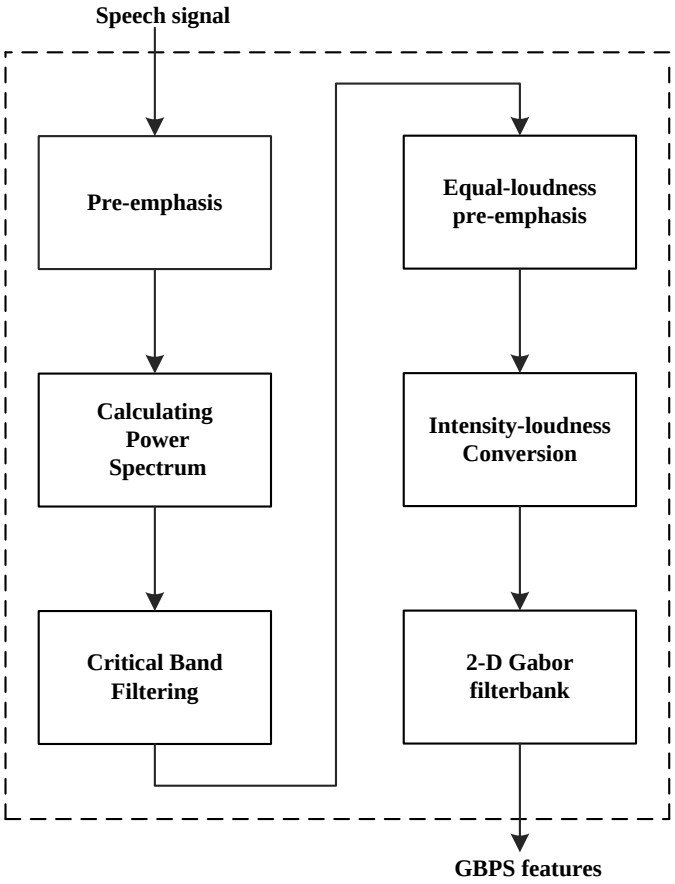

Fig. 2. Overview of the proposed feature extraction method based on two-dimensional Gabor filters

III. EXPERIMENTAL FRAMEWORK

A. The used Databases

The TIMIT database [19] was used for all ASR experiments reported in this paper. It is one of the standard databases used to evaluate the robustness and performance of any new method on an ASR task because it has a wide range of speakers and dialects. This database consists of speech signals with sampling frequency equal to 16 kHz of 630 (192 female and 438 male) different speakers from eight different major dialects of The United States, ten sentences spoken by each one of these speakers

In our experimental study, we used isolated words speech extracted from TIMIT database. A total of 9240 isolated speech words were exploited in the learning phase and 3294 isolated speech words were used for the recognition phase.

Furthermore, six background noises (restaurant, exhibition, babble, Car) drawn from the AURORA database [20] are used to evaluate the robustness of the proposed method under additive noise. The noisy isolated words used in this work were obtained by combining clean isolated words by each noise for various noise levels SNR.

B. The used Speech recognizer

The speech recognizer used in our experiments was based on HMM which have been built using the Hidden Markov Model Toolkit (HTK 3.4.1) [17]. This portable toolkit is developed by Cambridge University and used to construct and manipulate HMM optimized for speech recognition. An HMM is used to model a series of acoustic vectors. It represents a collection of stationary states which are connected by transition of Markov chain. At each state change, an observed acoustic vector o_t which described by an emitting probability distribution density $b_j(o_t)$ is generated. The transition between state s_i and state s_j is also probabilistic and has a discrete probability a_{ij} associated with it [21][22]. An example of an HMM consisting of five states with non-emitting entry and exit states is showed in Figure 3.

In the case of continuous density HMM, the most widely used output probability density $b_j(o_t)$ is the Gaussian mixture density which was defined as [17]

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} N(o_t; \mu_{jk}, \vartheta_{jk}) \quad (3)$$

Where $N(o_t; \vartheta_{jk}; o_{jk})$ is the multivariate Gaussian density with ϑ_{jk}, μ_{jk} and c_{jk} are the covariance matrix, the mean vector and weight associated with, the k^{th} Gaussian component at state j. "n" is the dimension of the vector o_t.

$$N(o_t; \vartheta_{jk}; o_{jk}) = \frac{1}{\sqrt{(2\pi)^n |\vartheta_{jk}|}} e^{(-\frac{1}{2}(o_t-\mu_{jk})^T \vartheta_{jk}^{-1}(o_t-\mu_{jk}))} \quad (4)$$

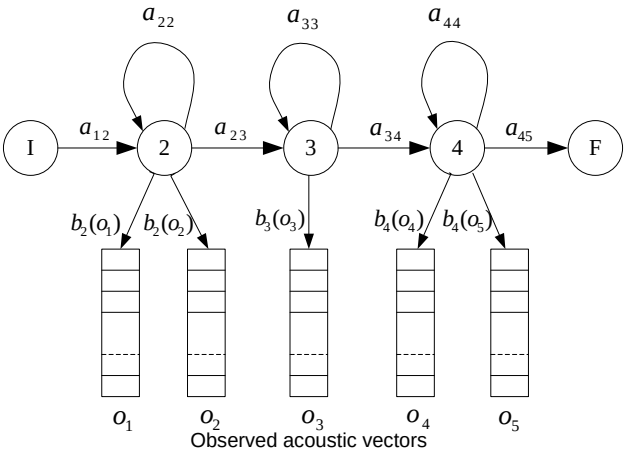

Fig. 3. Illustration of Hidden Markov models with five left-to-right states

The HMM topology exploited in our experiments is the left-to-right five-state HMM with Gaussian Mixture density and diagonal covariance matrix. Each HMM state is represented by four Gaussian Mixtures (HMM-4-GM).

C. Results and discussion

For all of our experiments, the proposed Gabor Bark Power Spectrum features or GBPS features are compared to four classic features combined with energy (E) such as Perceptual Linear Prediction (PLP_E), Linear Predictive Coding (LPC_E), Linear Prediction Cepstral coefficients (LPCC_E) and Mel Frequency Cepstral Coefficients (MFCC_E).

TABLE I. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE RESTAURANT NOISE CASE

Restaurant noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	48.15	15.24	14.63	15.42	14.63
5 dB	75.41	31.88	31.12	27.35	18.73
10 dB	91.26	60.35	60.53	48.15	27.41
15 dB	94.35	80.94	81.51	72.50	41.59
20 dB	95.60	88.77	89.13	82.91	54.07
25 dB	95.96	91.04	92.11	87.13	61.87

TABLE II. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE EXHIBITION NOISE CASE

Exhibition noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	44.44	5.65	6.34	5.98	5.04
5 dB	71.98	16.58	18.09	12.48	7.95
10 dB	88.37	38.49	18.09	30.42	17.30
15 dB	93.69	55.83	58.32	47.33	24.89
20 dB	95.23	73.95	74.13	62.96	33.24
25 dB	95.87	84.58	86.00	78.96	44.23

TABLE III. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE BABBLE NOISE CASE

Babble noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	45.87	18.94	18.94	15.66	13.36
5 dB	69.73	35.22	36.04	26.62	17.58
10 dB	87.80	60.60	60.50	50.00	24.74
15 dB	93.87	81.24	81.88	74.38	41.20
20 dB	95.29	88.92	89.53	83.58	53.64
25 dB	95.75	91.17	91.77	87.95	61.60

TABLE IV. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE CAR NOISE CASE

Car noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	49.73	11.63	13.11	14.85	8.23
5 dB	72.19	20.16	21.40	20.67	12.57
10 dB	89.74	37.37	38.40	37.28	24.32
15 dB	94.02	60.23	60.99	55.43	35.40
20 dB	95.39	80.66	82.48	73.62	42.53
25 dB	95.87	88.95	89.95	84.06	51.21

The result rates of recognition experiments with proposed Gabor features and the four classic features obtained using HMM-4-GM are summarized in the Tables I, II, III, and IV. Six noises (restaurant, exhibition, babble and car noises) drawn from the AURORA database and six specific signal-to-noise ratios (SNR) ranging from 0 dB to 25 dB in 5 dB steps were considered.

As illustrated in these tables, the proposed Gabor features outperform PLP_E, LPC_E, LPCC_E and MFCC_E features in the different cases. It can be observed that the highest percentage of the recognition rates is obtained using our Gabor features at almost all SNR levels, particularly at low SNR values. For example, in the car-noise case at SNR equal to 5 dB, the recognition rate of our Gabor features is higher than that of PLP_E, LPC_E, LPCC_E and MFCC_E features by 52.03, 59.62, 51.52 and 50.79 respectively. As can also be seen in the different tables, when decreasing the value of SNR level, the performance of all features degrade, but the proposed features remain robust and more performing than the classic features.

IV. CONCLUSION

A new physiologically motivated feature extraction method based on Gabor filterbank for isolated-word speech recognition under noisy conditions is presented in this paper. The proposed method takes into consideration the extraction of spectro-temporal modulation frequencies and the limitation of the redundancy on the feature level. The robustness of our Gabor Bark Power Spectrum features or GBPS features was evaluated on isolated speech words taken from TIMIT database using HMM. The obtained results show that our Gabor features have given the best results at all SNR levels compared to four classical features combined with energy: PLP_E, LPC_E, LPCC_E and MFCC_E features.

REFERENCES

[1] S.B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE T Acoust Speech, vol. 28, pp. 357–366, August 1980.

[2] D. O'Shaughnessy, "Linear predictive coding", IEEE Potentials, vol. 7, pp. 29–32, February 1988.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J Acoust Soc AM, vol. 87, pp. 1738–1752, April 1990.

[4] R.P. Lippmann, "Speech recognition by machines and humans", Speech Commun, vol. 22, pp.1–15, July 1997.

[5] B.T. Meyer, "Kollmeier B. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition", Speech Commun., vol.53, pp. 753–767, May 2011.

[6] Y. Zouhir, and K. Ouni, "A bio-inspired feature extraction for robust speech recognition", SpringerPlus, vol. 3, pp.651, November 2014.

[7] N. Mesgarani, and S. Shamma, "Speech processing with a cortical representation of audio", IEEE International Conference on Acoustics, Speech and Signal Processing; 22-27 May 2011; Prague, Czech Republic: IEEE. pp. 5872–5875.

[8] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: how the brain analyzes speech", IEEE International Conference on Acoustics, Speech and Signal Processing, 15-20 April 2007; Honolulu, Hawaii, USA: IEEE. pp. 765–768.

[9] M. Kleinschmidt, and D. Gelbart, "Improving word accuracy with Gabor feature extraction", International Conference on Spoken Language Processing; 16–20 September 2002; Denver, Colorado, USA: ISCA. pp. 25–28.

[10] H. Lei, B.T. Meyer, and N. Mirghafori, "Spectro-temporal Gabor features for speaker recognition", IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP; 25-30 March 2012; Kyoto, Japan: IEEE. pp. 4241–4244.

[11] S.V. Ravuri, and N. Morgan, "Using spectro-temporal features to improve AFE feature extraction for ASR", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH, 26-30 September 2010; Makuhari, Chiba, Japan: ISCA. pp. 1181–1184.

[12] M.R. Schädler, B.T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", J Acoust Soc AM, vol. 131, pp. 4134–4151, May 2012.

[13] C. Kim, and R.M. Stern, "Feature extraction for robust speech recognition using a power law nonlinearity and power-bias subtraction", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH; 6–10 September 2009; Brighton, United Kingdom: ISCA. pp. 28–31.

[14] I. Missaoui, and Z. Lachiri, An Extraction Method of Acoustic Features for Speech Recognition, Res. J. Appl. Sci. Eng. Technol, vol. 12, no. 9, 2016.

[15] I. Missaoui, and Z. Lachiri, Histogram equalization based front-end processing for noisy speech recognition, Journal of Theoretical and Applied Information Technology, 2016. in press.

[16] B.T. Meyer, and C. Spille, B. Kollmeier, and N. Morgan, "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH, 9-13 September 2012; Portland, Oregon, USA: ISCA. pp. 1259–1262.

[17] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK book (Revised for HTK version 3.4.1). Cambridge University Engineering Department, 2009.

[18] B.T. Meyer, S.V. Ravuri, M.R. Schädler, and N. Morgan, "Comparing Different Flavors of Spectro-Temporal Features for ASR", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH; 27-31 August 2011; Florence, Italy: ISCA. pp. 1269–1272.

[19] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S Pallett, "TIMIT acoustic-phonetic continous speech corpus CD-ROM", NIST speech disc 1-1., NASA STI/Recon Technical Report N 93, 27403, 1993.

[20] H. Hirsch, and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions", Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, 18–20 September 2000; Paris, France: ISCA. pp. 181–188.

[21] Y. Ephraim, and N. Merhav, "Hidden markov processes", IEEE T Inform Theory, vol. 48, pp.1518–1569, June 2002..

[22] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", P IEEE, vol. 77, pp. 257–286, February 1989.