ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HISTOGRAM EQUALIZATION BASED FRONT-END PROCESSING FOR NOISY SPEECH RECOGNITION

[1]IBRAHIM MISSAOUI, [2]ZIED LACHIRI
National Engineering School of Tunis (ENIT), Signal Image and Pattern Recognition Laboratory,
University of Tunis El Manar, BP. 37 Belvédère,1002 Tunis, Tunisia
E-mail: [1]brahim.missaoui@enit.rnu.tn, [2]zied.lachiri@enit.rnu.tn

ABSTRACT

In this paper, we present Gabor features extraction based on front-end processing using histogram equalization for noisy speech recognition. The proposed features named as Histogram Equalization of Gabor Bark Spectrum features, HeqGBS features are extracted using 2-D Gabor processing followed by a histogram equalization step from spectro-temporal representation of Bark spectrum of speech signal. The histogram equalization is used as front-end processing in order to reduce and eliminate the undesired information from the used spectrum representation. The proposed HeqGBS features are evaluated on recognition task of noisy isolated speech words using HMM. The obtained recognition rates confirm that the HeqGBS features yield interesting results compared to those of the Gabor features which are obtained from log Mel-spectrogram.

Keywords: *Front-end Processing, Histogram Equalization, 2-D Gabor processing, Feature extraction, Noisy Speech Recognition*

1. INTRODUCTION

Recognizing speech in noisy environment is the indispensable requirement for the efficiency of various speech applications. However, these applications remain less effective than human's ability to do this task, and no application that would equal this ability is developed [1][2]. Many research studies have been carried out for development of speech recognition system that would recognize speech using speech features based on some classic techniques or various auditory models. The performance of the classical techniques based features like the PLP [3], the MFCC [4] and the LPC [5] or the auditory model based features such as GFCC feature [6] AMFB feature [7] and GcFB feature [8] degrade in noisy environment and need further improvement. Other new features have integrated 2-D Gabor filters as a model of Spectro-Temporal Receptive Fields denoted STRF such as those developed in [9] and [10]. The STRF is the estimation of the activity of auditory cortex neurons [11]. The 2-D Gabor filters have applied to log Mel-spectrogram and PNCC spectrogram in [9], [10] and [12]. In [13], an auditory warped spectrum is processed by a bank of 2-D Gabor to generate a Gabor features.

The histogram equalization, which is an statistical adaptation normalization of features, has approved to be an important technique to improve the robustness of many speech recognition systems [14][15][16][17].

A new feature extraction method based on histogram equalization and 2-D Gabor processing for noisy speech recognition is presented in this work. The histogram equalization is used as front-end processing of Gabor features obtained from spectro-temporal representation of Bark spectrum of speech signal in order to improve their robustness performance. The 2-D Gabor processing consists to applying a set of 41 2-D Gabor filters to the spectro-temporal representation of speech. The experiments of isolated words recognition were carried out on TIMIT database [18] using HTK toolkit (HTK 3.5) [19]. The performance of the proposed features based on histogram equalization is compared to those of Gabor features which are obtained from log Mel-spectrogram.

Our paper is organized as follows: Section 2 will describe the proposed histogram equalization based features extraction method, while presenting the principle of histogram equalization technique. Then the section 3 will present the experimental recognition results. The section 4 will sum up the present paper.

2. HISTOGRAM EQUALIZATION BASED FEATURES EXTRACTION

2.1 Histogram Equalization

Histogram equalization (Heq) is a new normalization operation of speech features. It has been developed firstly to treat the digital image by normalizing the extracted visual features such as grey-level and contrast [20]. The Heq has been successfully employed in many features extraction approach to improve the performance and robustness of ASR applications [14][15][16][17]. It can be defined as a transformation of probability density function (pdf) of speech vectors (or testing) to a reference pdf (or training), i.e., this transformation equalizes the histogram by converting speech vectors histogram to reference histogram [21]. The Heq can be formulated and described as follows [22]: Let z is the features set with probability density function (PDF) denoted by P(z) and cumulative density function (CDF) denoted by $C_z(z)$. This set is transformed by a transformation function T into z_h with PDF $P(z_h)$ and CDF $C_{z_h}(z_h)$ which verified the following equation:

$$C_{z_h}(z_h) = C_{ref}(z_h) \qquad (1)$$

Where $C_{ref}(z_h)$ is the reference CDF.

The transformation function T is defined as:

$$z_h = T(z) = C_{z_h}^{-1}(C_z(z)) = C_{ref}^{-1}(C_z(z)) \qquad (2)$$

2.2 The proposed HeqGBS features

The proposed Histogram Equalization of Gabor Bark Spectrum features (HeqGBS features) is obtained using Histogram Equalization based proposed extraction method which can be summarized as a block diagram illustrated in figure 1. The input isolated words are firstly treated by a windowing operation, followed by a Discrete Fourier Transform of each frame. The corresponding power spectrum is obtained by calculating the square of the outputs. A filterbank which is designed to simulate the Bark-scale frequencies is applied to the power spectrum. The Bark-scale frequencies are given as [3]:

$$Bark(f) = 13 \arctan(76 * 10^{-5} f) + 3.5 \arctan\left((\frac{f}{7000})^2 \right) \qquad (3)$$

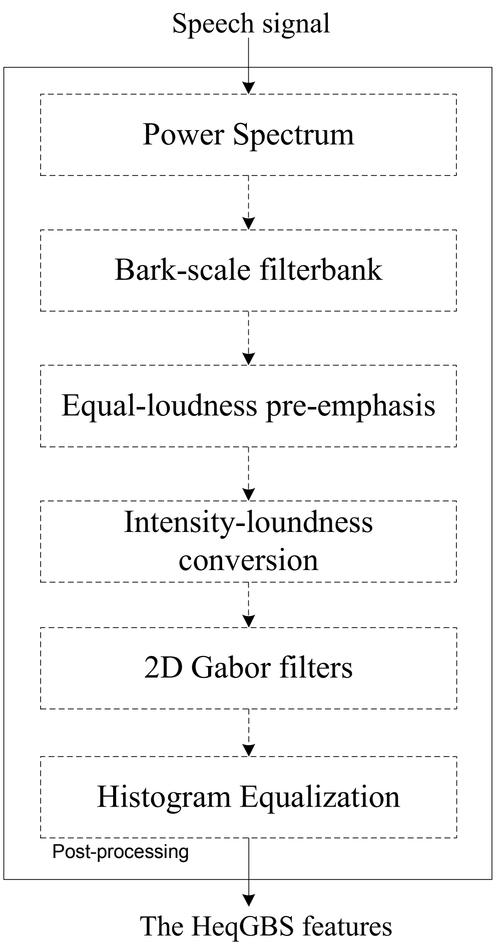

Figure 1: The Diagram Block Of The Histogram Equalization Based Proposed Extraction Method

The output Bark spectrum is pre-emphasized and then compressed by applying an equal loudness pre-emphasis function, followed by intensity loudness conversion. These two steps are employed to approximate and reproduce human sensitivity characteristic at various frequencies and the power law characteristic of hearing [3].

Then, 2-D Gabor processing is performed to the obtained representation of isolated speech word. The processing is done using 41 Gabor filters [9]. Each one is a product of $v(n,k)$ and $b(n,k)$,which are defined as [9][10][13]:

$$v(n,k) = \exp\left(i\omega_n (n - n_0) + i\omega_k (k - k_0)\right) \qquad (4)$$

$$b(n,k) = 0.5 - \cos\left(\frac{2\pi(n - n_0)}{W_n + 1}\right) \cos\left(\frac{2\pi(k - k_0)}{W_k + 1}\right) \qquad (5)$$

Where $v(n,k)$ is a complex sinusoid and $b(n,k)$ is Hanning envelope with the values of window lengths are $\mathrm{W_n}$ and $\mathrm{W_k}$.

The periodicity of $v(n,k)$ is definite by two terms ω_n and ω_k which represent the radian frequencies. These two terms allow to $v(n,k)$ to be tuned to various extent and orientation including diagonal one of spectro-temporal modulation.

Finally, the histogram equalization of the obtained Gabor features is done in order to improving the performance and robustness of these features, yielding the proposed features named as Histogram Equalization of Gabor Bark Spectrum features, HeqGBS features.

3. EXPERIMENT

The performance of the HeqGBS features obtained by the proposed extraction method is compared to the performance of Gabor features proposed by M.R. Schädler in [9]. These Gabor features are obtained by filtering the log Mel-spectrogram with 41 Gabor filters. The HeqGBS features are evaluated on a set of isolated words speech of TIMIT corpus [18]. 9240 words and 3294 words of this set are used respectively for the learning step and the recognition step. These words have been spoken by a set of speakers representing various dialects of the United States.

The noisy speech words used in our experiment are obtained by contaminated the clean one with three additive noises of AURORA corpus [23] at seven SNR levels. The sampling frequency of these words is 16 kHz. The used noises are «Car noise », «street noise » and «exhibition noise». The seven levels of SNR are -3 db, 0 db, 3 db, 6 db, 9 db, 12 db and 15 db. The temporal representations of the isolated speech word "All", the Car noise and the corresponding noisy isolated word and theirs spectrograms are illustrated in Figure 2.

The HMM recognizer is used as recognition system of speech words and have been implemented by HTK toolkit (HTK 3.5) [19]. It uses a left-to-right model of HMM with 5 states and 8 multivariate Gaussian mixtures characterized by full covariance matrix (HMM_8_GM) [24].

The recognition rates of HeqGBS features and Gabor features [9] for the three noises «Car noise », «street noise » and «exhibition noise» are reported in Tables 1, 2 and 3.

As shown in the three tables, we can observe that the HeqGBS features yield the best recognition results compared to those of Gabor features in almost values of SNR levels of the three used noises, especially for a range of SNR equals to 0 db, 3 db, 6 db and 9 db.

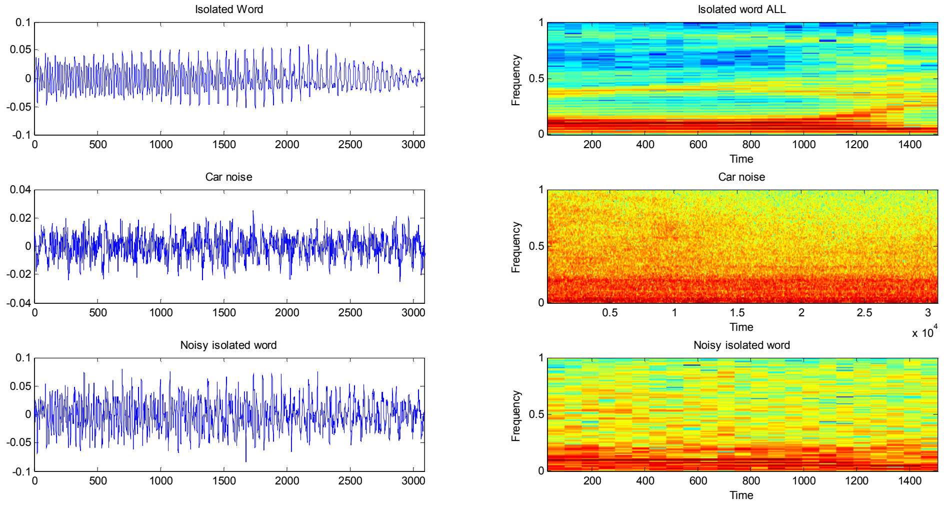

Figure 2 : The temporal representations of the isolated word "All", Car noise and the corresponding noisy isolated word and theirs spectrograms.

For example, in the case of exhibition noise at SNR value equals to 0 db, the recognition rate of HeqGBS features is higher than that of Gabor features by 20.36.

However, a slightly improvement is obtained in the three cases of noises at higher levels of SNR. For example, the HeqGBS features and Gabor features yields 96.05 and 92.82 respectively.

Table 1: Comparison of the Obtained Recognition Rates of the HeqGBS features and Gabor features in Car Noise Case

noise car	Recognition rate using HMM_8_GM	
SNR	Gabor features	HeqGBS features
-3 db	15.77	27.66
0 db	25.21	55.49
3 db	39.91	78.23
6 db	57.87	89.10
9 db	73.79	93.53
12 db	86.03	95.11
15 db	92.82	96.05

Table 2: Comparison of the Obtained Recognition Rates of the HeqGBS features and Gabor features in Exhibition Noise Case.

noise exhibition	Recognition rate using HMM_8_GM	
SNR	Gabor features	HeqGBS features
-3 db	23.53	34.03
0 db	37.75	58.11
3 db	58.57	76.81
6 db	76.65	88.43
9 db	87.12	93.11
12 db	92.09	95.14
15 db	94.89	96.08

Table 3: Comparison of the Obtained Recognition Rates of the HeqGBS features and Gabor features in Street Noise Case

noise street	Recognition rate using HMM_8_GM	
SNR	Gabor features	HeqGBS features
-3 db	11.96	28.81
0 db	25.27	59.53
3 db	48.22	79.96
6 db	67.55	89.59
9 db	80.61	93.53
12 db	89.41	95.08
15 db	94.40	95.99

4. CONCLUSION

We presented features extraction based on 2-D Gabor processing and histogram equalization for speech recognition. The histogram equalization is employed as front-end post processing for reducing the undesired information of isolated speech words. The proposed features HeqGBS is evaluated using TIMIT database and recognition system based on HMM. Comparing the recognition performance of HeqGBS features obtained using our method to those of Gabor features obtained from log Mel-spectrogram, it is shown that HeqGBS features yield the best recognition performance.

REFRENCES:

[1] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge", *Computer Speech & Language*. Vol. 27, No. 3, 2013, pp. 621-633.

[2] D. Yu, L. Deng, "Automatic Speech Recognition: A Deep Learning Approach", *Springer*, 2015.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, 1990, pp. 1738-1752.

[4] S.B. Davis, and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 28, No. 4, 1980, pp. 357-366.

[5] D. O'Shaughnessy, “Linear predictive coding”, *IEEE Potentials*, Vol. 7, No.1, 1988, pp. 29-32

[6] J. Qi, D. Wang, Y. Jiang, and R. Liu, “Auditory features based on gammatone filters for robust speech recognition”. *IEEE International Symposium on Circuits and Systems (ISCAS),* May 19-23, 2013, pp. 19-23.

[7] N. Moritz, “An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition Audio”, *IEEE/ACM Transactions on Speech, and Language Processing,* Vol. 23, No. 11, 2015, pp. 1926-1937.

[8] Y. Zouhir and K. Ouni,”A bio-inspired feature extraction for robust speech recognition”, *SpringerPlus,* Vol. 3, 2014, pp.651.

[9] M.R. Schädler, B.T. Meyer, and B. Kollmeier, “Spectro temporal modulation subspace-spanning filter bank features for robust automatic speech recognition”, *Journal of the Acoustical Society of America*, Vol. 131, No. 5, 2012, pp. 4134-4151.

[10] M.R. Schädler and B. Kollmeier, “Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition”. *Journal of the Acoustical Society of America,* Vol. 137, No. 4, 2015, pp. 2047-2059.

[11] S. Shamma, “Spectro-Temporal Receptive Fields”, *Encyclopedia of Computational Neuroscience, Springer*, 2014, pp 1-6.

[12] B.T. Meyer, 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*, September 9-13, 2012, pp. 1259-1262.

[13] I. Missaoui and Z. Lachiri, “Gabor Filterbank Features for Robust Speech Recognition”. *The International Conference on Image and Signal Processing ICISP, LNCS 8509, Springer*, 2014, pp. 665-671.

[14] V. Joshi, R. Bilgi, S. Umesh, L. García, and M. C. Benítez, “Sub-band based histogram equalization in cepstral domain for speech recognition”. *Speech Communication*, Vol. 69, 2015, pp. 46-65.

[15] H.J. Hsieh, B. Chen, J.w. Hung, “Histogram equalization of contextual statistics of speech features for robust speech recognition”, *Multimedia Tools and Applications*, Vol. 74, No. 17, 2015, pp. 6769-6795.

[16] R. Al-Wakeel, M. Shoman, M. Aboul-Ela, and S. Abdou, “Stereo-based histogram equalization for robust speech recognition”, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 15, pp.1-10.

[17] M.R. Schädler and B. Kollmeier, “Normalization of spectro-temporal Gabor filter bank features for improved robust automatic speech recognition systems”, *Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH*, September 9-13, 2012.

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

[19] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “The HTK book (HTK version 3.5)”, *Cambridge University Engineering Department*, 2015.

[20] T. Acharya, and A. K. Ray, “Image processing: principles and applications”. *Wiley-Interscience*, 2005.

[21] S. Molau, M. Pitz, and H. Ney, “Histogram-Based Normalization in the Acoustic feature space”, *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 21-24. [22] A. Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Bentez, and A.J. Rubio, “Histogram equalization of speech

representation for robust speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, 2005, pp. 355-366.

[23] 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*, September 18-20, 2000, pp. 181-188.

[24] Y. Ephraim, and N. Merhav, "Hidden markov processes", *IEEE Transactions on Information Theory,* Vol. 48, No. 6, 2002, pp. 1518-1569.