International Conference on Control, Engineering & Information Technology (CEIT'14)
Proceedings - Copyright IPCO-2014
ISSN 2356-5608

Auditory-Based Features Extraction Method for Speech Recognition

Youssef ZOUHIR, Kaïs OUNI

Research Unit: Signals and Mechatronic Systems, SMS,
Higher School of Technology and Computer Science (ESTI), University of Carthage, Tunisia
youssef.elc@gmail.com
kais.ouni@esti.rnu.tn

***Abstract*— In this paper we present a features extractor for speech recognition. The proposed features extraction method based on auditory filter modelling. The latter uses a Gammachirp Filterbank (GcFB), where their center frequencies are selected according to one of the three scales: the ERB-rate scale, the MEL scale or the BARK scale. The performance of the proposed features is evaluated, in the context of isolated words-recognition, on the TIMIT database. The recognition rate of our features extraction method with ERB-rate scale gives interesting results vs. the other two scales. The HTK platform (HMM Toolkit) recognizer is employed for the recognition system task. It's based on the Hidden Markov Models with Gaussian Mixture densities (HMM-GM).**

***Keywords*— Gammachirp auditory filterbank, Features extraction, Speech Recognition**

I. INTRODUCTION

Automatic Speech Recognition (ASR) is currently a significant research topic, since it can be serving as a man-machine interface in wide variety of applications [1]. The traditional features extractions used in the ASR applications as Mel-Frequency Cepstral Coefficients (MFCCs) [2] and Perceptual Linear Prediction (PLP) [3] are frequently based on human auditory system modeling. Better modeling of this system will improve the features robustness [4]. The auditory filterbank is generally used in the auditory model to simulate the human cochlear filtering. Specifically, a Gammatone filterbank has been exploited in various speech processing applications such as the ASR applications and the CASA systems [5].

Irino and Patterson [6] recently proposed a theoretically optimum auditory filter know as a Gammachirp filter, which represents an extension of the popular Gammatone filter with an additional analytic chirp term. This filter characterized by an asymmetric amplitude characteristic, offers an excellent fit to psychophysical masking Data [6], [7], [8].

In this paper, we propose features extraction method based on the human auditory filter and relies on the Gammachirp Filterbank (GcFB). The used filterbank is composed of 34 Gammachirp Filters covering the frequency ranges of 50-8000 Hz (sampling frequency equal to 16 kHz) [9]. The center frequencies of the used GcFB, are selected successively according to one of the three scales: the ERB-rate scale [10], the MEL scale [2] and the BARK scale [11], [3].

The Hidden Markov Model with Gaussian Mixture (HMM_GM)-based speech recognition system was performed using HMM toolkit (HTK.3.4.1) [12]. The Performance of the features extractor is assessed on the TIMT database.

The paper is presented as follows. After the introduction, we present the proposed features extraction method based on auditory filter modeling for speech recognition. Following this, in Section 3, we give the main evaluation results. Finally, conclusions are summarized in the last Section.

II. THE PROPOSED FEATURES EXTRACTION

The proposed features extraction approach is based on Gammachirp filterbank. The latter provides a spectrum reflecting the spectral properties of the human auditory system. The various steps of our features extractor (Perceptual Linear Predictive Gammachirp coefficients, PLPGc) are shown in Fig. 1. In our approach, the speech signal processing begins with pre-processing consisting in pre-emphasizing the signal using a high-pass filter characterized by a transfer function equal to (1-0.97z-1).

In the second step, the pre-processed speech signal is framed (typically 25 ms frames shifted about 10ms each time) and windowed using a Hamming-window. The power spectrum is then calculated by performing the square of DFT (Discrete Fourier Transform), for each frame of the speech signal.

In the third step, the power spectrum is decomposed into a 34-channel Gammachirp filterbank covering the frequency range of 50-8000 Hz [9], where the impulse response of the Gammachirp filter [7], [13], [14], is given by

$$g_c(t) = at^{n-1}e^{-2\pi bERB(f_c)t}e^{j2\pi f_c t + jc\ln t + j\varphi} \quad (1)$$

Here, time t>0, n and b are parameters defining the envelope of the gamma distribution, a is the amplitude, and f_c is the asymptotic frequency [15]. c is a parameter for the chirp rate, $ln(t)$ is the natural logarithm of time, φ is the phase, and $ERB(f_c)$ is the equivalent rectangular bandwidth of the auditory filter at f_c [16], [17].

The bandwidth of the Gammachirp filter is set according to its *ERB*, the value of *ERB* at frequency f in Hz [17], [5] is given by

$$ERB(f) = 24.7 + 0.108 f \quad (2)$$

The center frequencies of the Gammachirp auditory filterbank are chosen successively according to one of the three scales:

- The ERB-rate scale [10] :

$$ERBrate(f) = 21.4 + \log_{10}(0.00437 f + 1) \quad (3)$$

- The MEL scale [2] :

$$MEL(f) = 1127 \times 0.1048 \log(1 + f / 700) \quad (4)$$

- The BARK scale [11], [3].

$$BARK(f) = 13 \arctan(0.00076 f) + 3.5 \arctan((f / 7000)^2) \quad (5)$$

The figure 2 represents the mapping function from linear to logarithmic scale of the normalized scale of ERB-rate, Mel and Bark scale for a frequency range from 0 to 8000 Hertz

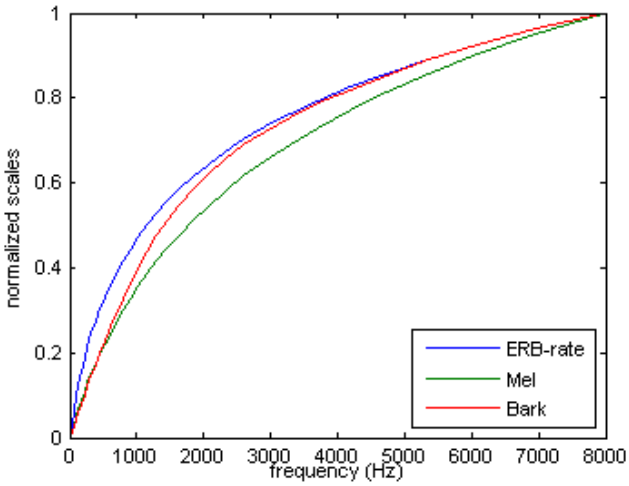

Fig. 2. Comparison between the ERB-rate, Mel and Bark scale

In the fourth and fifth steps, the output of the Gammachirp filterbank is weighted and compressed using equal-loudness pre-emphasized and the intensity-loudness conversion. The next step consists to use an autoregressive all-pole model in order to obtain an approximation of the simulated auditory spectrum [3]. The proposed features coefficients are obtained by applying a cepstral transformation in the last steps.

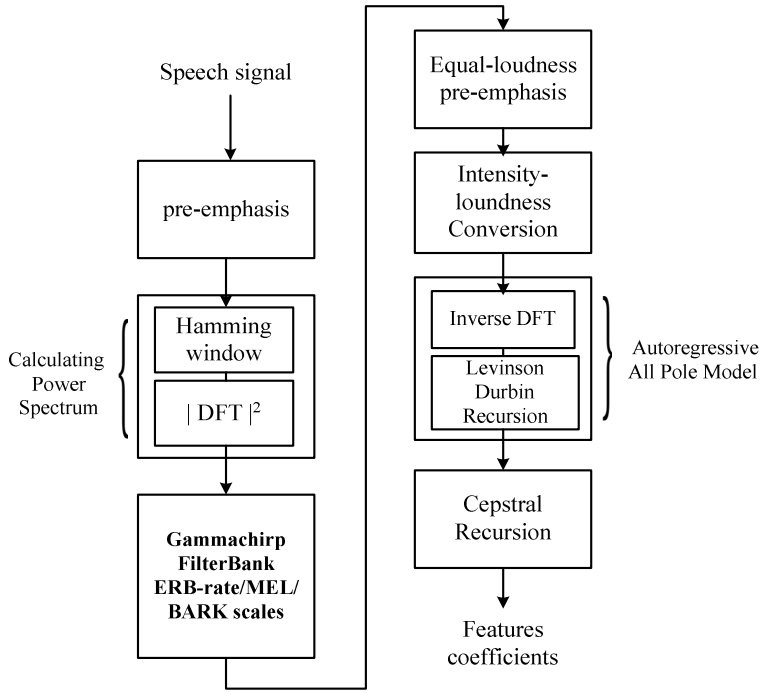

Fig. 1. Block diagram of our features extractor

III. Experimental results

To evaluate the performance of our features extraction method, we used isolated words extracted from the TIMIT database [18]. This database contains 630 speakers with sampling frequency of 16 kHz. A total of 9702 isolated-words were used for the learning phase. For the recognition phase, we used 3525 isolated-words. The Hidden Markov Model Toolkit (HTK 3.4.1) [12] recognizer is employed for the recognition task. The HMM topology is a 5-states and 4-Gaussian Mixtures per state were trained for each vocabulary isolated-words. The table 1 represents the parameters used in gammachirp filterbank.

Table 1 used parameters used of the gammachirp

Parameter	Value
n	4
b	1.019
c	2

The tables 2 and 3 represent respectively the recognition rates percentage (%) of the proposed features PLPGc (PLPGc_Brut) and the features PLPGc included the energy (E), the first (Δ) and second-order (A) temporal derivatives of the features coefficients (PLPGc+E+Δ+A) for the three frequency scales ERB-rate, MEL and BARK. In these two tables, we defines the number of correct words H, the number of deletions words D, the number of substitutions words S and the total number of words in the defining transcription files N.

As illustrated in table 2, we can observe that the recognition rate of the proposed features using the Gammachirp auditory filterbank (GcFB) with their center frequencies are chosen according to ERB-rate scale is performed better than that with the center frequencies of GcFB are selected according to the MEL and the BARK scales.

The recognition rate of our features with EBB-rate frequency scale achieved, 92.00 %, while the proposed features with MEL and BARK frequency scale had respectively, 91.69% and 91.29%.

The experimental results of the proposed features with the dynamic properties (PLPGc+E+Δ+A) are presented in table 3. We also observed that a small increase of recognition rate of proposed features (PLPGc+E+Δ+A) with ERB-rate scale that with MEL and BARK scales.

Table 2. Recognition rate (%) obtained using the proposed features (PLPGc_Brut) with center frequencies of Gammachirp filterbank (CF-GcFB) are chosen according to ERB-rate, MEL or BARK scales.

	HMM 4 GM					
	Scale (CF-GcFB)	%	N	H	S	D
Proposed features PLPGc Brut	ERB-rate	92.00	3525	3243	282	0
	MEL	91.69	3525	3232	293	0
	BARK	91.29	3525	3218	307	0

Table 3. Recognition rate (%) obtained using the proposed features (PLPGc + E + Δ + A) with center frequencies of Gammachirp filterbank (CF-GcFB) are chosen according to ERB-rate, MEL and BARK scales.

	HMM 4 GM					
	Scale (CF-GcFB)	%	N	H	S	D
Proposed features PLPGc +E + Δ + A	ERB-rate	98.16	3525	3460	65	0
	MEL	98.04	3525	3456	69	0
	BARK	97.93	3525	3452	73	0

IV. CONCLUSIONS

In this paper, we have presented a features extraction method that relies on the spectral analysis of the auditory filter. The proposed approach is based on the Gammachirp filterbank (GcFB), where the values the center frequencies of the GcFB, being chosen according to one of the three frequencies scales: the ERB-rate scale, the MEL scale or the BARK scale. The experimental results show that our features gives better recognition rates, with the center frequencies of the GcFB are based on the ERB-rate scale compared to those obtained using the MEL and the BARK scales

REFERENCES

[1] Furui, S., History and development of speech recognition. In: Chen, F., Jokinen, K. (Eds.), Speech Technology, Springer, USA, pp. 1–18(2010).

[2] Davis, S. B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, speech and Signal Processing, vol. 28, n. 4, pp. 357-366 (1980).

[3] Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer., vol. 87, n. 4, pp. 1738-1752 (1990).

[4] D. V. Compernolle, Development of a Computational Auditory Model , IPO Technical Report, Instituute voor Perceptie Onderzoek, Eindhoven, (1991).

[5] Wang, D. L., Brown, G. J.: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Publisher by IEEE Press / Wiley-Interscience (2006)

[6] Irino, T., Patterson, R. D.: A time-domain, level-dependent auditory filter: The Gammachirp. J. Acoust. Soc. Am., Vol. 101, n. 1, pp. 412-419 (1997).

[7] Irino, T., Patterson, R. D.: A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, n. 6 (2006), Author manuscript, available in PMC (2009).

[8] Unokia, M., Irino, T., Glasberg, B., Moore, B. C. J., Patterson, R. D.: Comparison of the roex and gammachirp filters as representations of the auditory filter. J. Acoust. Soc. Am., Vol. 120, n. 3, pp. 1474–1492,(2006), available in PMC(2010).

[9] Zouhir,Y., Ouni, K.: Speech Signals Parameterization Based on Auditory Filter Modeling. Advances in Nonlinear Speech Processing LNAI 7911, NOLISP 2013, Drugman, T. and Dutoit, T. (Eds), Springer. ISBN: 978-3-642-38846-0, pp. 60–66(2013)

[10] Moore, B. C., Glasberg, B. R.: Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am., Vol. 74, pp.750-753 (1983).

[11] Schroeder, M. R.: Recognition of Complex Acoustic Signals. Life Sciences Research Report 5, edited by T. H. Bullock (Abakon Verlag, Berlin), p. 324 (1977).

[12] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.4.1). Cambridge University Engineering Department (2009).

[13] Patterson, R. D., Unoki, M., Irino, T.: Extending the domain of center frequencies for the compressive gammachirp auditory filter. J. Acoust. Soc. Amer, vol. 114, n. 3, pp. 1529–1542 (2003).

[14] Irino, T., Patterson, R. D.: A compressive gammachirp auditory filter for both physiological and psychophysical data. J. Acoust. Soc. Am., vol.109, n. 5, pp. 2008-2022 (2001).

[15] Irino, T., Patterson, Unoki, M.: A time-domain, level-dependent auditory filter: An Analysis/Synthesis Auditory Filterbank Based on an IIR Gammachirp Filter. J. Acoust. Soc. Jpn.(E),. Vol. 20, n. 5, pp 397-406 (1999)

[16] Moore, B. C. J., Moore BC: An Introduction to the Psychology of Hearing. 5th ed., academic Press, London (2003).

[17] Glasberg, B. R., Moore, B. C. J.: Derivation of auditory filter shapes from notched-noise data. Hearing Research, vol. 47, pp. 103–138 (1990).

[18] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data and Speech Header Software NIST Speech Disc CD1-1.1 (1990).