Speech Signals Parameterization Based on Auditory Filter Modeling

Youssef Zouhir and Kaïs Ouni

Unité de Recherche Systèmes Mécatroniques et Signaux,
École Supérieure de Technologie et d'Informatique, Université de Carthage, Tunisie
youssef.elc@gmail.com, kais.ouni@esti.rnu.tn

Abstract. This paper presents a parameterization technique of speech signal based on auditory filter modeling by the Gammachirp auditory filterbank (GcFB), which is designed to provide a spectrum reflecting the spectral properties of the cochlea filter, which is responsible of frequency analysis in the human auditory system. The center frequencies of the GcFB are based on the ERB-rate scale, with the bandwidth of the Gammachirp filter is measured in Equivalent Rectangular Bandwidth (ERB) of human auditory filters. Our parameterization approach gives interesting results vs. other standard techniques such as LPC (Linear Prediction Coefficients), PLP (Perceptual Linear Prediction), for recognition of isolated words of speech from the TIMIT database. The recognition system is implemented on HTK platform (Hidden Toolkit) based on the Hidden Markov Models with Gaussian Mixture observation continuous densities (HMM-GM).

Keywords: Auditory filter modeling, Speech signal paramerization, Speech recognition.

1 Introduction

The statistical modeling of speech is used in most of speech recognition applications. In fact, the statistical approach provides an appropriate framework to model speech variability in both time and frequency domains. The best models commonly used nowadays are based on the Hidden Markov Models with Gaussian Mixture continuous densities (HMM-GM). In this case speech signal is classically represented as a sequence of acoustic vectors computed in synchronous way. The most efficient representations are based on spectral methods taking into account certain knowledge of speech production and perception proprieties [1].

Popular speech analysis techniques is based on simplified vocal tract models such as Linear Prediction Coefficients (LPC), whereas other techniques based on perceptual model of auditory system such as Mel-Frequency Cepstral Coefficients (MFCC) [2], and Perceptual Linear Prediction (PLP) [3]. Our technique is based on cochlear filter modeling in order to have a close parametric representation of the ear.

In fact, an acoustic signal entering the ear induced a complex spatiotemporal pattern of displacements along the length of the basilar membrane (BM) of the cochlear

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 60–66, 2013.

filter. These mechanical displacements at any given place of the BM can be viewed as the output signal of a band-pass filter whose frequency response has a resonance peak at frequency which is characteristic of the place [4]. Filters with a so-called gammatone impulse response are widely used for modeling of the cochlear filter [4.]

Recently, the auditory filter system is known to be level-dependent as evidenced by psychophysical data on masking, [5], [6]. The Gammachirp filter was proposed by Irino and Patterson is an extension of the gammatone filter with a frequency modulation term, or chirp term. Indeed, in the analytic Gammachirp, the level-dependency of the filter shape was introduced as the level-dependency of the chirp parameter. This filter provides a well-defined impulse response; it would appear to be an excellent candidate for an asymmetric, level dependent auditory filterbank. [7], [8], [5]

In this paper, we propose a parameterization technique based on the human auditory system characteristics and relying on the Gammachirp auditory Filterbank (GcFB). The filterbank has 34 filters with center frequencies equally spaced on the ERB-rate scale from 50 to 8 kHz, which gives a good approximation to the frequency selective behavior of the cochlea. The Model training and recognition were performed using speech recognition toolkit HTK.3.4.1 [9]. One Hidden Markov model (HMM) with five states and four Gaussian Mixtures per state were trained for each vocabulary word. The recognition performance of this approach was evaluated using the TIMIT database. The obtained evaluation results are compared to those of the standards techniques of parameterizations LPC and PLP.

This paper is organized as follows: It starts with, an auditory filter model in Section 2. Following this, section 3 gives the parameterization based an auditory filters modeling. The main results are presented in section 4. Finally, the major conclusions are summarized in section 5.

2 Auditory Filter Model

The objective of auditory modeling is to find a mathematical model which represents some perceptual aspects and physiological of the human auditory system [10]. In time-domain of auditory models, the spectral analysis performed by the basilar membrane is often simulated by the Gammachirp auditory filterbank [7], [6].

2.1 Gammachirp Auditory Filter.

The Gammachirp auditory filter is widely used for auditory speech analysis. Irino and Patterson have developed a theoretically optimal auditory filter [5], [11], [12], [7], in which the complex impulse response of the Gammachirp, is given as

$$g_c(t) = at^{n-1} e^{-2\pi bERB(f_0)t} e^{j2\pi f_0 t + jc\ln t + j\varphi} \tag{1}$$

Where time t>0, a is the amplitude, n and b are parameters defining the envelope of the gamma distribution, and f_0 is the asymptotic frequency [13]. ln(t) is the natural logarithm of time. c is a parameter for the frequency modulation or the chirp rate, φ is the initial phase, and ERB(f0) is the equivalent rectangular bandwidth of the auditory filter at f_0 [14], [15].

The bandwidth of the Gammachirp filter is set according to its equivalent rectangular bandwidth (ERB) of the human auditory filter. For auditory filter the ERB may be regarded as a measure of critical bandwidth [16], [14] and a good match with human data. The value of ERB at frequency f in Hz [15] is given by [16].

$$ERB(f\ ) = 24.7 + 0.108 f \tag{2}$$

The Fourier magnitude spectrum of the gammachirp filter is:

$$|G_c(f)| = \frac{a|\Gamma(n+jc)|e^{c\theta}}{(2\pi)^n\left[(bERB(f_0))^2 + (f-f_0)^2\right]^{\frac{n}{2}}} \tag{3}$$

Where $$\theta = arctg\left(\frac{f-f_0}{bERB(f_0)}\right) \tag{4}$$

And $\Gamma(n+jc)$ is the complex gamma distribution.

2.2 Gammachirp Auditory Filterbank

The used Gammachirp auditory filterbank (GcFB) is composed by 34 Gammachirp filters with center frequencies equally spaced between 50 Hz and 8 kHz on the ERB-rate scale of Glasberg and Moore [14]. This is a warped frequency scale, similar to the critical band scale of the human auditory system, on which filter center frequencies are uniformly spaced according to their ERB bandwidth. The ERB-rate scale is an approximately logarithmic function relating frequency to the number of ERBs, ERBrate(f), which is given by [16].

$$ERBrate(f) = 21.4\log_{10}(\frac{4.37f}{1000}+1) \tag{5}$$

The basilar membrane motion (BMM) produced by the GcFB in response of the waveform is presented in Fig.1 [17]. It is drawn as a set of lines, and each individual line is the output of one of the channels in the auditory filterbank [17]. As shown in Fig.1, the concentrations of activity in channels above 191 Hz show the resonances of the vocal tract which represents the 'formants' of the waveform.

Table 1. Used Gammachirp Parameters

Parameter	Value
n	4 (default)
b	1.019 (default)
c	2

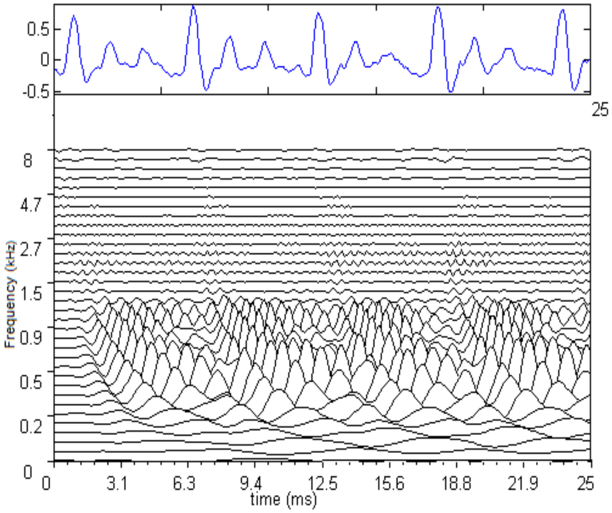

Fig. 1. The top panel shows the first 25ms -segment the waveform of the word 'ALL' extracted from TIMIT database. The bottom panel shows the Basilar membrane motion for the waveform produced by the GcFB.

3 Parameterization Based on Auditory Filter Modeling

The standard technique PLP analysis is based on an approximation of the basic psychophysical knowledge [3]. The proposed technique includes the use of a Gammachirp auditory filterbank for auditory spectral analysis.

3.1 The Standard PLP

The PLP technique is an LP-based analysis method that successfully incorporates a non-linear frequency scale and other known properties from the psychophysics of hearing. This technique uses three concepts from the psychophysics of hearing to extract an estimation of the auditory spectrum: the critical-band spectral resolution, the equal-loudness curve, and the intensity-loudness power law. The auditory spectrum is then approximated by an autoregressive all-pole model, followed by a cepstral parameterization. PLP analysis seems more consistent with human hearing, in comparison with conventional linear predictive analysis (LP) [3].

3.2 The PLPGc Technique

The proposed parameterization technique of speech signal, PLPGc (perceptual linear predictive Gammachirp) is illustrated in Fig. 2. After calculating the power spectrum of the windowed segment of speech signal, the result is passed to Gammachirp filterbank which is based on the cochlea filtering. The output is pre-emphasized by an equal loudness curve, which represents an approximation to the non-equal sensitivity of human auditory system at different frequencies. After that the Intensity loudness

Conversion step is done. This step consists in the cubic-root amplitude compression operation. It aims to simulate the non-linear relation between the intensity of speech signal and its perceived loudness. The next step of our approach is the computation of the autoregressive all-pole model which is done via the inverse DFT and the Levinson-Durbin recursion [3]. In the last step, the obtained coefficients are converted by cepstral transformation in order to obtain the PLPGc cepstral coefficients.

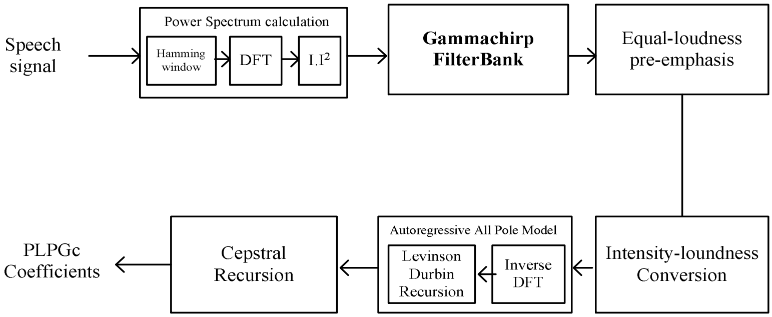
Speech signal
Power Spectrum calculation
Hamming window
DFT
I.I²
Gammachirp FilterBank
Equal-loudness pre-emphasis
Intensity-loundness Conversion
Autoregressive All Pole Model
Inverse DFT
Levinson Durbin Recursion
Cepstral Recursion
PLPGc Coefficients

Fig. 2. Block diagram of the proposed technique PLPGc

4 Experimental Results

The proposed technique has been evaluated on the TIMIT database [18], composed of 9702 isolated words for the learning phase. For the recognition phase, we used 3525 isolated words. All words were extracted from TIMIT database. The signals of this database are sampled at 16 kHz.

The Hidden Markov Model Toolkit (HTK) [9], is a portable toolkit for building and manipulating Hidden Markov Models with continuous Mixture Gaussian densities (HMM_GM). HTK is primarily used for speech recognition. The HMM topology is a 1st -order 5-states HMM model. The observation probability distribution is a 4 Gaussian mixture density with diagonal covariance matrix. 12 static coefficients vectors were computed using 25 ms hamming window, shifted with 10 ms steps. The Gammachirp filter was applied using the parameters given in the table 1.

The tables 2, 3 and 4 represent the recognition rates of different techniques: PLPGc (proposed technique), LPC and PLP [3]. Every time we add one of the following parameter to the first 12 coefficients: Energy (E), the first differential coefficients (Δ) and second order differential coefficients (A).

For HMM, we used 4 Gaussian Mixture, with 5 states of observation.

We define the parameters below: HMM 4 GM: Hidden Markov Models with 4-Gaussian-Mixtures. H is the number of correct words, D is the number of deletions words, S is the number of substitutions words and N is the total number of words in the defining transcription files. The percentage number (%) is the recognition rate of words.

As reported in the table 2 we can observe that an improvement of 2.52% relative increase of recognition rate is achieved with the PLPGc proposed technique of parameterization over the baseline PLP method. In tests the energy was also added to the feature vector. It can be seen in table 3 that the recognition rate improves slightly, with the PLPGc technique compared with PLP method.

The dynamic properties (E+Δ+A) were computed so that the final parameterization vector for techniques consisted of 39 coefficients (12 coefficients of the technique +E+Δ+A). The table 4 shows a small increase of recognition rate for the PLPGc proposed technique of parameterization compared to this of the standard technique PLP. We also observed that the standard LPC technique generally decreases the recognition scores compared to PLP and PLPGc, as shown in tables 2, 3 and 4.

Table 2. Recognition rate obtained by parameterization techniques in their brut state

Technique	HMM 4 GM				
_brut	%	N	H	S	D
PLPGc	92.00	3525	3243	282	0
PLP	89.48	3525	3154	371	0
LPC	58.55	3525	2064	1461	0

Table 3. Recognition rate obtained by parameterization techniques combined with energy (_E)

Technique	HMM 4 GM				
_E	%	N	H	S	D
PLPGc	93.67	3525	3302	223	0
PLP	93.33	3525	3290	235	0
LPC	70.07	3525	2470	1055	0

Table 4. Recognition rate obtained by parameterization techniques combined with energy, differential coefficients first and second order (_E_Δ_A)

Technique	HMM 4 GM				
_E_ Δ_A	%	N	H	S	D
PLPGc	98.16	3525	3460	65	0
PLP	97.93	3525	3452	73	0
LPC	78.24	3525	2758	767	0

5 Conclusion

In this paper, we have proposed paramerization technique PLPGc of speech signals based on the auditory filter modeling which uses the Gammachirp auditory Filterbank. Experimental results using the TIMIT database have shown that the PLPGc technique increases the recognition rate relatively according to conventional techniques such as PLP and LPC.

References

1. Frikha, M., Hamida, A.B.: A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition. American Journal of Intelligent Systems 2(1), 1–8 (2012)
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 28(4), 357–366 (1980)
3. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87(4), 1738–1752 (1990)
4. Ouni, K., Ellouze, N.: A Time-Frequency Analysis of Speech Based on Psychoacoustic Characteristics. In: Proceedings of the 17th International Congresses on Acoustics, ICA-ROME (2001)
5. Irino, T., Patterson, R.D.: A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE Transactions on Audio, Speech, and Language Processing 14(6) (2006); author manuscript, available in PMC (2009)
6. 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. 120(3), 1474–1492 (2006); available in PMC (2010)
7. Irino, T., Patterson, R.D.: A time-domain, level-dependent auditory filter: The Gammachirp. J. Acoust. Soc. Am. 101(1), 412–419 (1997)
8. Park, A.: Using Gammachirp filter for auditory analysis of speech. 18.327, Wavelets and Filter banks (2003)
9. 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)
10. Zoghlami, N., Lachiri, Z., Ellouze, N.: Speech Enhancement using Auditory Spectral Attenuation. In: Proceedings of the 17th European Signal Processing Conference, EUSIPCO, Glasgow, Scotland (2009)
11. Patterson, R.D., Unoki, M., Irino, T.: Extending the domain of center frequencies for the compressive gammachirp auditory filter. J. Acoust. Soc. Amer. 114(5), 1529–1542 (2003)
12. Irino, T., Patterson, R.D.: A compressive gammachirp auditory filter for both physiological and psychophysical data. J. Acoust. Soc. Am. 109(5), 2008–2022 (2001)
13. Irino, T., Patterson, U.M.: A time-domain, level-dependent auditory filter: An Analysis/Synthesis Auditory Filterbank Based on an IIR Gammachirp Filter. J. Acoust. Soc. Jpn (E) 20(5), 397–406 (1999)
14. Moore, B.C.J.: An Introduction to the Psychology of Hearing, 5th edn. Academic Press, London (2003)
15. Glasberg, B.R., Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hearing Research 47, 103–138 (1990)
16. Wang, D.L., Brown, G.J.: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. IEEE Press / Wiley-Interscience (2006)
17. http://www.acousticscale.org/wiki/index.php/AIM2006_Documentation
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)