Speech Compression Based on Frequency Warped Cepstrum and Wavelet Analysis

Francisco J. Ayala and Abel Herrera

Digital speech processing laboratory, Universidad Nacional Autónoma de México,
Facultad de Ingeniería
Ciudad Universitaria, 04510, Mexico City, Mexico
poasf01@cancun.fi-a.unam.mx,
abelhc@hotmail.com

Abstract. In this article it is described the process to extract a set of cepstral coefficients from a warped frequency space (mel and bark) and analyze the perceived differences in the reconstructed signal. We will try to determine if there is any audible improvement between these two most used scales for the purpose of speech analysis by synthesis. We will use the same procedure for parameter extraction and signal reconstruction for both functions, replacing only the warping scale. The proposed system is based on a basic cepstral analysis synthesis model on the mel scale, whose excitation signal generation process has been changed. The inverse MLSA filter was obtained in order to generate the analysis signal, then this signal is fed into a wavelet decomposition block and the resultant coefficients are sent to the decoding system where the excitation signal is reconstructed. Furthermore the mel scale is replaced by bark scale.

Keywords: speech compression, speech encoding, wavelet analysis, warped cepstrum.

1 Introduction

The mel scale was proposed at 1937, following a series of experiments used to establish a perceptual scale based on the perception of tones, the use of mel scale is almost standard for speech recognition application. The Bark scale, proposed by Eberhard Zwicker in 1961, divides the audible spectrum into 24 critical bands that try to mimic the frequency response of the human ear.

Given the characteristics of the human auditory system (nonlinear and time variant), the models needed to represent auditory perception are complex since they involve the use of non uniform frequency scales instead of linear scales such as the Hertz scale. The mel and bark scales are examples of non uniform scales, the use of them is desirable in low rate speech coding systems [1].

Cepstral analysis is performed for coefficients extraction, and then a frequency warping process is applied in order to change the frequency scale of the coefficients. The mel log spectrum approximation filter (mlsa) computes the synthetic

J.-F. Martínez-Trinidad et al. (Eds.): MCPR 2011, LNCS 6718, pp. 297–304, 2011.

signal [2]. This filter not only works on the mel scale but also on bark scale by just replacing the allpass parameter.

The inverse mlsa filter provides the analysis signal to which other coefficients are extracted to generate the excitation signal of the mlsa filter. This process involves the wavelet analysis. The wavelet coefficients perform well in the excitation signal generation process. Thus, as important is the quality of the cepstral coefficients used to model the mlsa filter, as the performance of the excitation signal.

2 Cepstral Analysis Synthesis

The filter coefficients are obtained through a linear transform from the warped cepstrum defined as the fourier cosine coefficients of the warped log spectrum of speech [1].

The nearest part to the origin of the cepstrum corresponds to the transfer function of the vocal tract and can be used to approximate the spectral envelope of the signal [3]. Hence the linear transform consist of a homomorphic filtering process in order to extract the first M cepstrum elements.

The mel and bark scales can be approximated by the phase characteristics of a first order allpass filter. The transfer function is given by [4]:

$$\tilde{z} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |z| < 1. \tag{1}$$

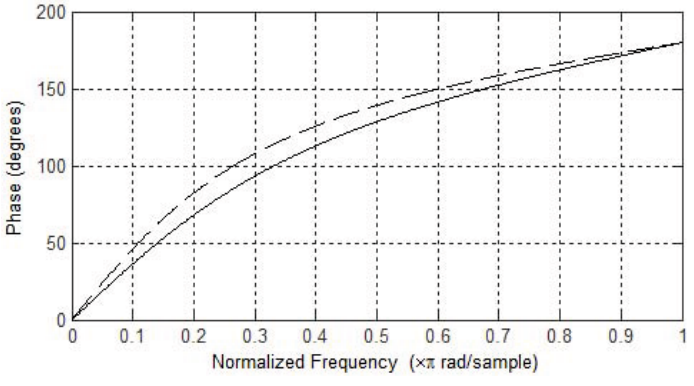

Fig. 1. Phase response of the all pass filter when α=0.4582 (dashed line) and α=0.35 (solid line)

The cepstrum is fed into an allpass filters chain so that the frequency scale becomes non uniform [1]. For certain values of α, the frequency transformation either resembles the mel scale or the bark scale. It is stated the value α=0.35 to approximate the mel scale and α=0.4582 for bark scale at 10kHz of sampling frequency [5].

The mlsa filter defined by

$$H(z) = exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m} \tag{2}$$

takes the cepstral coefficients and an excitation signal to generate the synthetic speech. Since the coefficients are transformed into a different frequency scale, the mlsa filter is designed over the warped frequency scale, thus each delay element is replaced by the first-order allpass filter. This substitution implements the unwarping while the filter is working [4].

From the inverse mlsa filter

$$\frac{1}{H(Z)} = exp \sum_{m=0}^{M} -\tilde{c}(m)\tilde{z}^{-m} \tag{3}$$

it is obtained the analysis signal which is sent to a wavelet analysis process. The purpose of this step is to generate the excitation signal.

3 Wavelet Analysis

The basis of this process is a set of quadrature mirror filters and the discrete wavelet transform [6]. In wavelet analysis, the signal is decomposed into approximations and details. The approximations are the low-frequency components of the signal. The details are the high-frequency components.

The decomposition process is achieved by iterations; one signal is broken down into many lower resolution (by dyadic decimation) components in each iteration [7].

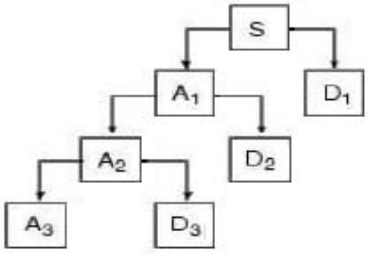

Fig. 2. Decomposition tree. For reconstruction: S=A3+D1+D2+D3.

Given a signal of length N, the discrete wavelet transform DWT consists of four levels at most (for a high quality reconstruction in this approach). The first step produces two sets of coefficients: approximation coefficients A1, and detail coefficients D1. These vectors are obtained by convolving the signal with the low-pass filter for approximation, and with the high-pass filter for detail, followed by decimation.

The second step divides the approximation coefficients A1 in two parts repeating the same procedure, replacing the input signal by A1 and producing A2 and D2, and so on. After these steps a wavelet decomposition four-level tree is obtained.

For the reconstruction process the inverse discrete wavelet transform IDWT is applied to the approximation and detail coefficients. The coefficients vectors are upsampled and filtered. The vectors are zero-padded for the upsampling step and for recovering the original size of the signal at the end of iterations. For filter design the wavelet standardized db2 coefficients are taken.

The filter coefficients are

(a) Low pass decomposition filter: h(n)=-0.1294,0.2241,0.8365,0.4830
(b) High pass decomposition filter: h(n)=-0.4830,0.8365,-0.2241,-0.1294
(c) Low pass reconstruction filter: h(n)=0.4830,0.8365,0.2241,-0.1294
(d) High pass reconstruction filter: h(n)=-0.1294,-0.2241,0.8365, -0.4830

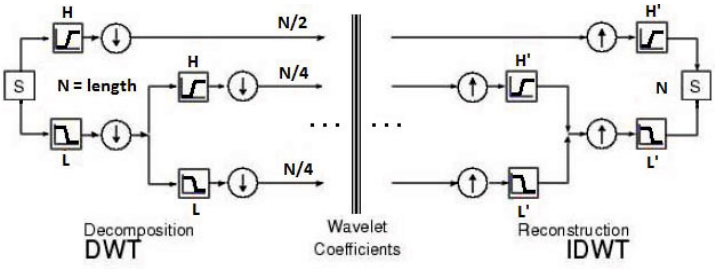

Fig. 3. (QMF) Quadrature mirror filters. Multiple level analysis-synthesis process.

In chapter 5 it is explained how this analysis is applied to the analysis signal.

4 Detection of Fricative Sound Block

Given the loss of details after compression, an improvement of voice naturalness and of intelligibility of fricative consonants can be achieved by adding noise of different band widths and amplitudes to the excitation signal.

For naturalness improvement, a low and constant amplitude noise is added to the entire excitation signal, and for intelligibility improvement of fricative sounds, a higher amplitude noise is only added to the corresponding frame. One bit is sent to indicate the presence of a high amount of zero crossings and a relative low energy of the current frame of the original speech.

In the decoder system, a white noise is divided in four bands of frequency. The band edges are given in Hertz as [500 1000 2500 3500 5000]. In original signals, the average of amplitudes corresponding to the lost information in synthetic speeches was estimated in the four bands in which the original signals were divided. Those amplitudes are the gain of the normalized noise to be added to

the excitation signal. Although the speech characteristics vary frame to frame, the results are actually good.

When the decoder receives a bit indicating the presence of fricative sounds, a white noise is added to the current frame. The noise amplitude is the average of amplitudes in original speeches that contain kind of unvoiced sounds. This block would not be necessary if a high level compression were not desired since the more compressed is the signal the higher the loss of details.

5 Design of the Coding-Decoding System

5.1 Encoding Process

A block diagram of the warped cepstral analysis synthesis system is shown in Fig. 4. First, the M cepstral coefficients are extracted as explained in chapter 2 to each 256-sample sequence. Also each sequence is fed into the inverse mlsa filter to obtain at its output the analysis signal. Then, the cepstral parameters are quantized and transmitted and the analysis signal is decomposed in wavelet coefficients.

After calculating the wavelet transform of the analysis signal it is found that most of the coefficients have small magnitudes, they are close to zero. Consequently, compression involves truncating coefficients below a threshold.

The low-frequency components are the most important part of human voice. When high-frequency components are removed the speech is still intelligible but sounds a little different. For that reason only the detail coefficients are truncated.

Around 90 % of the wavelet coefficients are found to be small and their truncation to zero make a barely perceptible difference to the signal.

In this work, for higher compression all the detail coefficients can be truncated leaving the approximation coefficients of the last level of the decomposition tree. If higher quality of synthetic speech is desired, less compression is needed and some of the details coefficients most be left.

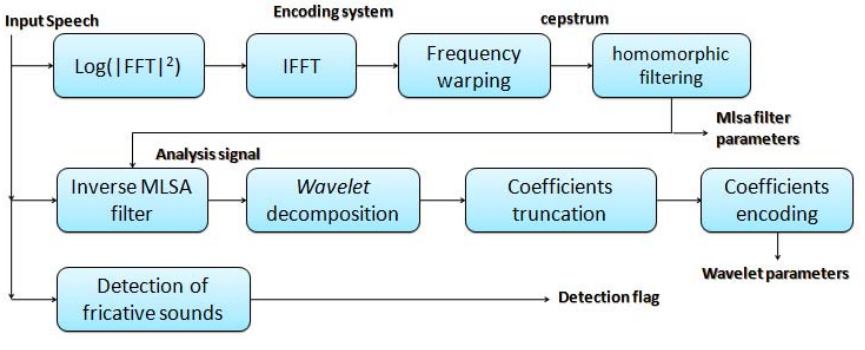

Fig. 4. Encoding system

The non-zero coefficients are stored in one vector; in a second vector is stored the starting position of a string of zeros and the number of zeros. Then the coefficients of the vectors are quantized and transmitted.

5.2 Decoding Process

The decoding system takes the cepstral and wavelet coefficients. Interpolation of the cepstral parameters of two successive frames is performed in order to smooth the transition of synthetic frames.

The interpolated cepstral coefficients are set as the mlsa filter parameters, which are unwarped in the filter using the negative magnitude of α [4].

The excitation signal is reconstructed from the wavelet coefficients by the wavelet reconstruction process where the vectors are zero padded and convolved with the reconstruction filters in each stage [7].

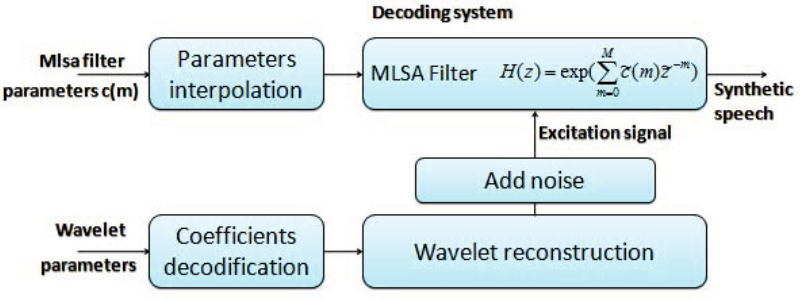

Fig. 5. Decoding system

6 Quantization of the Filter Parameters

MLSA filter parameters are quantized according to their characteristics. It is shown from experimental results that the maximal magnitude of the first coefficient is 8 (for m=1) and the maximal magnitude of the rest of the coefficients is 1. The first parameter is truncated to a 3-bit integer. The second set of parameters magnitudes are normalized according to the interval [0 1]. This allows the use of only one code book in the decoder. One additional bit is used to represent the sign. The maximal absolute value, of each set of coefficients, is transmitted to restore its original magnitude. No more than 8 bits should be taken for this value.

It is also shown from experimental results that the difference between the maximal and minimal values of each wavelet parameter does not exceed 1. The same normalization procedure is applied to these coefficients which are mapped into the interval [0 1]. The same code book is used. The maximal magnitude of each set of parameters is not transmitted. Experiments show that there is no distortion whether it is recovered or not the original magnitude of wavelet parameters.

Table 1. Bit allocation for this coder

Parameter	Resolution (bits/parameter)
Cepstral (sign included)	4
First cepstral parameter	3
Maximal value (cepstral)	8
Wavelet (sign included)	5
Bit of unvoiced frame	1

The selected number of bits for both the wavelet and cepstral coefficients depends on the level of desired compression and quality of synthetic speech. In Table 1 it is shown the proposed bit allocation.

7 Speech Quality and Bit Rate

In order to evaluate the quality of the synthetic speech, short (two seconds) and large (ten seconds) English sentences were recorded to be analyzed and synthesized.[1]

Given the parameters:

(a) Fs: sampling frequency. T: frame duration.
(b) M: cepstrum order.
(c) Bc: bits/cepstrum coefficient.
(d) W: number of wavelet coefficients.
(e) Bw: bits/wavelet coefficient.
(f) BM: bits/Maximal value.
(g) BF: bits/First cepstral parameter.
(h) α: warping parameter.

The overall bit rate B of this coder is calculated by

$$B = \frac{[(M-1)\cdot b_c + W \cdot b_w + b_M + b_F]}{T}. \tag{4}$$

For Fs=10kHz, T = 25ms, M=26, b_c=4, W=34, b_w=5, b_M=7, b_F=3 and α=0.35 (mel scale) the data rate is B=10.9 kbit/s. The speech quality is quite good and the intelligibility is very high. The speaker is clearly recognizable and the signal gets naturalness.

For Fs=10kHz, T = 25ms, M=26, b_c=4, W=18, b_w=5, b_M=7, b_F=3 and α=0.35 (mel scale) the data rate is B=7.8 kbit/s. The speech quality is good and the intelligibility is high. The speaker is clearly recognizable but the signal loses naturalness.

A MOS test was applied to 10 people. Each person heard (using headphones) and scored the seven recordings separately. The MOS score range is from 1 to 5, 1 being the worst and 5 the best. The average range of each evaluation is shown in Table 2.

[1] Recordings were performed in a quiet laboratory, using a general purpose unidirectional microphone and the sound card of a computer.

Table 2. MOS Test results

Signal	7.8 kbit/s (mel)	10.9 kbit/s (mel)	7.8 kbit/s (bark)	10.9 kbit/s (bark)
1 (male)	3.60	4.4	3.50	4.5
2 (female)	3.9	4.2	3.90	4.2
3 (male)	3.30	4.6	3.44	4.6
4 (male)	3.30	3.9	3.34	3.8
5 (female)	3.37	4	3.35	4
6 (female)	2.90	3.8	2.94	3.9
7 (female)	2.90	3.9	3.50	3.9
Average	3.32	4.11	3.42	4.12

8 Conclusion

The differences between the mel and bark scale is almost imperceptible. As the bit rate increases the diferences are completely imperceptible. For a too low bit rate there exist some audible differences between the scales but those are not statistically significant.

The quality of the synthetic speech is actually good; the performance of the coder is fair since the presence of noise does not affect it. The decoder is able to reconstruct any kind of noise but not with the same quality given to speech signals.

The spectral distortion of the parameters given the quantization process is to low. When comparing a synthetic speech generated with quantized parameters and another synthetic speech generated with non-quantized parameters there is not perceptible differences.

The proposed system performs quite good, this system is able to synthesize music, street noise (cars, airplanes, people, etc). No matter what is behind someone's voice, because it is always understandable after being compressed under this coding system.

References

1. Harma, A., Karjalainen, M.: Frequency-Warped Signal Processing for audio Applications. In: 108th AES convention Paris, France (2000)
2. Imai, S.: Cepstral analysis synthesis on the mel frequency scale. In: IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 1983, vol. 8, pp. 93–96 (1983)
3. Acero, A., Hon, H.-W.: Spoken language processing: a guide to theory. Algorithm and system development (2001)
4. Smith, J.O., Abel, J.S.: Bark and ERB bilinear transforms. In: IEEE Speech and Audio Processing, vol. 7, pp. 697–708 (1999)
5. Tokuda, K., Kobayashi, T.: Recursive Calculation of Mel-Cepstrum from LP Coefficients (April 1994)
6. Mallat, S.: A wavelet tour of signal processing, pp. 255–263 (1999)
7. Shivraman, G., Nilesh, N.: Speech compression using wavelets. Department of electrical engineering, Veermata Jijabai Technological Institute, University of Mumbai, pp. 29–54 (2002)