Distortion of Voicing and Vocal Tract Parameters After Codecs

Amr Nabil and M. Hesham
Engineering Mathematics and Physics Dept.
Faculty of Engineering, Cairo University
Giza,12613
Egypt
{anabil,mhesham}@eng.cu.edu.eg

Abstract: In this work, we present results on the effect of well-known mixed excitation linear prediction (MELP) and code-excited linear prediction (CELP) codecs (coder/decoder) on voicing and vocal tract parameters of Arabic sounds. The study shows that the spectral distortion is large compared to other studies and is largest for MELP1200. Vowel formants have a shift which may exceed one critical band below or above its reference value. Finally, it was found that the coded pitch period did not suffer any significant change through the coding/decoding process.

Key–Words: Spectral analysis, Arabic speech, Speech coders, Formants, Pitch period

1 Introduction

Speech coding has an important role in many applications. Bit rate, complexity, quality and delay are the main aspects to judge a speech codec (coder/decoder). Depending on the target application, one or more of these aspects has more weight than others to meet certain requirements. The quality of speech after coding/decoding process is relatively smaller than original [1]. MELP (Mixed Exited Linear Predictive) coders focus on bit rate while CELP (Code-Exited Linear Predictive) coders focus more on quality which causes some increase in the bit rate compared to low-bit rate coders such as MELP [2]. In this paper, we compare the spectra of speech signals. These comparisons include spectral energy and frequency, amplitude and bandwidth of vowel formants. Effect on pitch period is also studied. There is a strong correlation between preservation of these parameters and speech quality measures. Arabic like other languages contains a finite number of phonemes. These phonemes can be classified into vowels, semi-vowels and consonants.

According to large database statistics, vowels class represents about 60-70% of Arabic speech whereas the remaining percentage is distributed among other classes. Due to this high probability of occurrence of vowels in Arabic, special attention is given to the distortion in their basic formants.

The following sections are organized as follows: Section 2 explains the speech coders used in this work briefly. Section 3 lists the most important speech parameters usually used in quality measurements. Section 4 presents analysis results of speech coders on Arabic speech. Discussions and conclusions are given in section 5.

2 Speech Coders

Many coders were proposed to improve the original LPC coder [3]. In this section, the used speech coders in our work are described briefly. These coders are Melp2400, Melp1200 and CELP G723.1.

2.1 MELP2400

MELP at 2.4 kbps (Melp2400) is one of low bit rate coders. The key feature of the MELP is the mixed excitation of periodic pulses and random noise to match the spectrum of natural speech. This mixture extends the number of frame classes into voiced, unvoiced and jittery voiced. Another key feature is shape extraction of periodic excitation [2].

2.2 MELP1200

Based on the same concepts of Melp2400, Melp1200 was introduced but with half the bit rate of the original MELP. Encoding every 3 frames in one super frame is one of the main enhancements in the Melp1200. This approach allows the Melp1200 coder to have the same performance of Melp2400 at the price of additional processing delay [3].

2.3 CELP G723.1

CELP is the most widely used speech coder. By eliminating the strict classification of voiced/unvoiced frames in LPC coder, CELP uses codebook searching to select the most suitable excitation sequence to represent the speech frame.

Also, CELP preserves some phase information which was totally ignored in LPC coder. The complexity of the CELP coder comes from the search in the codebook. CELP is considered a higher bit rate coder when compared to MELP. G723.1 has a version which uses Algebraic Coded Excited Linear Prediction (ACELP) with a rate of 5.3 kbps. ACELP uses binary error correcting codes to represent N points on an M dimensional hyper sphere [3].

3 Speech Analysis Parameters

Through this section, an overview is given on the most important speech parameters and analysis measures used in this work.

3.1 Energy Distortion

In order to objectively measure the distortion between a coded and uncoded LPC parameter vector, the spectral distortion is often used in narrowband speech coding. For each frame, the spectral distortion (in dB), SD, is defined in [4] as:

$$SD = \sqrt{\frac{1}{F_s}\int_0^{F_s}[10\log_{10}(P_n(f)) - 10\log_{10}(\widehat{P}_n(f))]^2 df} \tag{1}$$

where, F_s is the sampling frequency, and $P_n(f)$ and $\widehat{P}_n(f)$ are the power spectra corresponding to the n^{th} original and processed frames.

3.2 Vowel Formants

The most important features of vowel phonemes are the formants. In most cases, the first 3-4 formants characterize the phoneme [5]. As derived in [6], the formant frequencies can be obtained from the roots of the LPC predictor polynomial as follows:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = \prod_{k=1}^{p}(1 - z_k z^{-1}) \tag{2}$$

where p is the order of the predictor. The roots $\{z_i, i = 1, 2, \ldots, p\}$ are converted from z-plane to s-plane using:

$$z_i = e^{s_i T} \tag{3}$$

where $s_i = \sigma_i + j\Omega_i$ the corresponding s-plane root and if we assumed $z_i = z_{ir} + jz_{ii}$ then,

$$\Omega_i = \frac{1}{T}\tan^{-1}\left(\frac{z_{ii}}{z_{ir}}\right) \tag{4}$$

By dividing Ω_i over 2π, we can get the formant frequencies in Hz. The spectrum of a phoneme can be extracted using equation (2) and accordingly the formants bandwidth.

3.3 Pitch Period Estimation

Pitch frequency is directly related to the speaker and sets the unique characteristic of a person. Voicing is generated when the airflow from the lungs is periodically interrupted by movements of the vocal cords. The time between successive vocal cord openings is called the fundamental period, or pitch period. The most famous technique in pitch estimation is the autocorrelation method [3]. In this method, the autocorrelation function is calculated between a frame and a time-shifted version of it using:

$$R[l, m] = \sum_{n=m-N+1}^{m} s[n]s[n-l] \tag{5}$$

where, N is the frame length, m is the end point of the frame and l is a positive integer representing a time lag. The range of lag is selected so that it covers a wide range of pitch period values [2].

4 Results on Arabic Speech

The results are obtained for Arabic speech database which was locally recorded. To represent each phoneme in an accurate manner, we need a sufficient number of frames through which the phoneme is produced. The source of all phonemes used is taken from Al Qur'an (the Holly book of Muslims) to ensure the correctness and standardization of all phonemes pronunciation. The database contains more than 10,000 phoneme records. These phonemes are sampled at a sampling rate of 8 kbps and each sample is represented in 16 bits. Matlab [7] was the mathematical tool which was used to generate our statistics. In the following subsections, we present our results on effect of the three speech coders on energy distortion, vowel formants and pitch period.

4.1 Spectral Energy Distortion

The first group of experiments demonstrates the percentage of frames which has certain spectral distortion. Similar experiments were done in [8] for Japanese phonemes and in [9] and [10] for English phonemes. Firstly, the speech samples in the time domain are converted into frequency domain using Fourier Transform applied on frame-by-frame basis. Each frame length is $32ms$. Overlapping is half of the

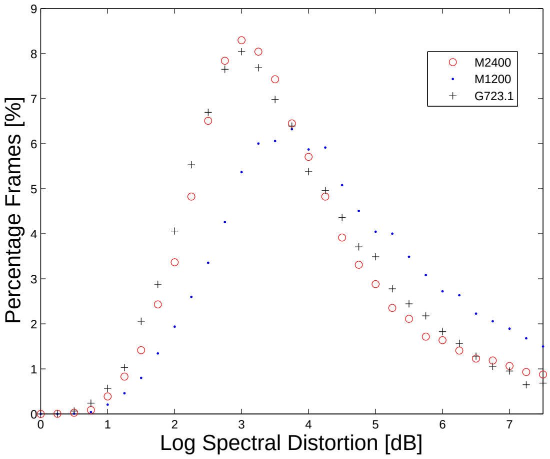

Figure 1: Effect of speech coders on speech Arabic vowels

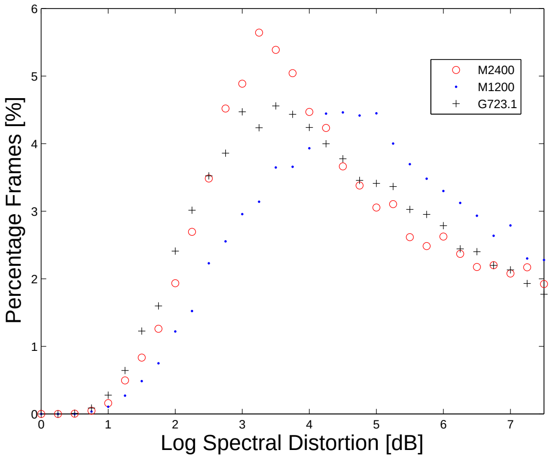

Figure 2: Effect of speech coders on Arabic consonants

frame length. Secondly, the spectral distortion is calculated on frame-by-frame basis using equation (1). Fig. 1 shows the effect of three coders, Melp2400, Melp1200, and G723.1, on the value of spectral distortion of Arabic vowels. Horizontal axis represents the log spectral distortion calculated using equation (1). The vertical axis represents the percentage of frames of a certain value of distortion. The mean value of distortion for all coders is around 3-4 dB. The highest value comes with Melp1200 coder. Also, the standard deviation of spectral distortion distribution of Melp1200 coder is the highest. From this and on the average, the processed frames by Melp1200 coder suffer from larger values of distortion than those of the other two coders. The same experiment was performed on the Arabic consonants. Results are demonstrated in Fig. 2. As shown in it, the three coders give approximately the same relative performance for consonants as those of vowels. However, the means of distorted spectra increase to be around 3.5-4.5 dB compared to original ones. Also, the standard deviations of the three distributions are higher than those in Fig. 1.

4.2 Vowel Formants

In this section, the effect of speech coders on the vowel formants is studied. The most important parameters which characterize a vowel formant are the formant frequency, its amplitude and bandwidth. These three parameters are studied extensively below for Arabic vowels.

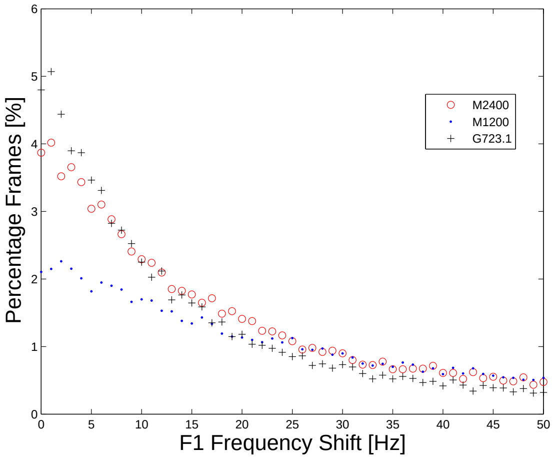

Figure 3: Effect of speech coders on the first Arabic speech vowels formant frequency

4.2.1 Formants Frequency Shift

Here, the shift in the formant frequency is measured. This shift is calculated using the absolute difference between the formant frequency of the original and processed speech. Fig. 3 presents the relation between the frequency shift of the first formant and the percentage of frames suffering from this shift for all studied vowels of Arabic. As illustrated in the figure, G723.1 coder has the least frequency shift of the first formant. The worst case belongs to Melp1200 coder. Other formants exhibit almost the same performance.

For all coders, there is a percentage of frames which suffer from more than 50 Hz frequency shift for first formant (F1). This value can be considered as large one compared to the real values of that formant

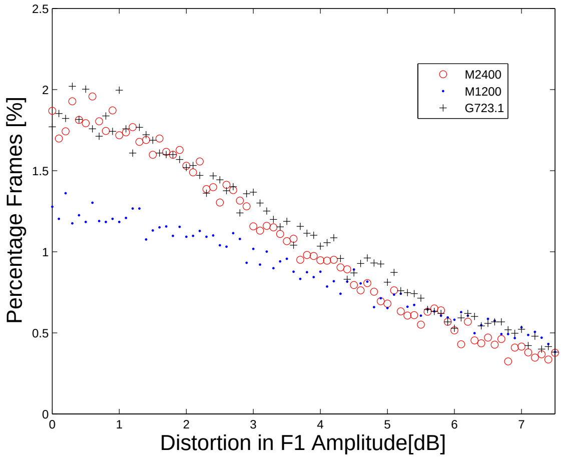

Figure 4: Effect of speech coders on the first Arabic speech vowels formant amplitude

which range from 300-500 Hz. This notice motivated us to study the frequency translation of the formant from the its original critical band to another one on vowel-by-vowel basis as indicated below.

4.2.2 Formants Amplitude Distortion

The amplitude of formants is calculated to see the effect of speech coders on them. Fig. 4 gives the results of distortion in the first formant amplitude due to the three speech coders. It's clear that the effect of Melp2400 and G723.1 are almost the same. The worst effect is due to Melp1200 coder. Other formants suffer from almost the same effect.

4.2.3 Effect of First-Formant Translation

Due to large shifts in the first formant frequency, we study here the coders' effect on the translation of formant frequency from a critical band to another one. We selected ”FATHA” vowel as a representative to Arabic vowels to perform this experiment. As shown in Fig. 5, a high percentage of frames have a frequency shift more than 40 Hz which may cause the formant to be transferred to another critical band. Because of the importance of the first formant in characterizing the reference vowel, this effect may cause the vowel to be perceived differently.

The following table shows the percentage of frames for ”FATHA” vowel that suffers from formant frequency translation to another critical band. The first formant frequency was calculated and the corresponding critical band was determined. These calculations were made for both original and processed signal. If the critical band of the processed frame was different than that of the corresponding original frame, it was counted.

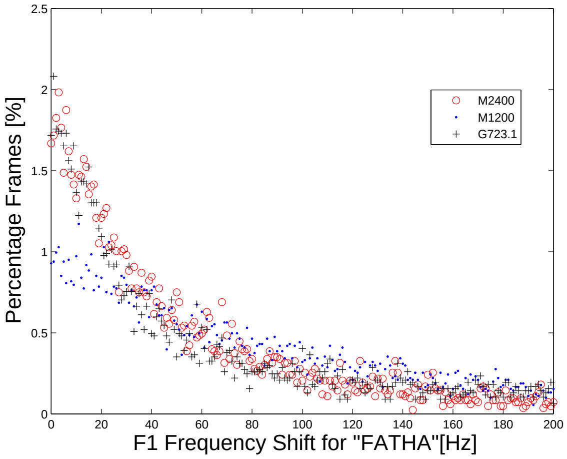

Figure 5: Effect of speech coders on ”FATHA” first formant frequency

Table 1: Percentage of frames with F1 translation to another critical band

Speech coder	Percentage of frames
Melp2400	45.02%
Melp1200	44.00%
G723.1	36.35%

4.2.4 Formant BW Distortion

This set of experiments was targeting the effect of speech coders on the bandwidth of vowels formants. We found that more than 75% of the frames suffer from zero BW shift. The remaining 25% of the frames has a shift in the range of 1-20 Hz which can be considered negligible when compared to the original formant frequency.

4.3 Pitch Period Distortion

Using all voiced frames in our Arabic database (vowels & voiced consonants), the pitch period distortion was calculated and plotted in Fig. 6. From the figure and corresponding results, it was found that more than 95% of voiced frames suffer from approximately no distortion.

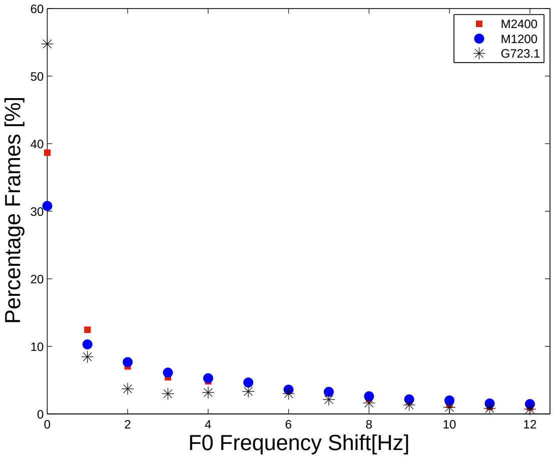

Figure 6: Effect of speech coders on pitch period

5 Conclusion

In this work, the effect of three speech coders on Arabic was reported. The experiments show that the Melp1200 results in larger values of energy distortion than those of the other two coders. Melp2400 coder has the least destructive effect on Arabic. Another set of experiments on Arabic vowels shows that the first formant of a vowel, on the average, suffers a frequency shift in the order of 50 Hz from its original value. This frequency shift can be considered large compared to the average value of the first formant value which ranges in (300-500) Hz. Also, the amplitude distortion of the vowel formants due to Melp1200 coder was the highest, compared to the distortion of the other coders. High percentage (35-45%) of formants is translated to another critical band due to coding effect for 'FATHA' vowel. This translation may lead to a change in the vowel perception. On the other hand, BW of that formant suffers from no change for more than 75% of the frames and a tinny change (compared to the average formant frequency) for the remaining percentage.

Studying the effect of coders on the pitch period shows that there is almost no distortion in pitch period value. This can be justified that the speech coder extracts the pitch period from each frame, encodes it and transfers their bits to the decoder directly. This special care of pitch period led to the tiny values in the distortion values.

Acknowledgements: Special thanks go to Prof. Mohamed Fathy, Elec. and Comm. Dept., Cairo Univ., for providing us with the Arabic speech database, his support and valuable comments.

References:

[1] Cisco, Internetworking Technologies Handbook. Cisco Press, 2003.

[2] W. C. Chu, Speech Coding Algorithms, foundation and evolution of standardized coders, John Wiley & Sons, Hoboken, New Jersey, 2003

[3] J. Benesty, M. M. Sondhi and Y. Huang, Springer Handbook of Speech Processing, Springer-Verlag, Berlin Heidelberg, 2008

[4] S. So and K. K. Paliwal, ”Comparison of LSF and ISP representations for wideband LPC parameter coding using the switched split vector quantiser”, in Signal Processing and Its Applications, Proceedings of the Eighth International Symposium, vol. 2, pp. 595-598, 2005.

[5] S. Sakayori, T. Kitama, S. Chimoto, L. Qin, and Y. Sato, ”Critical spectral regions for vowel identification”, Neuroscience Research, vol. 43, no. 2, pp. 155-162, 2002.

[6] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[7] Matlab Version 7.0: The Mathwork, Inc., 2004

[8] P. C. Nguyen, M. Akagi, and B. P. Nguyen, ”Limited error based event localizing temporal decomposition and its application to variable-rate speech coding”, Speech Communication, vol. 49, no. 4, pp. 292-304, 2007.

[9] V. Parsa and D. Jamieson, ”Interactions between speech coders and disordered speech”, Speech Communication, vol. 40, no. 3, pp. 365-385, 2003.

[10] A. McCree, ”Reducing speech coding distortion for speaker identification”, in INTERSPEECH-Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 2006.