CELP Coder Modification for the Voice Conversion

A. Guerid and A. Houacine
LCPTS, Faculty of Electronics and Informatics, USTHB, BP 32, EL Alia, BabEzzouar, Algiers, Algeria
Email: abguerid@gmail.com, ahouacine@usthb.dz

***Abstract*—Voice Conversion (VC) consists in modifying a source voice to a target speaker voice. In our approach, we modified only the Code excited linear Predictive (CELP) coder by introducing a pre-processing before the coder for the voice conversion. The decoder part of CELP was not modified. This allows maintaining the transmission rate. Our approach for conversion consists in separating the voiced and unvoiced frames, and thus two different conversion functions are associated. The Spectral Frequency Parameters LSF parameters are adopted to represent the vocal tract and Gaussian Mixture Models (GMM) are used to calculate the conversion functions. The pitch for the voiced frames is transformed by linear conversion. The model was tested for conversions between male and female voices.**

***Index Terms*—voice conversion, gaussian mixture model (GMM), CELP, LPC, speech coder**

I. Introduction

Voice Conversion (VC) process consists in the modification of the speaker source voice *x* toward another target speaker voice *y*, while preserving the natural aspects of the transformed speech. Several methods were introduced for the calculation of the voice conversion function, among which we mention the vector quantization (VQ) [1], the Hidden Markov Model (HMM) approach [2] and the Gaussian Mixture Model (GMM) method [3].

In this paper we applied the VC to the CELP speech coder. The CELP algorithm is widely used in the field of communications [4], [5]. It can be used with some modification to calculate the excitation such as such as Algebraic (ACELP) [6] or Low Delay (LD-CELP) [7].

Introduced by B. Atal [8], this algorithm works according to the analysis by synthesis procedure [9]. A pre-processing is associated only to the CELP coder part to perform the computation of the conversion functions. It consists, in separating the voiced and unvoiced frames and performing LPC analysis for these two kinds of frames. Thereafter two conversion functions are calculated; one transforms the LPC coefficients and the residue of the voiced frames, while the other transforms the unvoiced frames. We used the Line Spectral Frequency parameters (LSF) to represent the vocal tract and GMM to calculate the conversion function.

Manuscript received June 11, 2014, revised December 12, 2014.

II. The CELP Coder

Most modern speech codecs are based on the principle of CELP coding [9]. They exploit a simple source/filter model of speech production, where the source corresponds to the vibration of the vocal cords or/and to a noise produced at a constriction of the vocal tract, and the filter corresponds to the vocal/nasal tracts. Based on the quasi-stationarity property of speech, the filter coefficients are estimated by linear prediction and regularly updated (typically at 20ms time-intervals).

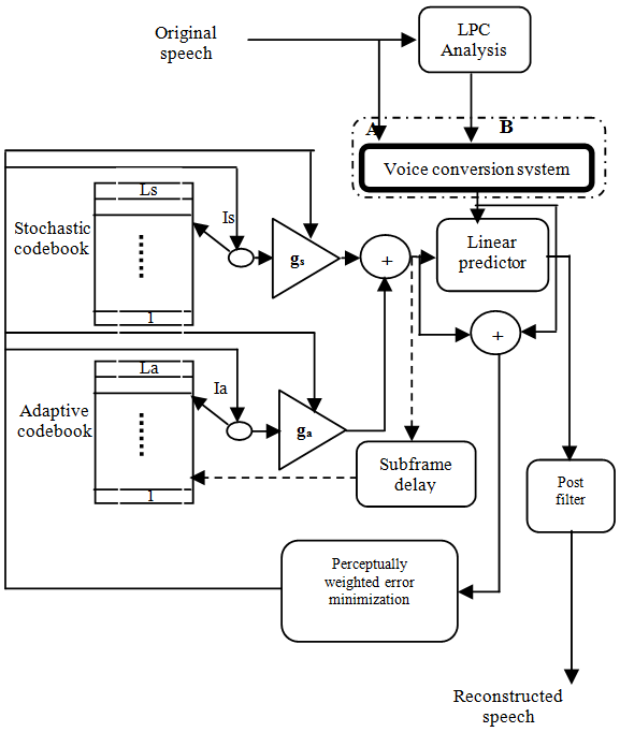

Figure 1. Diagram of CELP speech coder with the suggested conversion system

In the basic diagram of a CELP coder (Fig. 1), the analysis window is divided into several sub-windows and the excitation is calculated for each analysis sub-window by vector quantization [10]. The excitation signal is modeled by a linear combination of vectors, extracted from the adaptive and stochastic codebooks of well defined size. This global excitation signal is the result of the addition of the two elementary excitations: the first is a vector-code extracted from the adaptive codebook of

doi: 10.12720/ijsps.4.2.124-127

dimension L_a, with the index I_a, and weighted by the gain g_a. The second is the vector-code extracted from the stochastic codebook of dimension L_s, with the index I_s and weighted by the gain g_s.

The principle of the analysis of CELP speech coder consists thus in finding the parameters of the excitation (indices and gains) to minimize the Perceptual error.

III. The CELP Decoder

For the synthesis of the speech (Fig. 1), the decoder uses, for forming the excitation signal, the parameters transmitted by the CELP coder and which are the indices I_a, I_s, g_a and g_s, and uses the same adaptive and stochastic codebooks, and the LPC parameters a_i.

To minimize the noise due to the quantification and to improve quality of synthesis a post filtering is used with the same parameters of the LPC filter of the processed window. The transfer function $H(z)$ corresponding to this post filtering is defined by:

$$H(z) = A(z) / A(z/\alpha) \tag{1}$$

with

$$A(z) = 1 - \sum_{i=1}^{k} a_i z^{-i} \tag{2}$$

IV. The Voice Conversion System

The conversion system that we associate to the coder CELP (Fig. 1) is introduced just after the LPC analysis, so that the configuration of the decoder remains unchanged.

A. Calculation of the Conversion Functions

The first operation in our system consists in calculating, in the training phase, the GMM parameters used in our case as models for the voiced and unvoiced frames. The theory of the conversion system proposed consists in separating the speech source signal into voiced and unvoiced frames, and the fundamental frequency is thus computed only for the voiced frames. The analysis window used in our experiments is of 20ms with an overlap of 10ms.

For the calculation of the conversion function of the spectrum representing the vocal tract represented by the LSF parameters, we consider the Gaussian mixture model GMM for the joint probability $P(x, y)$ of the source and target training speech [11]. The parameters involved are (α, μ, Σ) corresponding respectively to the weighting factors, the means, and the covariance matrices. The two last parameters can be represented as follows:

$$\Sigma = \begin{bmatrix} \Sigma^{xx} & \Sigma^{xy} \\ \Sigma^{yx} & \Sigma^{yy} \end{bmatrix}$$

and

$$\mu = \begin{bmatrix} \mu^{x} \\ \mu^{y} \end{bmatrix}$$

The GMM model parameters are estimated using the EM algorithm. Then the spectral envelop conversion is determined by the function that transforms the LSF parameters of the source x to those of the target y, and given by:

$$F_1 C(x) = \sum_{i=1}^{M} h_i(x) \left[\mu_i^y + \Sigma_i^{yx} \left(\Sigma_i^{xx} \right)^{-1} \left(x - \mu_i^x \right) \right] \tag{3}$$

where $h_i(x)$ is the posterior probability that a given vector x can be generated by the class index i, and is calculated by applying the Bayes rule, as :

$$h_i(x) = \frac{\alpha_i N\left(x; \mu_i^x, \Sigma_i^{xx}\right)}{\sum_{j=1}^{M} \alpha_j N\left(x; \mu_j^x, \Sigma_j^{xx}\right)} \tag{4}$$

For the transformation of the fundamental frequency we use the linear conversion function defined by:

$$F_L = \left(\frac{(f_x - \mu_x)}{\sigma_x} \right) \sigma_y + \mu_y \tag{5}$$

where μ_y and σ_y represent the mean and variance of the target pitch, and μ_x and σ_x represent those of the source pitch.

For the conversion of the unvoiced frames, we used the same technique as that introduced in [3]. After alignment, achieved by a Dynamic Time Warping algorithm (DTW), on the LSF parameters, we consider, as training vectors, only the pairs corresponding to unvoiced frames of both the source and the target speech. We used then a Gaussian mixture model for the joint probability for this training step.

The parameters of this model are estimated with the EM algorithm. And the conversion function F_2C will have the same form as that represented in (3).

B. The Conversion Process

Our conversion system realized is detailed in (Fig. 2). The first version ou our conversion system was tested without including the part (C). That means that we have not transformed the fundamental frequency of the voiced frames.

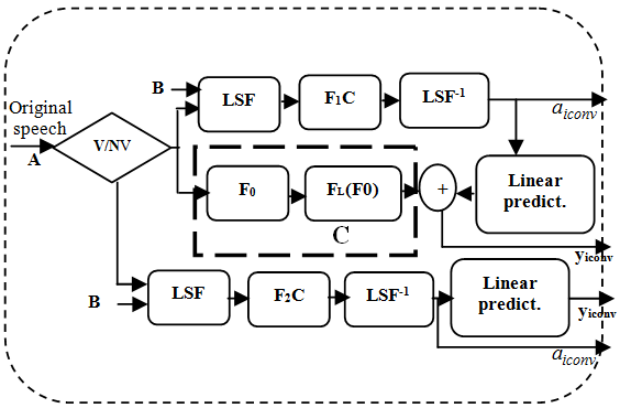

Figure 2. Overview of the proposed system

The whole process of conversion can be performed through the following steps:

1) *Case of voiced frames:*

- Transform the a_i parameters of the LPC analysis into LSF parameters;
- Use the conversion function F_1C, as defined in (3), to transform the LSF parameters;
- Reconvert the transformed LSF parameters to a_{iconv} parameters;
- Filter the voiced frame with a_{iconv} to obtained the residue y_{iconv};
- Use the (a_{iconv}, y_{iconv}) parameters and continue with the CELP coder process.

2) *Case of unvoiced frames:*

- Transform the a_i parameters of the LPC analysis into LSF parameters;
- Use the conversion function F_2C, as given similarly to (3) to transform the LSF parameters;
- Reconvert the transformed LSF parameters to a_{iconv} parameters.
- Filter the voiced frame with a_{iconv} to obtain the residue y_{iconv};
- Use the (a_{iconv}, y_{iconv}) parameters and continue with the CELP coder process.

From our experiments, we noticed that the transformed frames, when synthesized with only the residue y_{iconv} of the linear predictor resulting from the obtained coefficients a_{iconv}, are very disturbed and lose their quality of voicing.

To improve the voicing quality, we considered a periodic signal, generated by the transformed fundamental frequency (part C of Fig. 2), and added to the residue y_{iconv}. This procedure concerns only the voiced frames and is applied before using the parameters (a_{iconv}, y_{iconv}). This led to a neat improvement in quality for the resulting converted speech signal.

The treatment of conversion used in our approach aims to adding the task of voice conversion to a target speaker, while maintaining the rate of transmission used by the CELP coder. Only the parameters of excitation and the LPC coefficients are transmitted. For each window of index i (Fig. 3) consisting of N samples (N is equal to 320), from n1 to n2. We calculate the fundamental frequency f_i^0 (for the case of voiced frames), the parameters a_i and the residue y_i of the LPC analysis filter. Their transformation is performed through the respective conversion functions as estimated in a training phase.

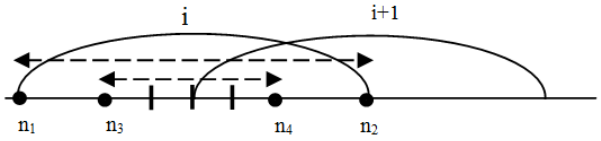

Figure 3. Frames and sub-frames of the signal to be converted

Thereafter, we calculate the parameters defining the excitation, and which are the indices Ia, Is, and the gains ga and gs, by applying the theory of the CELP coder. Each frame [n3 to n4] is divided into 4 sub-frame of N" samples (N" equal to 80). The content of the adaptive codebook is updated after processing each sub-frame, while the stochastic codebook is calculated only at the beginning of the frame. For our case we took the dimensions of the adaptive and stochastic codebooks equal to 256 vectors. The next parameters f_{i+1}^0 and y_{i+1} of the window index $i+1$ are obtained by sliding of the window index i by N' samples (N' equal to 160), and performing the computational process in the same manner.

V. Evaluation

To evaluate our conversion system, we tested it on the Arabic language with the same training and testing corpus used in [3]. The corpus of speech recorded for the calculation of the parameters of the various functions of conversions is taken equal to five minutes for each speaker (one male and one female). The test set is composed of 8 sentences, 4 for each speaker. The sampling rate is taken equal to 16 kHz and the order of the LPC model is taken equal to 10. The system is tested only for two types of conversions that are male to female (m-f), and female to male (f-m).

To measure the performances we used objective and subjective tests. The objective test consisted in calculating the normalized spectral distortion given by the ratio of the spectral distance (transformed, target) signals $d(t,Y)$ and (source, target) signals $d(X,Y)$ defined as following:

$$Dist = \frac{\frac{1}{N}\sum_{1}^{N} d(t,Y)}{\frac{1}{N}\sum_{1}^{N} d(X,Y)} = \frac{\sum_{1}^{N}\left(\sqrt{\sum_{i=1}^{M}|t_i - y_i|^2}\right)}{\sum_{1}^{N}\left(\sqrt{\sum_{i=1}^{M}|x_i - y_i|^2}\right)} \quad (6)$$

The subjective evaluation that we used for our conversion system is given by a metric evaluation. We used the "Perceptual Evaluation of Speech Quality" [12], ABX test, to evaluate the converted voice at the perceptive level, and the MOS (Mean opinion score) test, which informs about the quality of the listened voice.

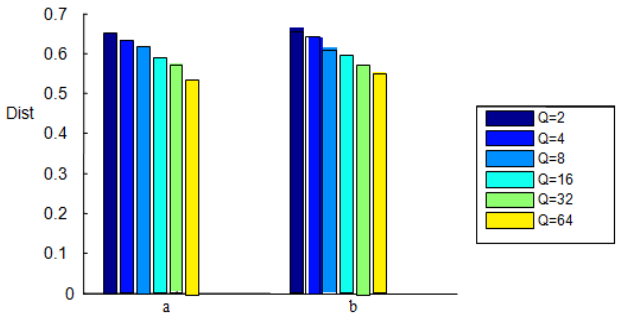

Figure 4. Normalized spectral distortions between the targets and converted envelopes (a) male-female transformation, (b) female-male.

Fig. 4 represents a comparison of performances for different values of gaussian mixture models, as M=[2 4 8

16 32 64], for the two types of conversions (m-f) and (f-m). We notice that the spectral distortion error is inversely proportional of the number M of gaussians, and the distortion error obtained for the conversion (m-f) is better than of that obtained for (f-m).

The results of PESQ presented for each sentence in the following Table I and Table II are obtained for the number of gaussien M equal to 64. We notice that the results of the two types of conversion (m-f) and (f-m) present almost the same values in average.

The objective test obtained (Table III), for different values of gaussian mixtures model M [8 16 32 64], shows that the distortion estimated for m-f conversion is better than that of type f-m. We also notice that this distortion is inversely proportional to the number of gaussian mixture model M.

The results obtained through the subjective evaluation are presented in the Table IV. We notice that the conversion (m-f) is better than that obtained for the type (f-m), and this applies for the two tests carried out. For ABX test the listeners judged the (m-f) conversion is more successful.

TABLE I. MALE TO FEMALE CONVERSION (M-F)

	Sentence 1	Sentenc 2	Sentence 3	Sentence 4
PESQ	1.77	1.70	1.64	1.56

TABLE II. FEMALE TO MALE CONVERSION (F-M)

	Sentence 1	Sentence 2	Sentence 3	Sentence 4
PESQ	1.67	1.63	1.58	1.53

TABLE III. RESULTS OF OBJECTIVE TEST

Gaussien value / Types of conversion	M=8	M=16	M=32	M=64
m-f	0.62	0.58	0.56	0.54
f-m	0.64	0.60	0.58	0.56

TABLE IV. RESULTS OF SUBJECTIVE TESTS

Gaussien Value / Types of conversion	M=8	M=16	M=32	M=64
ABX (m-f) %	25	20	18	25
MOS (m-f)	2.1	2.1	2.1	2.1
ABX (f-m) %	20	25	23	23
MOS (f-m)	2.3	2.2	2.1	2.1

VI. CONCLUSION

The system we proposed performs the voice conversion task without modifying the excitation parameters (Ia, Is, ga, gs) and the coefficients a_i of the CELP coder. The structure of CELP decoder remains thus unchanged.

This approach has advantage in maintaining the transmission rate of the CELP speech coder. That means the modifications provided at CELP coder does not increase the transmission rate.

The results obtained through this article motivated us to initiate a work which consists in implementing and evaluating our system on a real DSP environment.

REFERENCES

[1] M. Abe, S. Nakanura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1998, pp. 655-658.

[2] C. Wu, C. Hsia, T. Liu, and J. Wang, "Voice conversion using duration embedded bi-HMMs for expressive speech synthesis," *IEEE Transaction on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1109-1116, 2006.

[3] A. Guerid, A. Houacine, R. Andre-Obrecht, and H. Lachambre, "Performance of new voice conversion systems based on GMM models and applied to Arabic language," *International Journal of Speech Technology*, vol. 15, no. 4, pp. 477-485, 2012.

[4] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The proposed federal standard 1016 4800 bps voice coder: CELP," *Speech Technology Magazine*, vol. 5, pp. 58-64, Apr. 1990.

[5] ITU-T Recommendation G.723.1, Speech Coders: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s, ITU, Mar. 1996.

[6] R. Salami, *et al.*, "Design and description of CS-ACELP: A toll quality 8kb/s speech coder," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 116-130, Mar. 1998.

[7] N. Y. Ku1, C. Y. Yeh, and S. H. Hwang, "An efficient algebraic codebook search for ACELP speech coder," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, 2014.

[8] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. on ASSP*, vol. 27, no. 3, pp. 247-254, Jun. 1979.

[9] N. Moreau, *Tools for Signal Compression*, ISTE Wiley, 2011

[10] N. Moreau, *Techniques de Compression des Signaux*, Paris: Edition Masson, 1995.

[11] A. Kain, "High resolution voice transformation," Phd. Dissertation, OGI School of Science and Engineering at Oregon Health and Science University, 2001.

[12] ITU-T Recommendation P.862.2, ITU-T Study Group 12 (2005-2008) under the ITU-T Recommendation A.8 Procedure, 13 Nov. 2007.

Abdelkader Guerid was born in Algeria. He received his magister in 1999 from polytechnic national school of Algeries. He received his doctorate in 2012 from USTHB. He is currently an associate research worker at the LCPTS laboratory, USTHB. His fields of interest are signal processing, analysis and synthesis of speech and voice conversion.

Amrane Houacine was born in Algeria. He is a professor at the faculty of Electronics and Informatics, University of Sciences and Technology Houari Boumédienne of Algiers, Algeria. Previous research areas were on fast algorithms, adaptive filering, and image processing applied to remote sensing. His current research interests include speech, signal and image processing and their applications to human-machine interaction.