HAL
archives-ouvertes.fr

Prediction of Cepstral Excitation Pulses for Voice Conversion

Fadoua Bahja, Joseph Di Martino, El Hassan Ibn Elhaj, Driss Aboutajdine

▶ **To cite this version:**

Fadoua Bahja, Joseph Di Martino, El Hassan Ibn Elhaj, Driss Aboutajdine. Prediction of Cepstral Excitation Pulses for Voice Conversion. 5th. International Conference on Information Systems and Economic Intelligence - SIIE 2012, Feb 2012, Djerba, Tunisia. hal-00761776

HAL Id: hal-00761776
https://hal.inria.fr/hal-00761776

Submitted on 10 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prediction of Cepstral Excitation Pulses for Voice Conversion

Fadoua BAHJA*, Joseph Di Martino†, El Hassan Ibn Elhaj‡ and Driss Aboutajdine*
*LRIT laboratory, Unit associated to CNRST, URAC 29 Faculty of Sciences, Rabat, Morocco
†INRIA Nancy - Grand Est / LORIA-UHP Vandoeuvre-les-Nancy, France
‡INPT, Rabat, Morocco

***Abstract*—Voice conversion is one of useful techniques to enhance pathological speech to be perceived as normal speech, although it concerns also the modifications of normal source speaker's speech to be perceived as if a target speaker had uttered it. The parameters to be converted are obtained by matching the spectral envelope of the vocal tract for the source and the target speech. Gaussian Mixture Models (GMMs) parameters are determined for providing conversion functions. The main contribution of our study consists in the prediction of Fourier cepstrum coefficients related to the excitation signal. Such a prediction leads to a satisfactory voice conversion system. Subjective perceptual results indicate that the proposed approach yields significant improvements in quality of the converted voice.**

I. Introduction

The goal of voice conversion is to modify a source speaker's speech to be perceived as if a target speaker had uttered it.

The scope of voice conversion is large enough and varied: the personalization of Text-To-Speech (TTS) systems [11], speech synthesis [9] and the enhancement of pathological speech [15], [3], [7].

In the work proposed in this paper, we focus specially on a conversion technique based on Gaussian Mixture Models (GMMs).

The GMMs model robustly the acoustic space of speech and it has been used successfully as a method for spectral transformation. Hence the algorithm proposed in this paper evaluates the GMMs model whose parameters are calculated using a joint density estimation [8], [11] and are used for determining the voice conversion function.

In the literature, several methods for converting voices, from a source speaker to a target one, using the GMM modelling to provide probabilistic classification functions, were presented [19], [18], [17], [14].

The originality of the proposed algorithm concerns the prediction of cepstral excitation pulses we use for determining the target excitation spectrum. The real cepstrum of a signal is defined as the inverse fourier transform of the log magnitude spectrum. We focus, in this approach, on three principal steps: the first one based on the Dynamic Time Warping (DTW) algorithm, in order to create the mapping vector list between the source and target cepstral vectors and on Vector Quantization (VQ); the second step consists in calculating the conversion function based on GMM parameters; and the last step consists in predicting the cepstral excitation pulses.

In this algorithm, we use two main stages to convert the voices: the first one consists in a training procedure in order to estimate the conversion functions; the second stage consists in a test evaluation procedure where we use these functions to convert the source voices into the target voices.

To evaluate the performance of the proposed algorithm, we use 3 parallel corpora composed of 50 sentences pronounced by 3 male speakers. The Signal to Error Distortion (SED) was evaluated and perceptual tests were also carried out to assess the effectiveness of our voice conversion approach.

Objective speech quality measurements and subjective listening test results for the proposed algorithm are given in the sequel.

The paper is organized as follows. In section II, we describe all the different stages implemented in our algorithm. In section III, we presents our experimental results. And finally, we draw some conclusions.

II. Voice Conversion algorithm based on cepstral excitation pulses prediction

Our voice conversion system has three major stages: speech analysis, spectral conversion and speech synthesis (see figure 1). In this section, we present a description of each component according to our approach as follows.

A. Speech analysis

Let $x = [x_1, x_2, \cdots, x_M]$ be the sequence describing spectral vocal tract vectors of a succession of speech sound pronounced by the source speaker and $y = [y_1, y_2, \cdots, y_N]$ be the spectral vocal tract vectors characterizing the same speech sound pronounced by the target speaker.

1) Dynamic Time Warping (DTW): The main idea of the approach proposed in this paper consists in aligning the same phonetic content vector from the source speech features and the target features.

Figure 2 illustrates a warping parallelogram where M and N are the number of frames of the two spectral patterns.

For each couple vector (i, j) we associate three possible ways:

1) The way 1 comes from the couple vector $(i-2, j-1)$;
2) The way 2 comes from the couple vector $(i-1, j-1)$;
3) The way 3 comes from the couple vector $(i-1, j-2)$.

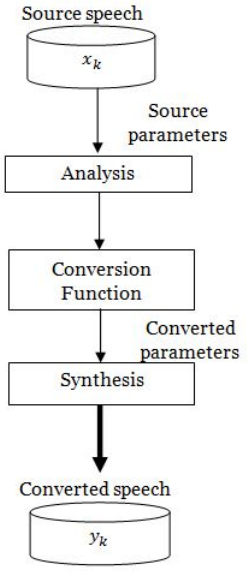

Fig. 1. Bloc diagram of voice conversion system components

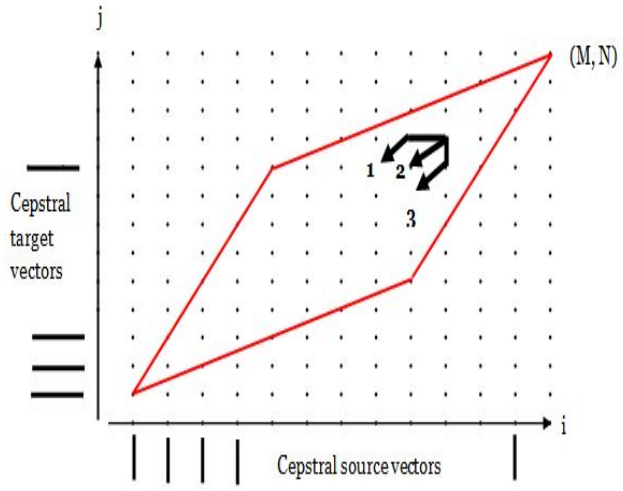

Fig. 2. The implicit parallelogram used in DTW matching.

Because of the difference between the duration of the source and target speech, the warping region is adjusted adaptively to accommodate the spectral patterns to be matched.

2) Vector Quantization: The Vector Quantization (VQ) [16] process is performed on all vectors in the entire database using the source and target speaker vectors.

In the VQ-based method, proposed by [11], each cepstral vector of the source voice and its associated vector (by DTW) are concatenated. These extended vectors are classified using an accelerated LBG (Linde Buzo and Gray) vector Quantization algorithm [12]. The main idea of VQ conversion technique is to classify each vector by finding a codebook representing the concatenated source and target cepstral vectors. In our work, we use a codebook composed of 32 codevectors. The LBG VQ design algorithm is an iterative technique which uses a splitting method. Firstly, an initial codevector is set as the average of the entire training sequence. It is split into 2 vectors. The 2 final codevectors are then split into 4 and the algorithm is repeated until we obtain the desired number of codevectors.

B. Spectral conversion

1) The Gaussian Mixture Model (GMM): The distribution density of vectors x are represented as the sum of Q multivariate Gaussian densities, given by

$$p(x) = \sum_{i=1}^{Q} \alpha_i N(x; \mu_i; \Sigma_i), \sum_{i=1}^{Q} \alpha_i = 1, \alpha_i \geq 0 \quad (1)$$

Where N(x;μ;Σ) denotes a Gaussian distribution with a mean vector μ and a covariance matrix Σ. The total number of mixture components is Q. The α_i are the weighting coefficients of the Gaussian distributions.

α_i is estimated as the ratio between the number of vectors of class i ($N_{s,i}$) and the total number of vectors (N_s):

$$\alpha_i = \frac{N_{s,i}}{N_s}$$

μ_i denotes the vectors average of class i. It is calculated as follow:

$$\mu_i = \frac{\sum_{k=1}^{N_{s,i}} x_i^k}{N_{s,i}}$$

Σ_i represents the covariance matrix of source vectors of class i and Γ_i represents the cross-covariance matrix of target/ source vectors of class i. The classical estimation formula for calculating the elements of Σ_i and Γ_i are given by formula 2 and 3 [19]:

$$\sigma_{i,j} = E(x_i x_j) - E(x_i)E(x_j) \quad (2)$$

and

$$\gamma_{i,j} = E(y_i x_j) - E(y_i)E(x_j) \quad (3)$$

2) The conversion function: From the GMM probability distribution $N(x; \mu; \Sigma)$, the conversion function is estimated to be a regression of the following form:

$$F(x) = E[y/x] = \sum_{i=1}^{Q} h_i(x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)] \quad (4)$$

where

$$h_i(x) = \frac{\frac{\alpha_i}{(2\pi)^{\frac{N}{2}}|\Sigma_i^{xx}|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu_i^x)^T(\Sigma_i^{xx})^{-1}(x-\mu_i^x)]}{\sum_{j=1}^{Q} \frac{\alpha_j}{(2\pi)^{\frac{N}{2}}|\Sigma_j^{xx}|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu_j^x)^T(\Sigma_j^{xx})^{-1}(x-\mu_j^x)]} \quad (5)$$

with

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

and

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

In order to transform source spectra to the corresponding target spectra, we estimate conversion functions given by formula 4.

In this paper, we use, for calculating the GMM parameters, the same approach as the one proposed by [19]. This algorithm is called "Iterative Statistical Estimation Directly from Data" (ISE2D).

C. Speech synthesis

1) Signal reconstruction: From the trained GMMs parameters, the conversion function is performed in order to calculate the converted vectors $\widehat{y}_k$.

$$\widehat{y}_k = F(x_k)$$

The main contribution of our approach consists in the prediction of the cepstral excitation pulses. The classical cepstrum coefficients c_p, with $p = 0, ..., O_c - 1$ (O_c is the number of vocal tract cepstral coefficients), are defined as the first coefficients of the inverse fourier transform of the log amplitude.

We explain in the following how to extract the cepstrum excitation signal from the input signal $x(n)$. Let

$$x(n) = h(n) \otimes e(n) \tag{6}$$

where $x(n)$ represents the convolution between the vocal tract $h(n)$ and the excitation signal $e(n)$.

We start by multiplying the speech signal $x(n)$ with a normalized Hamming window (see formula 7) of 512 samples followed by a Fast Fourier Transform (FFT) and the modulus of this output signal.

$$H(n) = \begin{cases} \frac{2\sqrt{\frac{L}{N}}}{\sqrt{4a^2+2b^2}}(a + b\cos(\frac{\pi(2n+1}{N})), if 0 \leq n < N \\ 0 \qquad\qquad\qquad otherwise \end{cases} \tag{7}$$

where L is the analysis step size, N is the length of the analysis window, a=0.54 and b=-0.46.

By calculating the log we get the Fourier log spectrum. By applying an Inverse Fast Fourier Transform (IFFT) we get the real log cepstrum related to the analysed frame.

From the real log-cepstrum, we can separate the vocal tract from the cepstrum excitation signal by eliminating the $O_c = 26$ first coefficients (which represent the vocal tract) as shown in figure 3.

Figure 4 describes all steps to obtain the cepstrum excitation signal.

To predict the cepstral excitation pulses, we propose the following steps:

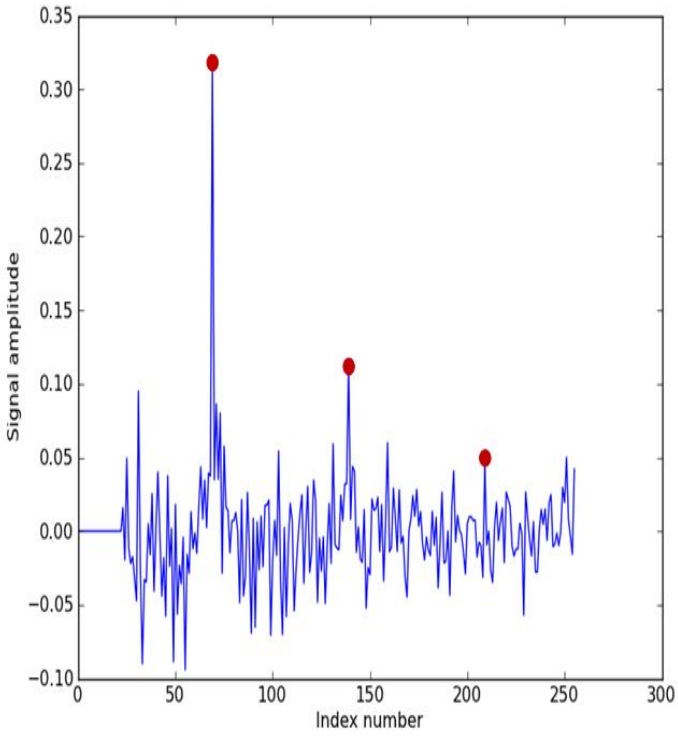

Fig. 3. log cepstrum of an excitation signal for a female voice. The red dots are related to the F0 and harmonic peaks.

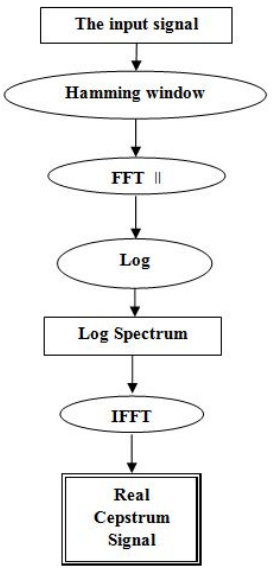

Fig. 4. Bloc diagram of the cepstrum excitation signal.

- We match the excitation cepstral coefficients of the source and target speech;
- We sort these coefficients in order to focus only on the maximum (absolute) amplitudes (positive and negative);
- We eliminate the zero amplitudes;
- We normalize the excitation pulses with the following formula:

$$log\_pulse\_index = \frac{\log \frac{Index}{N/2}}{C} \tag{8}$$

where:

  - N is the length of the analysis window, in our work N is 512 samples, that corresponds to 32 ms;
  - $Index$ represents the pulse index;
  - C is a normalization constant given by:

$$C = 2 \log \frac{O_c}{N/2} \tag{9}$$

This constant has been introduced in order to reduce the range of variation of the log indexes and to mimic a possible vocal tract cepstrum coefficient.

- Furthermore, we calculate the transformation functions for:
  - The vocal tract vectors;
  - The first cepstral coefficients $c0$;
  - The greatest positive amplitudes of the cepstral excitation pulses and their indexes;
  - The lowest negative amplitudes of the cepstral excitation pulses and their indexes.

  The conversion functions concerning the c0, the amplitude coefficients and the log indexes of the excitation pulses are calculated by mapping vectors composed of the concatenation of the first coefficient to be transformed and the $(O_c - 2)$ cepstral vocal tract coefficients (source and target).

The spectral synthesizer we used is based on the iterative sequential Nawab extrapolation approach [13]. The main interest of our algorithm is that it is not iterative and can operate in real time [4], [6].

III. ANALYSING AND TRAINING

A. Dataset

In order to train the proposed algorithm, the database used contains 3 parallel corpora of 50 sentences uttered by 3 male speakers (AL, CB and NG).

The utterances of each corpus are sampled at 16 kHZ. The average duration of each sentence is approximatively 2 seconds.

B. Experimental conditions

Table I summarizes the conditions of the experiments:

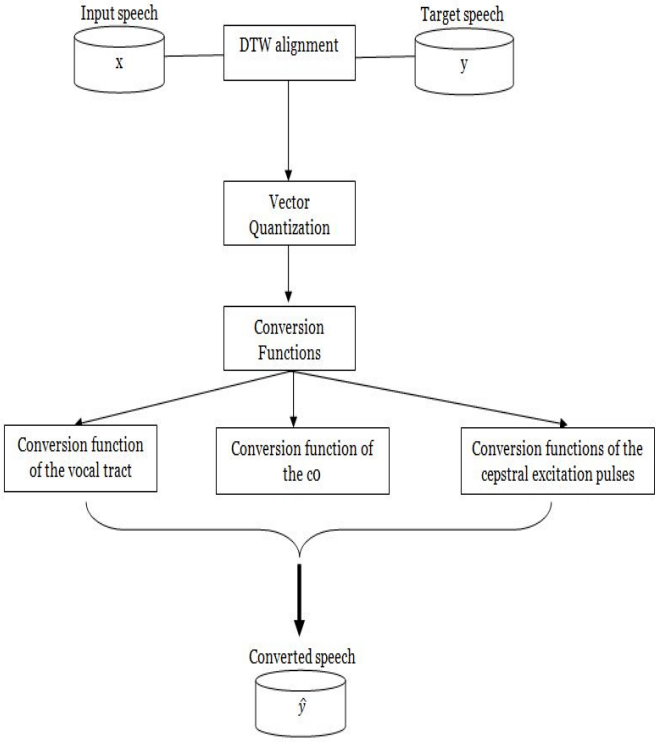

Fig. 5. Bloc diagram of the algorithm.

TABLE I
EXPERIMENTAL CONDITIONS

Number of training utterances	50 sentences / speaker
Number of test utterances	10 sentences
Analysis window	Hamming
Window length	32 ms
Shift length	4ms
Number of GMMs	32
Number of positive trained pulses	15
Number of negative trained pulses	15
Number of vocal tract cepstral coefficients	26
Number of classes in LBG VQ design	32

C. Objective evaluation

To objectively evaluate the performance of our spectral conversion, we use the Signal to Error Distortion (SED). It is estimated by the following formula:

$$SED = \frac{\sum_k \|y_k\|^2}{\sum_k \|y_k - \widehat{y}_k\|^2} \quad (10)$$

where y_k and $\widehat{y}_k$ are respectively the target and converted cepstral vectors.

TABLE II
SED VALUES ACCORDING THE ITERATION NUMBER FOR (CB→ AL) AND (NG→ AL) PARALLEL CORPORA

Number iteration	SED(dB) CB to AL corpus	SED(dB) NG to AL corpus
1	5.145	4.560
2	9.021	8.947
3	9.181	9.121
4	9.201	9.168
5	9.228	9.178
6	9.245	9.196
7	9.246	9.207
8	9.250	9.209
9	9.259	9.212
10	9.264	9.215

In our experiments, we used two parallel corpora CB→AL and NG → AL. And as shown in Table II, the SED is stabilized in around 10 iterations.

We can calculate the relative decrease SED rate by the given formula:

$$Relative\ decrease\ SED\ rate = \frac{SED_f - SED_i}{SED_f} \quad (11)$$

where SED_i and SED_f are the values of SED at the first and the last iteration respectively.

8.72% and 8.71% are respectively the relative decrease SED rate for the NG to AL corpus and for the CB to AL corpus.

D. Subjective evaluation

To make subjective tests, we evaluate our approach by extracting 10 utterances from the test dataset. The results of informal perceptual evaluations indicated that the converted speech is more and more closer to the target utterance by varying the number of excitation pulses we predict. The more we predict the better we perceive.

IV. CONCLUSION

We propose an algorithm of voice conversion by estimating the probabilistic conversion functions based on Gaussian mixture models and a joint density estimation. The originality of the proposed algorithm lies in the prediction of the cepstral excitation pulses. Otherwise, experimental results of the proposed algorithm are very promising concerning objective as well as subjective tests.

ACKNOWLEDGMENT

The authors wish to thank the INRIA institute for its support through the Euro-Mediterranean 3+3 M09/02 Oesovox project and the European COADVISE- IRSES (FP7) program.

REFERENCES

[1] M. Abe. A Study on Speaker Individuality Control. *PhD thesis, NTT Human Interface Laboratories*, March 1992.

[2] F. Bahja, J. Di Martino, E. Ibn Elhaj and D. Aboutajdine. An Improvement of the eCATE Algorithm for F0 Detection. *in Proceedings of the International Symposium on Communications and Information Technologies (ISCIT)*, pages 24-28, 2010. Tokyo, Japan.

[3] N. Bi and Y. Qi. Application of Speech Conversion to Alaryngeal Speech Enhancement. *in IEEE Transactions on Speech and Audio Processing, vol. 5 Issue:2, pp. 97-105*, March 1997.

[4] M. Chami, J. Di Martino, L. Pierron and E. Ibn Elhaj. Real-time Signal Reconstruction from Short-Time Fourier Transform Magnitude Spectra using FPGAs. *submitted for publication in SIIE conference, 2012*. Tunisia.

[5] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Dated via the EM Algorithm. *Journal of the Royal Statistical Society Serie B, flight 39, pp. 1-38*, 1977.

[6] J. Di Martino and L. Pierron. Synthétiseur numérique audio amélioré. *INRIA, Université Henri Poincaré Nancy 1, INPI, Paris, brevet "Oesovox" 10/02674*, 2010.

[7] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Statistical Approach to Enhancing Esophageal Speech based on Gaussian Mixture Models. *in Proceeding ICASSP, pp. 4250-4253, Dallas, USA*, March 2010.

[8] T. En-najjary, O. Rosec and T. Chonavel. A Voice Conversion Method Based on Joint Pitch and Spectral Envelope Transformation. *in Proceeding of International Conference on Spoken Language Processing. Jeju Island, South Korea*, October 2004.

[9] T. En-najjary. Conversion de voix pour la synthèse de la parole. *PhD thesis, Université de Rennes I*, France, 2005.

[10] N. Iwahashi and Y. Sagisaka. Speech Spectrum Transformation by Speaker Interpolation. *in IEEE International Conference on Acoustics, Speech and Signal Processing*, Australia, April 1994.

[11] A. Kain . High Resolution Voice Transformation. *PhD thesis, Oregon Health and Science University*, October 2001.

[12] Y. Linde, A. Buzo and R. M. Gray. An Algorithm for Vector Quantizer Design. *in IEEE Transactions on Communications, vol. COM-28, No. 1*, January 1980.

[13] S. H. Nawab, T. F. Quatieri, and J. S. Lim. Signal Reconstruction from Short-Time Fourier Transform Magnitude. in IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-31, No. 4, pp. 986-998, August 1983.

[14] K.-Y. Park and H.S. Kim. Narrowband to Wideband Conversion of Speech using GMM based Transformation. *in Proceeding ICASSP*, 2000. Istanbul , Turkey.

[15] Y. Qi. Replacing Tracheoesophageal Voicing Sources using LPC Synthesis. *in Journal of Acoustical Society of America, vol. 88, pp. 1228-1235*, 1990.

[16] D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker, Identification using Gaussian Mixture Speaker Models. *in IEEE Transactions on Speech and Audio Processing, vol. 3, No. 1, pp. 72-83*, January 1995.

[17] Y. Stylianou. Harmonic plus Noise Model for Speech, Combined with Statistical Methods, for Speech and Speaker Modification. *PhD thesis, Ecole Nationale Supérieure des Télécommunications*, Paris, France, 1996.

[18] H. Valbret. Système de conversion de voix pour la synthèse de la parole. *PhD thesis, ENST Paris*, September 1992.

[19] A. Werghi, J. Di Martino and S. Ben Jebara. On the Use of an Iterative Estimation of Continuous Probabilistic Transforms for Voice Conversion. *in Proceedings of the 5th International Symposium on Image/Video Communication over fixed and Mobile Networks (ISIVC)*, September 2010. Rabat, Morocco.