ON THE TRANSFORMATION OF THE SPEECH SPECTRUM FOR VOICE CONVERSION

G. Baudoin, Y. Stylianou

ESIEE, PSI Laboratory, BP 99, Noisy Le Grand, 93162 CEDEX, FRANCE
baudoing@esiee.fr, styliani@esiee.fr

ABSTRACT

In many speech applications, control of the speech individuality is required. These applications include the personalization of the voice of speech synthesizers, the restoral of voice individuality for interpreting telephony, the improvment of abnormal speech intelligibility. It is generally admitted that both prosodic and spectral parameters have to be changed in order to modify the speech individuality.

Several algorithms have recently been proposed for the spectrum control. This paper presents some improvments added to these previously proposed methods and compares 4 approaches in the same common framework of voice conversion for application to text to speech synthesizers.

1. GENERAL FRAMEWORK

This study has been realized in the general framework of voice conversion and was restricted to the transformation of spectral parameters. The main objective was to create new voices for text to speech synthesizers using speech segments concatenation.

Two approaches can be used to create new voices for speech synthesizers : record new speakers and apply automatic segmentation algorithms to segment the recorded signals in elementary segments as diphones, or modify existing segments from one original speaker by voice conversion techniques. We have studied the possibilities of the second approach.

We have used 2 speech databases, corresponding to 2 different speakers called the source speaker and the target speaker. The voice conversion consists in transforming the speech of the source speaker so that it sounds like the speech of the target speaker. During a training phase, a spectral transformation is learned using the training database. Then this transformation is applied to the test data. We have developped and compared several techniques of spectral transformation for voice conversion using the same analysis and synthesis technique, the same spectral parameters and the same databases.

1.1. Databases

We have used databases of the CNET (France Telecom) text to speech synthesizer. They are made of the same French logatomes for each speaker. They are sampled at 16 Khz with 16 bits.

From these logatoms, we have extracted the diphones which are effectively used by the synthesizer. By spectral analysis and temporal alignment we obtained 2 spectral vector databases, 80% of which were kept for the training and 20% for the test. The total number of cepstral vectors was 35000.

We have divided these databases in 2 parts, corresponding to voiced and unvoiced frames. All the tests were done in 2 situations : training and transformation on the total training database including voiced and unvoiced vectors, separate training and transformation for the voiced and unvoiced databases.

We also disposed of 2 sets of phonetically balanced sentences for the same speakers.

1.2. The Analysis and Synthesis method

The above databases were analyzed using the Harmonic + Noise Model (HNM) which allows high quality speech synthesis and prosodic modifications. The HNM performs a pitch synchronous decomposition of the speech signal making use of a sum of purely harmonic signals and of a modulated noise [1,2].

For voiced sounds, the speech spectrum is divided into two bands delimited by a time-varying frequency called the maximum voiced frequency. The lower band of the spectrum is represented solely by harmonically related sine waves. The upper band is modeled as a noise component modulated by a time-domain amplitude envelope.

The amplitudes of the harmonics that constitute the voiced part of speech are determined by a time-domain weighted least-squares technique [1]. For the unvoiced part (unvoiced frames and the upper band of the spectrum for the voiced frames) the energies of a bank of filter are calculated. The frequencies of the harmonic and of the filter bank are converted to a Bark frequency scale using the analytical formulas reported in [3].

In this study, during the training procedure the analysis is performed at a constant frame rate of 10 ms in order to allow time-alignment by a DTW algorithm. During the transformation procedure, first a synchronous analysis is done and the spectral envelope is calculated from the HNM parameters, the spectral parameters are modified, then the HNM amplitudes of sinusoïds and noise parameters are derived from the modified spectral envelope and the HNM synthesis is achieved with these modified parameters.

1.3. Spectral parameters

A continuous model of the spectral envelope that connects the obtained harmonics and the energies of the filter bank is estimated using the discrete regularized cepstrum method [4]. The spectral envelope is thus described by parameters that are analogous to the

standard Mel-Frequency Cepstrum Coefficients. They are noted c_i in the following text. An order p=16 was used.

The first cepstral coefficient c_0, which represents the energy of the frame was omitted from the training parameters. For some tests, the cepstral coefficients were normalized, using means and standards deviations calculated on the training databases.

The square euclidian distance d was used in the different algorihms to evaluate distances between 2 cepstral vectors $\mathbf{c}^S$ and $\mathbf{c}^T$, of size p.

$$d(\mathbf{c}^S, \mathbf{c}^T) = \frac{1}{p}\sum_{i=1}^{p}\left(c_i^S - c_i^T\right)^2$$

2. SPECTRAL TRANSFORMATIONS

It has already been demonstrated that a linear conversion of the frequency axis is not sufficient to transform the spectrum from one speaker to another and that the transformation should depend on the type of sound to be modified.

To take into account this dependency on the sound class, we have for some experiments classified the cepstral vectors in differents classes using vectorial quantization.

Four spectrum conversion methods have been compared :

1. Vector Quantization Mapping (VQM) [5] : Conversion is done by mapping two vector quantisation codebooks.
2. Statistical conversion (GMM) [6] : The spectral transformation uses a Gaussian Mixture Model (GMM) to modelize the acoustic space of the source speaker and is linear in each acoustic class.
3. Neural Networks conversion (NNETS): Multilayer perceptrons convert source spectral vectors.
4. Linear Multivariate Regression (LMR) [7] : Linear transformations converts source vectors.

2.1. VQM, spectral conversion by Mapping of Vector Quantization codebooks

Training phase : in the training phase, for each speaker training database a vector quantization (VQ) codebook of size N is constructed by the LBG algorithm. The cepstral vectors are then quantized with the appropiate VQ codebook and time aligned by DTW (Dynamic Time Warping).

From the aligned cepstral vectors couples, histograms of correspondances are calculated. These histograms indicate for each vector of the source VQ codebook, the numbers of associations (by DTW) of this vector with each vector of the target VQ codebook.

A Mapping codebook is calculated from the histograms. The mapping codebook associates to each vector of the source VQ codebook a transformed vector. Two kinds of mapping codebooks have been evaluated : max_mapping and weighted_mapping codebook. To each vector of the source VQ codebook, the max_mapping codebook associates the vector of the target VQ codebook corresponding to the maximum in the histogram, while the weighted_mapping codebook associates a vector which is a linear combination of vectors in the target VQ codebook, with weights given by the values in the histograms.

In the case of the weighted_mapping codebook, we have improved the subjective quality of the transformed voice by replacing this codebook by a natural_codebook which is obtained by replacing each vector of the weighted_mapping codebook by the closest unquantised target vector.

Transformation phase : a source cepstral vector is transformed by first vector quantizing it with the source VQ codebook and then substituting it with the corresponding vector in the mapping codebook.

In the VQM approach, the converted voice quality is limited by the vector quantization of the cepstral vectors.

2.2. GMM statistical spectral conversion method

Training phase : The first step of the statistical approach consists in fitting a gaussian mixture model (GMM) to the source vectors $\mathbf{c}^S$. Thus, the probability distribution of the observed parameters can be written as :

$$P(\mathbf{c}^S) = \sum_{i=1}^{m}\alpha_i N(\mathbf{c}^S; \mu_i, \Sigma_i) \qquad (1)$$

where $N(\mathbf{c};\mu,\Sigma)$ denotes the p-dimensional normal distribution with mean vector μ and covariance matrix Σ and α_i are normalized positive scalars. Note that the gaussian mixture model has the ability to model the acoustic space of a speaker as a combination of several components Ω_i (i=1,...m), where m is the number of mixture components. The conditional probabilities $P(\Omega^i / \mathbf{c}^S)$ that a given observation vector $\mathbf{c}^S$ belongs to one of the acoustic classes Ω_i is given by :

$$P(\Omega_i / \mathbf{c}^S) = \frac{\alpha_i N(\mathbf{c}^S; \mu_i, \Sigma_i)}{\sum_{j=1}^{m}\alpha_j N(\mathbf{c}^S; \mu_j, \Sigma_j)} \qquad (2)$$

The parameters of the GMM are estimated using the classic Expectation-Maximization (EM) algorithm of [9].

The following parametric form [2, 6] is assumed for the conversion function F :

$$F(\mathbf{c}^S) = \sum_{i=1}^{m} P(\Omega_i / \mathbf{c}^S)\left[\nu_i + \Gamma_i \Sigma_i^{-1}\left(\mathbf{c}^S - \mu_i\right)\right] \qquad (3)$$

The conversion function F is entirely defined by the p-dimensional vectors ν_i and the pxp matrices Γ_i, for i=1,...m. The

parameters of the conversion function are obtained by least squares optimization on the learning data so as to minimize the total squared conversion error ε between the N converted data $F(\mathbf{c}^{S,k})$ and the N target data $\mathbf{c}^{T,k}$:

$$\varepsilon = \sum_{k=1}^{N} d\left(\mathbf{c}^{T,k}, F(\mathbf{c}^{S,k})\right) \quad (4)$$

The complete learning procedure using the above statistical approach is depicted on figure 1, and a detailed description of the statistical approach using GMM can be found in [2, 6,10].

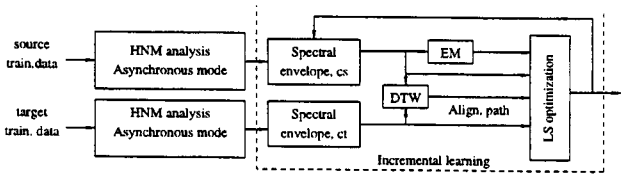

Figure 1: learning procedure for the GMM method

We distinguish three particular types of conversion functions : *Full* conversion where no constraints are applied either to the parameters of the GMM or to those of the conversion function, *Diagonal* conversion where matrices Σ_i and Γ_i are diagonal and *VQM-type* conversion if we omit the above two matrices from (3).

Transformation phase : To transform a source spectral vector, the conversion function F is applied directly to this vector.

2.3. NNETS spectral conversion method

NNETS have already been applied for voice conversion [11], transforming formants. Here multilayer perceptrons are used to convert source cepstral parameters. The best results were obtained with networks with 2 hidden layers of size 15, or with 3 hidden layers of size 12.

Training phase : The 2 cepstral vector databases are time aligned by DTW. This alignment gives the (input, output) couples of associated (source, target) cepstral vectors for the training of the neural networks. The cepstral vectors are normalized, using the corresponding source or target means and standard deviations.

For the source speaker a VQ codebook of size M is constructed. Using this codebook, the couples of cepstral vectors are classified in M different classes and a neural network is trained for each spectral class. The maximum number of classes was 64.

The criterium which is optimised during the training is the average quadratic cepstral distance d between the target cepstral vectors and the corresponding transformed source vectors.

In the case of a single class, a single neural net has to be trained with a large amount of data. we have tested 3 types of gradient back propagation learning algorithms for the neural net : global learning, stochastic learning and semistochastic learning.

In the global learning, at each iteration, the gradient is corrected by evaluation on the complete database. In the stochastic learning, at each step, the gradient is corrected by evaluation on a single vector randomly chosen in the database. In the semi-stochastic learning, the gradient is corrected by evaluation on a randomly chosen subset of N vectors containing one vector of each class of a VQ codebook of size N. The semi-stochastic approach was the best method both for speed and quality of results.

The initialization of the neural nets is done in such a way that at the beginning of the training, the output vectors of the network are very closed to the input vectors. So the original criterium is nearly equal to the distance between the source and target vectors. Biaises are initialised with very small random values. Weigths are fixed to one plus a small random noise on the direct connectns, they are set to zero plus a small random noise on the crossed connectns.

Transformation phase : a source cepstral vector is transformed by normalizing it with the source means and standard deviations, then applying the normalized vector to the neural network corresponding to its class and finally denormalising the output of the network with the target means and standard deviations.

2.4. LMR, Linear Multivariate Regression spectral conversion method

Training phase : The 2 cepstral vector databases are time aligned by DTW. The cepstral vectors are then normalised as previously described for NNETS. Then, the vectors are classified in different classes (using VQ) and a linear transformation is learned for each class.

Let $\mathbf{c}^{T,k}$ be a target cepstral vector and $\mathbf{c}^{S,k}$ be a source cepstral vector. The source vectors $\mathbf{c}^{S,k}$ are converted in $\mathbf{c}^{C,k}$ vectors by a matrix **M** calculated in order to minimize the criterium J:

$$\mathbf{c}^{C,k} = \mathbf{M}\mathbf{c}^{S,k}$$

$$J = \min \sum_{k=1}^{N} \sum_{i=1}^{p} \left(c_i^{T,k} - c_i^{C,k}\right)^2$$

The optimal solution is : $\mathbf{M} = \mathbf{C}_T'\mathbf{C}_S\left(\mathbf{C}_S'\mathbf{C}_S\right)^{-1}$

where $\mathbf{C}'$ represents the transposed matrix **C**, and $\mathbf{C}_S$ and $\mathbf{C}_T$ are the matrixes (Nxp) of the N source and target cepstral vectors.

Transformation phase : A linear transformation is used to convert source spectral vectors. A source cepstral vector is transformed by normalizing it with the source means and standard deviations, then applying the LMR matrix M corresponding to its class and finally denormalizing the transformed vector with the target means and standard deviations.

2.5. NNETS or LMR with context conversion

For NNETS and LMR methods we have also done the training using the context of each cepstral vectors. The transformation of a vector is done taking into account the preeceding and the following vector. For each actual vector of the source we have formed « context vectors » by concatening 3 cepstral vectors : the

actual vector in the middle, with the preeceding vector at the beginning and the following vector at the end.

In the LMR case, for the calculation of M, the matrix C_T is unchanged, but the matrix C_S is formed of the context vectors and it is 3 times larger as before.

In the NNETS case, the cepstral input vectors, which are context vectors are 3 times bigger as the output cepstral vectors.

3. EXPERIMENTS AND RESULTS

3.1. Quantitative results

The table 1 gives the normalized distances (average quadratic cepstral distances) obtained with the different methods, and distinguishes the cases voiced V, and unvoiced UV. A normalized distance is a distance divided by the distance between the souce and the target. The distances were calculated on the test databases.

Dtc = Normalized distances between target and converted vectors, Dsc = Normalized distances between source and converted vectors The numbers in brackets represent the numbers of classes or of components. (c_1 apart) means that c_1 is linearly transformed.

	Dtc V	Dsc V	Dtc UV	Dsc UV
VQM WeightedMap	0.30	0.77	0.20	0.84
VQM maxMap	0.41	0.88	0.28	0.92
Full (64)	0.28	0.75	0.20	0.83
Diag (128)	0.30	0.70	0.20	0.79
VQMtype (256)	0.30	0.77	0.20	0.82
NNETS (1)	0.35	0.65	0.23	0.85
NNETS (64)	0.32	0.79		
LMR (1)	0.36	0.64	0.22	0.77
LMR (64)	0.31	0.74		
LMR(1,context)	0.34	0.66		

Table 1 : Normalized distances, for voiced and unvoiced cases.

It can be notice that non linear neural networks do not give better results as pure LMR, when the transformations are learned on 64 classes.

3.2. Subjective results

Subjective tests were done on the sentences to evaluate the quality of the transformed speech.

It gave GMM > NNETS ≅ LMR >VQM.

But the obtained speech quality is too poor even for a text to speech synthesizer.

3.3. Work in progress

In order to improve the speech quality, the transformation of the phase of the harmonics of the HNM model is under study.

Acknowledgment : This work was supported by a grant from France Telecom (CNET).

4. REFERENCES

1. Stylianou Y., Laroche J. , Moulines E. « High-quality speech modification based on a harmonic plus noise model », *proc. EUROSPEECH, Madrid, Spain,*1995.

2. Stylianou Y. « Harmonic plus Noise Models for speech, combined with statistical methods for speech and speaker modification », *Ph.D Dissertation, ENST Paris, Jan. 1996.*

3. Zwicker E., Terhardt E. « Analytical expressions for critical-band rate and critical bandwidth as a function of frequency », *J. Acoust. Soc. Am, 68 (5), 1523-1525,1980.*

4. Cappe O., Laroche J., Moulines E. « Regularised estimation of cepstrum enveloppe from discrete frequency points », *IEEE ASSP Workshop on Ap. of sig. proc to audio and acous., Mohouk, October 95*

5. Abe M., Nakamura S., Shikano K., Kuwabara H. « Voice conversion through Vector Quantisation », Int. Conf. on Acoust. Speech Signal Processing, pp 565-568, 1988

6. Stylianou Y., Cappe O., Moulines E. « statistical methods for voice quality transformation », *Proc. Eurospeech, Madrid, Spain, 1995*

7. Valbret H. « Système de conversion de voix pour la synthèse de parole », *Ph.D Dissertation, ENST Paris, 1992*

8. Duda R. O., Hart P. E. « Pattern classification and scene analysis », *John Wiley & sons, Inc. , New York, 1973*

9. Dempster A. P., Laird N. M., Rubin D. B. « Maximum Likehood from incomplete data via the EM algorithm », *J. Roy. Stat. Soc. Ser. B, 39 (1) :1-22 and 22-38, 1977*

10. Stylianou Y., Cappe O., Moulines E. « Continuous probabilistic transform for voice conversion », *submitted to IEEE, Speech and Audio Processing.*

11. Narendranath M., Murthy H. A., Rajendran S., Yegnanarayana B. « *Transformation of formants for voice conversion using artificial neural networks, Speech communications,Eurasip, pp 207-216, vol. 16, N°2,Feb 95.*