A SYSTEM FOR VOICE CONVERSION BASED ON PROBABILISTIC CLASSIFICATION AND A HARMONIC PLUS NOISE MODEL

Yannis Stylianou and Olivier Cappé†

AT&T Labs-Research, SIPS 180 Park Avenue, Florham Park, NJ 07932
email : styliano@research.att.com
† ENST, Signal Department, 46 Rue Barrault 75634 Paris Cedex 13
email : cappe@sig.enst.fr

ABSTRACT

Voice conversion is defined as modifying the speech signal of one speaker (source speaker) so that it sounds as if it had been pronounced by a different speaker (target speaker). This paper describes a system for efficient voice conversion. A novel mapping function is presented which associates the acoustic space of the source speaker with the acoustic space of the target speaker. The proposed system is based on the use of a Gaussian Mixture Model, GMM, to model the acoustic space of a speaker and a pitch synchronous harmonic plus noise representation of the speech signal for prosodic modifications. The mapping function is a continuous parametric function which takes into account the probabilistic classification provided by the mixture model (GMM). Evaluation by objective tests showed that the proposed system was able to reduce the perceptual distance between the source and target speaker by 70%. Formal listening tests also showed that 97% of the converted speech was judged to be spoken from the target speaker while maintaining high speech quality.

1. INTRODUCTION

Voice conversion is a subject of considerable importance. Applications include text-to-speech synthesis based on acoustic unit concatenation, interpreted telephony, and very low rate bit speech coding. In speech synthesis, voice conversion is a simple and efficient way to create the desired variety of voices while avoiding recording of different speakers. Voice conversion is useful in interpreted telephony where it is important, for the naturalness of the conversation, that the characteristics of each speaker's voice are to be maintained. Maintenance of speaker's characteristics is also important in the context of high-quality very low rate speech coding based on text-to-speech synthesis and speech recognition.

The voice conversion problem has recently attracted a lot of research effort. An approach to this problem was the mapping codebook of Abe *et al.* [1]. The mapping codebook method is based on a discrete description of the spectral parameters spaces of both speakers obtained through Vector Quantization (VQ). A variation of this basic scheme is the fuzzy vector quantization approach described in [2]. In [3], a different approach is proposed which is based on the interpolation between the spectra of several speakers to determine the converted spectrum. Other recent works suggest that a possible way to improve the quality of the converted speech consists of modifying only some specific aspects of the spectral envelope, such as the location of the formants [4],[5]. Spectral conversion techniques have been also proposed for speaker/environment adaptation that map speech features of the same speaker between clean and noisy acoustic spaces [6], [7].

This paper describes a new system for voice conversion. Compared to the methods mentioned above, the main contributions of the proposed system are: 1)*Probabilistic classification*: the acoustic space of a speaker is described by a parametric Gaussian mixture model, GMM, which, in contrast with VQ-based methods, provides continuous and smooth classification indexes avoiding unnatural discontinuities. 2)*Mapping Function*: a novel function is proposed to associate the acoustic space of the source speaker with the acoustic space of the target speaker. The proposed function makes use of the complete description of each component of the GMM, considering these components as complete clusters rather than as single vectors, as is the case in VQ approaches. 3)*High-quality prosodic modifications*: a pitch synchronous harmonic plus noise (HNM) representation of the speech signal is used for prosodic modifications. Objective tests and formal listening tests were carried out and the results show that by using the proposed conversion system effective voice conversion is achieved.

The paper is organized as follows. First, the probabilistic classification and the proposed mapping function is presented. Next, the front-end analysis/synthesis system, HNM, is briefly described. This is followed by the implementation of the proposed conversion system. Finally, results from a formal listening test as well as from an objective test are presented to support our conclusions.

2. PROBABILISTIC CLASSIFICATION AND MAPPING FUNCTION

In this section, we assume that the available data consists of two sets of paired p-dimensional spectral vectors $\{\mathbf{x}_t,\ t = 1, \ldots, n\}$ (source) and $\{\mathbf{y}_t,\ t = 1, \ldots, n\}$ (target) with the same length n.

2.1. Gaussian mixture model

The first step consists in fitting a gaussian mixture model to the source vectors $\{\mathbf{x}_t\}$. The modeling of the acoustic space of a speaker by a Gaussian mixture model (GMM) has been illustrated by recent studies [8] to be efficient for text-independent speaker recognition. The GMM assumes that the probability distribution

This work was mainly done when the authors were with ENST-Paris, Signal Department. The work was supported by the Centre National d' Etudes des Télécommunications under contract no. CNET France Telecom 91-7126.

of the observed parameters takes the following parametric form [9]

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (1)$$

where m is the number of the Gaussian components and $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ defined by

$$N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad (2)$$

In (1) the terms α_i are normalized positive scalar weights. A fundamental assumption of the GMM states that the observation vectors $\{\mathbf{x}_t\}$ are independent of one another. The mixture weights $\{\alpha_i\}$ represent the statistical frequency of each class in the observations. The conditional probability that a given observation vector $\mathbf{x}$ belongs to an acoustic class $\mathcal{C}_i$ of the GMM is easily derived from (1) by direct application of Bayes' rule as

$$P(\mathcal{C}_i|\mathbf{x}) = \frac{\alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (3)$$

The parameters of the GMM are estimated from the set of source vectors $\{\mathbf{x}_t\}$ using the Expectation-Maximization (EM) algorithm [10]. An important implementation issue associated with the EM algorithm is its initialization. In this study, the EM algorithm is initialized by use of a standard binary splitting VQ procedure [11]: the weight, mean vector and covariance matrix of each component are initialized independently using the clusters obtained by VQ of the source vectors $\{\mathbf{x}_t\}$.

2.2. Mapping function

In the limit-case where the GMM is reduced to a single class and assuming that the source vectors $\mathbf{x}_t$ follow a Gaussian distribution $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that the source and target vectors are jointly Gaussian, the minimum mean square error (mmse) estimate of the target vector after observing $\mathbf{x}$ is given by [12](p. 325)

$$E[\mathbf{y}|\mathbf{x} = \mathbf{x}_t] = \boldsymbol{\nu} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}), \quad (4)$$

where $E[]$ denotes expectation, and $\boldsymbol{\nu}$ and $\boldsymbol{\Gamma}$ are respectively the mean target vector

$$\boldsymbol{\nu} = E[\mathbf{y}],$$

and the cross-covariance matrix of the source and target vectors

$$\boldsymbol{\Gamma} = E[(\mathbf{y}-\boldsymbol{\nu})(\mathbf{x}-\boldsymbol{\mu})^T],$$

where the superscript T denotes transposition. It was decided to extend this result to the GMM by weighting terms that are analogous to the Gaussian conditional expectation. It seems logical to choose as weighting terms the conditional probabilities that the vector $\mathbf{x}_t$ belongs to the different classes $\mathcal{C}_i$ (Eq. (3)). Thus, the proposing mapping function has the following parametric form:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i|\mathbf{x}_t)\left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right] \quad (5)$$

The parameters of the mapping function are computed by least squares optimization on the learning data so as to minimize the total squared conversion error

$$\epsilon = \sum_{t=1}^{n} ||\mathbf{y}_t - \mathcal{F}(\mathbf{x}_t)||^2 \quad (6)$$

In this paper, the covariance matrices of the GMM $\boldsymbol{\Sigma}_i$ and the conversion matrices $\boldsymbol{\Gamma}_i$ are full matrices. The mapping function will be referred to as *full mapping*.

3. THE HARMONIC PLUS NOISE MODEL, HNM

The voice conversion system is based on the use of the Harmonic + Noise Model (HNM) which allows high-quality modifications of speech signals [13]. HNM performs a pitch-synchronous harmonic plus noise decomposition of the speech signal. For voiced sounds, the speech spectrum is divided into a low and a high band delimited by the so-called maximum voiced frequency. The low band of the spectrum (below the maximum voiced frequency) is represented solely by harmonically related sine waves. The upper band is modeled as a noise component modulated by a time-domain amplitude envelope. Due to the pitch-synchronous scheme of HNM, time-scale and pitch-scale modifications are quite straightforward[13].

4. IMPLEMENTATION OF THE CONVERSION SYSTEM

HNM decomposes speech into a harmonic and a noise part so the conversion procedure could be different for each part. There are several reasons for this decision. Spectral envelopes associated with the noise part exhibit large variations and the corresponding GMM components are characterized by large variances and significant overlap. Moreover, the contribution of the noise part to the individuality of the speaker was found to be far less important than that of the harmonic part. As a consequence, the conversion methodology presented in the previous section is only applied to the transformation of the harmonic part of the signal. For the conversion of the noise part two corrective filters are defined; one for voiced and another for unvoiced frames.

4.1. Conversion of the noise part

The conversion of the noise part is simply achieved by the use of two different correction filters (one for voiced frames and one for unvoiced frames). These correction filters, implemented as 6*th* order all-pole filters, model the difference between the average noise spectra of the source and target speaker.

4.2. Conversion of the harmonic part

For the conversion of the harmonic part an efficient parameterization of the spectral envelope is desirable. Because the harmonic amplitudes will be computed from the converted spectral envelopes, it is desirable to use a spectral representation method that leads to an envelope that passes through the measured harmonic amplitudes. Such a representation has already been developed in [14] where a regularization technique has been proposed to achieve a well-behaved spectral envelope using discrete cepstrum coefficients. For a better fit in low frequencies, the harmonic frequencies are converted to a Bark frequency scale. The cepstral

parameters obtained are similar to the usual Mel-Frequency Cepstrum Coefficients (MFCC) except for the fact that they are obtained from the minimization of a discrete set of frequency constraints.

The learning procedure for the conversion of the harmonic part is depicted in Fig. 1. Note that for the training of the conversion function, the source and target signals are analyzed with a fixed 10ms frame rate in order to allow time-alignment by the *Dynamic Time Warping*, DTW, algorithm. The optimization of the conversion function (rightmost block in Fig. 1) makes use of the time-aligned spectral envelopes $\{\mathbf{x}_t\}$ (source) and $\{\mathbf{y}_t\}$ (target) as well as the parameters of the GMM as estimated by the EM algorithm.

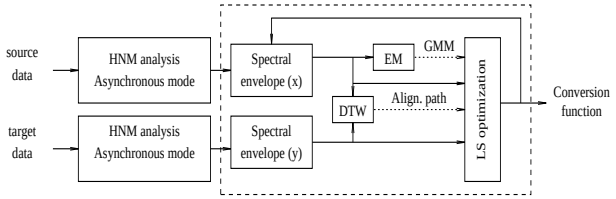

Figure 1: Block diagram of the learning procedure.

4.3. The conversion system

Once the spectral conversion function has been estimated, the voice transformation is performed as indicated in Fig.2. Note that for voice transformation, the HNM analysis is performed pitch-synchronously because this mode enables higher quality time-scale and pitch-scale modifications [13]. The noise part is modified with two different fixed filters (so called "corrective filters") depending on whether the frame is voiced or not. In the present

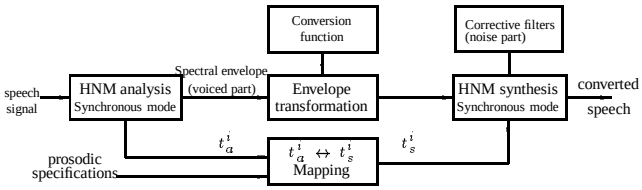

Figure 2: Block diagram of the voice conversion system. t_a^i: analysis time-instants, t_s^i: synthesis time-instants.

system, we do not consider the problem of matching the prosodic characteristics of both speakers. As a consequence, the prosodic modifications performed are merely intended to match the average fundamental frequency and articulation rhythm of both speakers.

5. RESULTS AND DISCUSSION

The proposed conversion system has been tested on the conversion task between two male speakers. The speech databases have been provided by the Centre National d' Etudes des Télécommunications (CNET), which cover all the diphones of the French language. The sampling frequency was $16kHz$. Approximately 20000 voiced vectors have been obtained per speaker (3.5 minutes of speech). The frame size for the asynchronous HNM analysis was 10 msec and the cepstrum order was 20. In the present study, the first cepstrum coefficient c_0 was omitted as a form of energy normalization. In practice, it was found that it is not advisable to include c_0 in the training parameters because it biases the classification achieved by the GMM. The spectral parameters are thus 20-dimensional vectors which contain the discrete MFCC coefficients $c_1, c_2, \ldots, c_p$. For a simplification of HNM, the maximum voiced frequency was fixed at a constant value of $4kHz$. An independent corpus of about 1.5 min (with more than one minute of voiced speech) was used to evaluate the performance of the proposed method.

5.1. Objective test

In this section results from an objective test are presented and the *Full mapping* function is compared with the VQ-mapping approach [1]. For the objective test, the rms log-spectral distortion is computed using the warped frequency scale as

$$d_{\text{rms}}{}^2 = 2\sum_{k=1}^{p}[c_1(k) - c_2(k)]^2 \qquad (7)$$

Fig. 3 compares VQ-mapping (Fig. 3-(a)) and *full mapping* (Fig. 3-(b)) based on the frame rms log spectral distortion measured for one second of natural speech. Note that $0dB$ value refers to the *initial average* distortion between the source and target envelopes and their frame based distortion is represented by solid line. Full mapping makes it possible to achieve an average distortion reduction of $5dB$ while the average distortion reduction for VQ-mapping is $4dB$. It is also worth noting that the reduction of the log spectral distortion by VQ-mapping is very non uniform (see Fig. 3-(a)) in contrast with *full mapping* where the reduction is almost always greater than $2dB$ (see Fig. 3-(b)). The VQ-mapping system has also been used with a codebook of 512 centroids. However, the average distortion reduction was slightly higher than $4dB$.

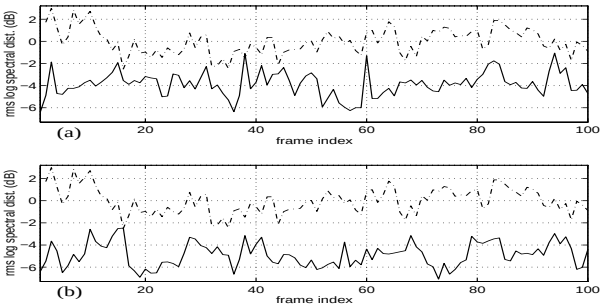

Figure 3: Normalized warped rms log-spectral distortion in dB for 100 consecutive frames of voiced speech. **(a)**: Conversion by VQ-mapping (128 centroids).**(b)** Full mapping (128 GMM). Dash-dot line: distortion between source and target envelopes; Solid line: distortion between converted and target envelopes.

5.2. Formal listening test

The proposed conversion system has also been assessed during formal listening tests on sentences uttered by the source and the target speakers. To evaluate only the spectral conversion aspect the prosody of the source speaker has been altered (using HNM) to match as closely as possible the prosody of the target speaker. The evaluation has been carried out on approximately 12 seconds of continuously uttered sentences using a 16 GMM and a 64 GMM. Two listening tests have been designed; XAB test, and opinion test. Twenty listeners have been participated in each of these experiments.

5.3. XAB test

In the XAB test, A and B were either the target or the source speaker and X was either the prosody-only modified speech, the 16 GMM, or the 64 GMM converted speech. Subjects were asked to select either A or B as being most similar to X. Table 1 summarizes the results from this test giving the percentage of correct answers; the converted/modified speaker is recognized as the target speaker. For the first three columns of the Table 1 speakers A and B uttered the same sentence which was different from the sentence uttered by X, while for the last column of the table all speakers uttered the same sentence. It is worth noting from Table 1 the difference in score between prosody-only modification and 16 GMM. The score continues to increase as the number of GMM parameters increases and the score becomes higher when X, A and B utter the same sentence (easier task for the listeners).

	PO	16 GMM	64 GMM	64 GMM(2)
Correct answers	18%	83%	88%	97%

Table 1: Results from the XAB test. PO stands for *prosody only* modification.

5.4. Opinion test

To evaluate the overall performance of the proposed method an opinion test was designed. Pairs of speech signals including all possible combinations of original speaker, target speaker, "prosodic modified" speaker and converted speaker using 16 and 64 GMM components were presented to the listeners. Different sentences were used to make these pairs. Listeners were asked to rate the similarity of each pair of speakers on a scale with ten values between 0 for "identical" and 9 for "very different". Fig.4 presents the results from this test. The symbols used in this figure stand for the distances: "TT", target-target, "SS", source-source, "M2", converted speaker using 64 GMM components-target, "M1", converted speaker using 16 GMM components-target, "PT", prosodic modified speaker- target and "ST" source-target. For each of the distances the median value is given (noted by "x") as well as the variation of the decisions using as estimator the mean absolute deviation rather than the standard deviation. This figure clearly shows the efficiency of the proposed method and confirms the results of the first test.

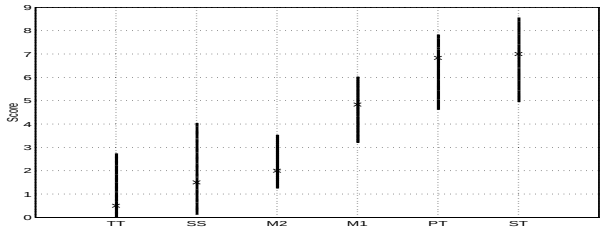

Figure 4: Results form the opinion test.

6. CONCLUSIONS

This paper has presented a voice conversion system based on probabilistic classification and the Harmonic plus Noise Model, HNM. The proposed system has been evaluated by objective and formal listening tests. The results show that the proposed mapping function which takes into account the probabilistic classification provided by the mixture model (GMM) is more robust and efficient than methods based on VQ. The proposed system is able to reduce the perceptual distance between the source and target speaker by 70%. Formal listening tests also show that 97% of the converted speech is judged to be spoken from the target speaker.

7. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 655–658, 1988.

[2] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, pp. 165–173, Feb. 1995.

[3] N. Iwahashi and Y. Sagisaka, "Speech spectrum transformation by speaker interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[4] H. Mizuno and M. Abe, "Voice conversion based on piecewise linear conversion rule of formant frequency and spectrum tilt," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[5] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA techniques," *Speech Communication*, vol. 11, pp. 175–187, Jun 1992.

[6] C. Mokbel and G. Chollet, "Speech recognition in adverse environments: speech enhancement and spectral transformations," *Proc. IEEE ICASSP-91*, pp. 925–928, May 1991.

[7] L. Neumeyer and M. Weintraub, "Probabilistic optimum filtering for robust speech recognition.," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Adelaide, Australia), pp. 417–420, 1994.

[8] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 293–296, 1990.

[9] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*. New York: John Wiley & Sons, Inc., 1973.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B (methodological)*, vol. 39, no. 1, pp. 1–22 et 22–38 (discussion), 1977.

[11] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communication*, vol. 28, no. 1, 1980.

[12] S. M. Kay, *Fundamentals of statistical signal processing: Estimation theory*. PH signal processing series, Prentice-Hall, 1993.

[13] Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, 1995.

[14] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3(4), pp. 100–102, April 1996.