Statistical methods for voice quality transformation

Yannis Stylianou, Olivier Cappé, Eric Moulines
Ecole Nationale Supérieure des Télécommunications
Département Signal / CNRS-URA 820
46 Rue Barrault 75634 Paris Cedex 13, FRANCE

Abstract

This paper presents a new method for the statistical learning of the correspondence between spectral parameters measured from two different speakers uttering the same text. This method is based on the use of a gaussian mixture model of the speaker's spectral parameters. It is shown to be more efficient and robust than previously known techniques based on the use of vector quantization. The results obtained on large speech database demonstrate effective high-quality transformations of the voice characteristics.

1 Introduction

Text-independent speaker recognition aims at extracting from the speech waves the information relative to the speaker individuality without explicit references to what is uttered. It proceeds by extracting from the speech samples the information which is characteristic of the way a particular speaker utters individual phonemes, articulates string of phonemes, and then produces words and sentences. The information which is pertinent for such purpose are of course related to the physiological and the behavioral characteristics of the speaker. These characteristics exist both in the short-term spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech.

Voice conversion consists exactly in the reverse operation: starting from the speech signal uttered by a speaker, it aims at transforming the characteristics of the speech signal, in such a way that a human listener could believe that the transformed speech is produced by another (target) speaker. In other words, the machine disguises the voice of the speaker to mislead the listener. The potential applications of such techniques are numerous; at the first place, voice conversion would be an essential component of future text-to-speech systems based on concatenation of acoustical units. It is now more or less admitted that high-quality synthesis system will use huge amounts of speech data. These data are difficult to collect and uneasy to handle (segment, store, access in real time). Developing a new voice will be an extremely expensive process. Voice conversion could be an alternative. Other applications include: voice individuality disguise for secure communications, voice individuality restoral for interpreting telephony, and so on... [1]

This work was supported by the Centre National d'Etudes des Télécommunications (CNET).

2 Main features of the method

In this contribution, we concentrate on the transformation at the segmental level. Our aim is to learn a conversion function that maps the acoustic space of a source speaker to the acoustic space of the target speaker. Since it is likely that the conversion should depend on the 'class' (in a broad sense), an initial clustering is performed. Specific transformations can then be learned for each class.

This kind of approach to the voice conversion problem was first pioneered by Abe et al. using the mapping codebook method [2]. In this approach, the clustering is achieved through Vector Quantization (VQ) of the acoustic spaces of the two speakers. The main shortcoming of this method is the fact that the acoustic space of the converted signal is limited to a discrete set of envelopes. As of today, none of the subsequent developements of the original mapping codebook method has been entirely successful in overcoming this drawback [3]. Compared to these methods, the main originalities of our system are the following:

- **Soft clustering**: During the learning phase, the spectral data are modelled by a mixture of Gaussian densities. In contrast with VQ, the mixture model allows to obtain continuous 'smooth' classification indexes (the classification is probabilistic and is a continuous function of the spectral parameters). This characteristic improves the synthesis quality, avoiding the artefacts generated by unnatural discontinuities in the transformation which typically occurs in the VQ model, when a vector jumps from one class to the other.
- **Incremental learning**: In oder to minimize the influence of local errors in the time alignment path between the two speakers, it is proposed to learn the conversion function incrementally. In a first iteration, the data from the source speaker

and the target speaker are aligned by a standard Dynamic Time Warping (DTW) algorithm. This alignment is used to obtain the initial parameters of the conversion function. In subsequent iterations, the DTW procedure is applied to the converted data and the target data in order to refine the alignment path.

- **Continuous transform**: The proposed parametric conversion function makes use of the probabilistic classification achieved by the mixture model as well as the characteristics of each class (mean vector and covariance matrix). Each class is thus considered as a complete cluster rather than as a single vector as is the case in VQ-based methods. This conversion function drastically reduces the unwanted spectral distortions observed with most current voice conversion techniques.

3 Training of the conversion function

This section is concerned with the learning of the spectral conversion function from the time-aligned spectral data corresponding to both speakers. We consider 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.

3.1 Gaussian mixture model

The first step consists in fitting a gaussian mixture model to the source vectors $\{\mathbf{x}_t\}$. The gaussian mixture model implies two fundamental assumptions:

1. The probability distribution of the observed parameters can be written as [4]

$$p(\mathbf{x}_t) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}_t; \mu_i, \mathbf{\Sigma}_i), \qquad (1)$$

   where $N(\mathbf{x}; \mu, \mathbf{\Sigma})$ denotes the p-dimensional normal distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$ and α_i are normalized positive scalar weights ($\sum_{i=1}^{m} \alpha_i = 1$ and $\alpha_i \geq 0$).

2. The observation vectors $\mathbf{x}_t$ are independent from one another.

The gaussian mixture model is used for its ability to model the acoustic space of a speaker as a combination of several components [5]. Each component, or class, $\mathcal{C}_i$ ($i = 1,\ldots,m$) is characterized its center (mean vector μ_i) as well as by a characteristic spreading around the center of the class (covariance matrix $\mathbf{\Sigma}_i$). The mixture weights α_i represent the statistical frequency of each class in the observations.

The 'soft clustering' mentioned above is achieved by computation of the conditional probabilities $P(\mathcal{C}_i|\mathbf{x}_t)$ that a given observation vector $\mathbf{x}_t$ belongs to each one the acoustic classes $\mathcal{C}_i$. A straightforward application of Baye's rule yields [4]

$$P(\mathcal{C}_i|\mathbf{x}_t) = \frac{\alpha_i N(\mathbf{x}_t; \mu_i, \mathbf{\Sigma}_i)}{\sum_{j=1}^{m} \alpha_j N(\mathbf{x}_t; \mu_j, \mathbf{\Sigma}_j)} \qquad (2)$$

In the present work, the parameters of the gaussian mixture model (α_i, μ_i, $\mathbf{\Sigma}_i$) are estimated using the classic Expectation-Maximization (EM) algorithm of [6]. The EM algorithm provides a general framework for iteratively finding local maximum-likelihood estimates of the unknown parameters. The EM methodology as found numerous applications which include the cases of mixture densities and hidden Markov models. The details of the EM algorithm in the case of gaussian mixture models can be found in [5].

3.2 Conversion function

The following parametric form is assumed for the conversion function $\mathcal{F}()$:

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{m} P(\mathcal{C}_i|\mathbf{x}_t) \left[\nu_i + \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{-1}(\mathbf{x}_t - \mu_i)\right] \qquad (3)$$

The unknown parameters of the conversion function $\mathcal{F}()$ are the p-dimensional vectors ν_i and the $p \times p$ matrices $\mathbf{\Gamma}_i$, for $i = 1,\ldots,m$ (where m is the number of mixture components). In (3), the term between brackets is easily recognized as the conditional expectation of $\mathbf{y}_t$ given the observed value of $\mathbf{x}_t$ in the jointly gaussian case. If the gaussian components of the mixture could be separated, the proposed conversion function would thus result in a modification of the posterior mean and covariance of each component. In practice, the components of the mixture cannot be separated since the conditional probabilities $P(\mathcal{C}_i|\mathbf{x}_t)$ are not restricted to be equal to either 0 or 1. It was thus decided to weight the gaussian conditional expectation term mentioned above by the conditional probability that the observed vector $\mathbf{x}_t$ belongs to the acoustic classes $\mathcal{C}_i$.

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 between the converted and the target data

$$\epsilon = \sum_{t=1}^{n} ||\mathbf{y}_t - \mathcal{F}(\mathbf{y}_t)||^2 \qquad (4)$$

Note that since we use a cepstral parametrization for the spectral vectors, ϵ can also be interpreted as the total quadratic log-spectral distortion between the converted and the target spectra.

It can be shown that the optimal values of the parameters of the conversion function can be computed by resolving the following set of normal equations:

$$\left(\begin{bmatrix} \mathbf{P}^T \\ \cdots \\ \mathbf{D}_x^T \end{bmatrix} \cdot \begin{bmatrix} \mathbf{P} & \vdots & \mathbf{D}_x \end{bmatrix} \right) \cdot \begin{bmatrix} \nu \\ \cdots \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T \\ \cdots \\ \mathbf{D}_x^T \end{bmatrix} \cdot \mathbf{y} \tag{5}$$

Where $\mathbf{y}$ is a $n \times p$ matrix that contains the target spectral vectors ordered the following way

$$\mathbf{y} = \left[\mathbf{y}_1 \vdots \cdots \vdots \mathbf{y}_n \right]^T$$

$\mathbf{P}$ is a $n \times m$ matrix that features the conditional probablities

$$\mathbf{P} = \begin{bmatrix} P(\mathcal{C}_1|\mathbf{x}_1) & \ldots & P(\mathcal{C}_m|\mathbf{x}_1) \\ P(\mathcal{C}_1|\mathbf{x}_2) & \ldots & P(\mathcal{C}_m|\mathbf{x}_2) \\ \vdots & & \vdots \\ P(\mathcal{C}_1|\mathbf{x}_n) & \ldots & P(\mathcal{C}_m|\mathbf{x}_n) \end{bmatrix}$$

$\mathbf{D}_x$ is a $n \times pm$ matrix that depends on the conditional probabilities, the source vectors and the parameters of the GMM

$$\mathbf{D}_x = \begin{bmatrix} P(\mathcal{C}_1|\mathbf{x}_1)(\mathbf{x}_1 - \mu_1)^T \mathbf{\Sigma}_1^{-1} & \cdots & P(\mathcal{C}_m|\mathbf{x}_1)(\mathbf{x}_1 - \mu_m)^T \mathbf{\Sigma}_m^{-1} \\ P(\mathcal{C}_1|\mathbf{x}_2)(\mathbf{x}_2 - \mu_1)^T \mathbf{\Sigma}_1^{-1} & \cdots & P(\mathcal{C}_m|\mathbf{x}_2)(\mathbf{x}_2 - \mu_m)^T \mathbf{\Sigma}_m^{-1} \\ \vdots & & \vdots \\ P(\mathcal{C}_1|\mathbf{x}_n)(\mathbf{x}_n - \mu_1)^T \mathbf{\Sigma}_1^{-1} & \cdots & P(\mathcal{C}_m|\mathbf{x}_n)(\mathbf{x}_n - \mu_m)^T \mathbf{\Sigma}_m^{-1} \end{bmatrix}$$

And the two matrices

$$\nu = \left[\nu_1 \vdots \nu_2 \vdots \cdots \vdots \nu_m \right]^T \quad (m \times p)$$

and

$$\mathbf{\Gamma} = \left[\mathbf{\Gamma}_1 \vdots \mathbf{\Gamma}_2 \vdots \cdots \vdots \mathbf{\Gamma}_m \right]^T \quad ((m \times p) \times p)$$

contains the parameters of the conversion function.

As is usual with least-squares problems, the matrix that need to be inverted (between parentheses in (5)) is symmetric and positive definite so that the normal equations can advantageously be solved using the Cholesky decomposition. The attention of the reader is drawn however on the fact that the dimension of this matrix ($(m + mp) \times (m + mp)$) becomes pretty large as the number of components of the mixture m increases. Simplified versions of (5) can be obtained in special cases such as when the matrices $\mathbf{\Sigma}_i$ and $\mathbf{\Gamma}_i$ are diagonal.

4 Voice conversion system

This conversion function was tested using the Harmonic + Noise Model (HNM) which allows high-quality modifications of speech signals [7]. The general principle of HNM consist in decomposing the speech signal as the sum of a purely harmonic signal and of a modulated noise [7].

The present work uses a simplified version of HNM in which the harmonic part of the signal is supposed to cover the frequency range 0-4kHz (for voiced frames) and the analysis is performed at a constant frame rate of 10 ms. The method presented in the previous section is used in order to modify the harmonic part of the signal. The noise part is only roughly modified by use of two separate time invariant filters (one for voiced frames and the other for unvoiced frames). This simple transformation scheme was found to be sufficient since the exact frequency content of the noise part does not seem to contribute significantly to speaker individuality.

The amplitudes of the harmonics that constitute the voiced part of speech are determined by a time-domain weighted least-squares technique [7]. A continuous model of the spectral envelope that connects the obtained harmonics is then estimated using the discrete regularized cepstrum method [8]. For this purpose, the frequencies of the harmonics are first converted to a non-linear Bark frequency scale. The spectral envelope is thus described by parameters that are analogous to the standard Mel-Frequency Cepstrum Coefficients (MFCC). The discrete cepstrum method has the advantage of providing a very good match of the spectral envelope with the amplitudes of the harmonics for reasonable values of the order of the cepstrum [8]. In this study an order of the cepstrum of p=20 was used and the first cepstral coefficients $c(0)$ was omitted from the training parameters.

5 Results

The database used in order to train the conversion function consists of the diphones of the french language uttered in context by two different male speakers. The time alignment between the source and target signals was performed on each diphone separately using a standard DTW technique and discarding the unvoiced potions of the signal. The remaining time-aligned data consisted of approximately 20 000 spectral vectors which corresponds to more than 3 mn of speech.

Figure 1 presents the relative spectral distortion (average quadratic spectral distortion of (4) normalized by the initial distortion between the two speakers) measured on the training data sets. The proposed method, featured on part (a) of figure 1, is compared with a VQ-based system on part (b). What

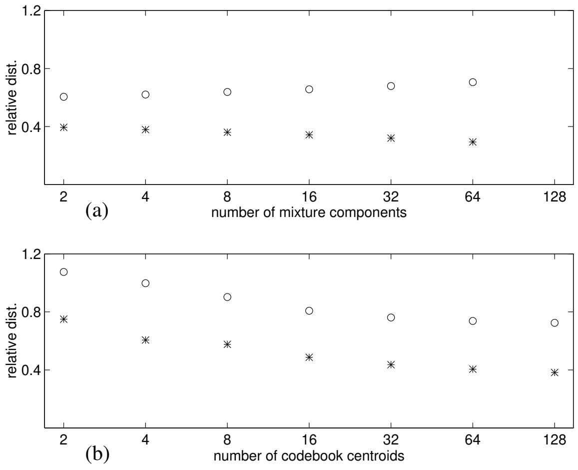

Figure 1: Relative spectral distortion between the converted and target data (stars) and the converted and source data (circles) for different sizes of the underlying model. (a) conversion with the proposed conversion function. (b) conversion with the VQ-based system.

is striking is the way the proposed conversion scheme steadily approaches the target data as the number of mixture component increases: on part (a) of figure 1, the distortion between the converted and target vectors (stars) decreases while the distortion between the converted and source vectors (circle) gradually increases. This is in total contrast with what is observed on part (b) of figure 1 for the VQ-based system where the unwanted spectral distortion due to the discretization of the speaker space causes the converted vectors to be very different from both the target *and* source vectors.

Moreover the reduction of the spectral distortion (between the converted and target vectors) achieved by the proposed system is by far superior to that of the VQ system: the VQ-based system used with a codebook of 512 vectors still produces a distortion that is 17% higher than that obtained with the proposed system when using a mixture of 64 components.

The iterative refinement of the time alignment path (the above mentioned incremental learning) was found to produces only a marginal improvement of the results: the conversion distortion obtained after the second iteration of the whole learning process is generally around 5% lower than that obtained for the first iteration. This is believed to be due to the fact that the DTW is only performed on very short segments of speech in our case.

6 Conclusion

Informal listening tests indicate that the proposed system produces speech signals which are free of artefacts (burbles and other oddities) associated with the VQ technique. The obtained conversion effect is impressive and the general quality of the transformed speech signal is satisfying although a muffling effect is perceptible when the number of mixture components is too small. Future developments of the conversion method introduced in this paper will include its evaluation with other sets of speech parameters (such as formant parameters) which may be best suited for voice conversion.

Acknowledgments

The author are indebted to their colleague Jean Laroche for his insightful suggestions and to Olivier Boeffard of the CNET Lannion for his help with the speech database.

References

[1] E. Moulines and Y. Sagisaka, editors. *Voice conversion: state of the art and perspectives (special issue of Speech Communication)*. Elsevier, 16(2), Feb. 1995.

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

[3] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: Control and conversion. *Speech Commun.*, 16(2):165–173, February 1995.

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

[5] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1):72–83, January 1995.

[6] 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)*, 39(1):1–22 et 22–38 (discussion), 1977.

[7] Y. Stylianou, J. Laroche, and E. Moulines. High-quality speech modification based on a harmonic + noise model. In *EUROSPEECH*, Madrid, September 1995.

[8] O. Cappé, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. In *IEEE ASSP Workshop on App. of Sig. Proc. to Audio and Acoust.*, Mohonk, October 1995.