

Including Dynamic and Phonetic Information in Voice Conversion Systems

Helena Duxans*, Antonio Bonafonte*, Alexander Kain**, Jan van Santen**

*Department of Signal Theory and Communication, TALP Research Center
Technical University of Catalonia (UPC), Barcelona, Spain

**Center for Spoken Language Understanding, OGI School of Science Engineering
Oregon Health Science University, Portland, Oregon, USA

Abstract

Voice Conversion (VC) systems modify a speaker voice (*source speaker*) to be perceived as if another speaker (*target speaker*) had uttered it. Previous published VC approaches using Gaussian Mixture Models [1] performs the conversion in a frame-by-frame basis using only spectral information. In this paper, two new approaches are studied in order to extend the GMM-based VC systems. First, dynamic information is used to build the speaker acoustic model. So, the transformation is carried out according to sequences of frames. Then, phonetic information is introduced in the training of the VC system. Objective and perceptual results compare the performance of the proposed systems.

1. Introduction

Voice Conversion (VC) systems modify a speaker voice (*source speaker*) to be perceived as if another speaker (*target speaker*) had uttered it. Applications of VC systems can be found in several fields, such as TTS (text-to-speech systems) customization, automatic translation keeping speaker's voice individuality [2], education tools for foreign language learning [3], medical aids to improve the voice of people with speech impairments [4] and in the entertainment field [5].

The goal of this paper is to build a VC system as a post-processing block for a TTS, in order not to have to produce and store several speech databases, one for each speaker. So, the amount of training data is not a severe problem. Only high quality is required.

Previous approaches to VC have assumed the glottal flow + vocal tract model for speech production, learning a mapping function between source and target speakers for vocal tract features and predicting the converted residual signal from the vocal tract [1]. An already proposed mapping function is based on GMM as a model for joint source and target acoustic space. To estimate the GMM aligned source-target feature vectors are used. GMM-based systems work frame by frame, using only spectral information to learn the mapping and transform voices. In this paper the following topics are studied, introducing two new approaches to vocal tract conversion:

- The effects of including dynamic characteristics in the acoustic model used to build local vocal tract mapping functions. For this reason, GMM-based systems are extended to HMM-based systems, which can model dynamic characteristics.
- The effects of including phonetic information in the learning of the mapping function and in the transformation. The learning will be carried out in an unsupervised way by CART decision trees.

The outline of this paper is as follows. In section 2 a new approach based on HMM is introduced. Then, in section 3 it is explained how to apply a CART decision tree to build a VC system. Finally, in section 4 the results are discussed and conclusions can be found in section 5.

2. HMM-based voice conversion

Previous GMM-based systems work in a frame-by-frame basis. It means that to convert one frame the information about past and future frames is not relevant. This is a simplification of the real speech production mechanism. Our propose is to include dynamic information in the voice conversion task. HMM are well-known models which can capture the dynamics of the training data using states. A HMM can model the probability distribution of any feature vector, according to its actual state, and also it can model the dynamics of sequences of vectors with transition probabilities between states.

The model parameters $(a_{ij}, b_i(\mathbf{x}), \pi_i)$, where a_{ij} indicates the transition probability matrix, $b_i(\mathbf{x})$ the emission probability function of the i th state and π_i the initial probability of the i th state, can be estimated using the Baum-Welch algorithm.

In this paper, all the studied HMM are ergodic, i.e. all the states are connected, and the emission probability function for each state is a Gaussian function. LSF parameters have been used as a vocal tract features.

The block diagram of a HMM-based VC system is presented in figure 1. In the training step, a HMM is estimated from training data, and then a conversion function is calculated for each state. In the transforming step, HMM is used twice. First, source data is segmented

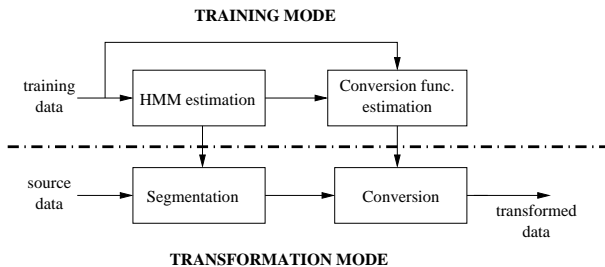


Figure 1: HMM-based VC system block diagram.

according to the HMM states. Then, each frame is transformed applying the state-dependent conversion function.

2.1. Source HMM-based system

The basic idea of this system is to model the dynamics of the source speaker with an ergodic HMM. The transition probabilities of this model will be used as dynamic characteristics in the conversion. This system is similar to the one proposed in [6], but using continuous transformation functions in order to avoid spectral jumps in the converted features that, as it was reported, degrades the quality of the transformed speech.

The steps for training the conversion function are the following. First, a source HMM is estimated from source data. Then, using the estimated HMM, source training vector sequences are segmented according to the optimal state path (using Viterbi search). All the vectors, with their target alignments, are collected for each state, and N (number of states) joint Gaussian functions are estimated. Finally, regressing the function for each state we have:

$$F_s(x) = \mu_s^y + \Sigma_s^{yx} \Sigma_s^{xx}{}^{-1} (x - \mu_s^x) \quad (1)$$

as a conversion function, where s indicates the state, \mathbf{x} and \mathbf{y} aligned source and target vectors, and μ and Σ means and covariance matrices. To transform a new sequence, we need to segment it according to the source HMM. Then, the conversion function of each state is applied to each state parameters.

2.2. Joint HMM-based system

As it has been previously done with GMM systems, we introduce joint information in order to allocate the distribution functions more judiciously, and also to use both source and target dynamic information. So, using aligned source-target features vectors a joint HMM is estimated. Like in joint GMM, there is no need of an extra step to calculate the mapping function for each state. Since there is a joint Gaussian per state, we can calculate the regression function straightforward.

Once the joint HMM is estimated, there are two different ways of transforming new vectors. In method A, the new sequence can be segmented according to the op-

timal state path s^* :

$$s^* = \arg \max_s p(\mathbf{x}, \mathbf{s} / \lambda) \quad (2)$$

$$s^* = \arg \max_s p(\mathbf{x} / \mathbf{s}, \lambda) p(\mathbf{s} / \lambda) \quad (3)$$

where $\lambda = (a_{ij}, b_i(\mathbf{x}), \pi_i)$ $i = 1 \dots N$, for a HMM with N states. Then, as in source HMM, each vector is transformed according to its segmentation state. Note that now transition probabilities take into account not only source speaker, but also target speaker information.

Another way of transforming a new sequence, method B, is to include the regression in the search of the optimal path.

$$s^* = \arg \max_s p(\mathbf{y}, \mathbf{x}, \mathbf{s} / \lambda) \quad (4)$$

$$\approx \arg \max_s p(\tilde{\mathbf{y}}, \mathbf{x}, \mathbf{s} / \lambda) \quad (5)$$

$$\approx \arg \max_s p(\tilde{\mathbf{y}} / \mathbf{x}, \mathbf{s}, \lambda) p(\mathbf{x} / \mathbf{s}, \lambda) p(\mathbf{s} / \lambda) \quad (6)$$

where $\tilde{\mathbf{y}}$ indicates the transformed frame. We have approximated the solution using the transformed frame instead of the target frame, which is unknown. Although a priori the transformed frame is also unknown, the decomposition 6 allows to compute it applying the regression function of the state s to the source frame. The equation 6 can be easily solve using dynamic programming.

3. CART-based voice conversion

Previous GMM-based systems work with spectral features to estimate the conversion function and to transform new source spectral vectors. In this section, the inclusion of phonetic information for each frame, such as the phone, a vowel/consonant flag, point of articulation, manner and voicing, is studied. Note that all this information is available in a TTS, so its inclusion in the conversion system is straightforward.

To estimate the mapping function a CART decision tree has been used. It is based in the idea that the acoustic space of both speakers is organized in acoustic classes, and a conversion function can be estimated for each class. Using GMM or HMM, we only have spectral information to identify the classes. But using decision trees we can also use phonetic information. The tree extracts, at each splitting step, overlapping regions of the acoustic space that can be represented by only one acoustic class, modeled by a joint probability function.

The procedure to grow the tree is as follows. A GMM-based VC system with one component is estimated from a training data set for the parent node (the root node in the first iteration), and an error index for all the elements of a validation data set is calculated. The error index used is:

$$E = \frac{1}{M} \sum_{m=0}^{M-1} \frac{D(\tilde{\mathbf{y}}_m, \mathbf{y}_m)}{D(\mathbf{x}_m, \mathbf{y}_m)} \quad (7)$$

where x_m , y_m and \tilde{y}_m are the source, target and converted m th frames respectively, and $D(\cdot)$ indicates an Inverse Harmonic Mean Distance [7] calculated as:

$$D(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{p=1}^P c(p)(x(p) - y(p))^2} \quad (8)$$

$$c(p) = \frac{1}{w(p) - w(p-1)} + \frac{1}{w(p+1) - w(p)} \quad (9)$$

with $w(0) = 0$, $w(P+1) = \pi$ and $w(p) = x(p)$ or $w(p) = y(p)$ so that $c(p)$ is maximized (p is the vector dimension). The features used are LSF. Using this distance measurement we weight more the mismatch in spectral picks than the mismatch in spectral valleys.

Then, all the possible question of the form ‘‘phonetic property n=value’’ are evaluated, and two child nodes are populated for each question. For each child node, a GMM with one component is estimated, and the error index for the vectors of the validation set corresponding to this child node is calculated. The decision to grow the tree is:

$$E_{parent} - \frac{(E_{child1} * elem_{child1}) + (E_{child2} * elem_{child2})}{(elem_{child1} + elem_{child2})} \quad (10)$$

where $elem_{child}$ indicates the number of spectral vectors of the validation set belonging to the child node. Only when this decision rule is positive and the number of training frames is higher than 25 this node is a candidate to be split with this question. At each iteration, the node with the decision rule with higher value for any question is split according to that question. The tree is grown until there is no node candidate to be split.

To transform new source vectors, they are classified into leafs according to their phonetic features by the decision tree. Then, each vector is converted according to the GMM-based system belonging to its leaf.

Taking into account that phonetic information may be not enough to split all the acoustic classes, an alternative is to model the final leafs with more than one Gaussian. So, once the tree has been grown, for each final leaf a GMM with a number of components from 1 to 5 is estimated, until the error in the validation set increases.

4. Experiments

The corpus used for the experiments was built to generate a Spanish unit selection TTS system. Speech and laringograph signals were recorded in an acoustically isolated room. A sample frequency of 32kHz and 16 bits per sample were used. For this study, signals were decimated to 8kHz. The corpus has been segmented (manually supervised) into phones. Two speakers, one male and one female, read the same corpus.

The frame alignment used is lineal inside each phoneme. Only phonetic transcription matching sentences without pauses are used.

4.1. Objective Tests

The performance index used for test is:

$$P = 1 - \frac{1}{M} \sum_{m=0}^{M-1} \frac{D(\tilde{\mathbf{y}}_m, \mathbf{y}_m)}{D(\mathbf{x}_m, \mathbf{y}_m)} \quad (11)$$

where the distance function is defined in equation 8. It can be seen that $0 \leq P < 1$.

Three sets of experiments have been carried out: using 5 sentences (about 1,500 aligned vectors), using 20 sentences (about 7,100 aligned vectors), and using 100 sentence (about 38,500 vectors) for the training. For validation 5, 20 and 20 different sentences were used. In figure 2 and figure 3 the results of systems based on GMM, source HMM, joint HMM method A and B, CART and CART allowing more than one Gaussian per leaf (CART+) are presented for the three sets of training data, converting the male speaker to the female and vice-versa. To estimate GMM and HMM systems several number of components and states have been tested. In the figures it is indicated the optimal number. Also, it is indicated the number of final Gaussians for CART systems.

When few training data is available GMM, source HMM and both CART’s, performs in a similar way. But when the amount of training data increases, CART systems outperform GMM and source HMM. So, the inclusion of phonetic information allows a better splitting of the acoustic space. On the other hand, both CART and CART+ systems performs similar, so there is no need to tune the number of Gaussians. This is a very computational expensive part in GMM, and also in HMM tuning the number of states. However, CART systems need training data phonetically labeled, what restricts their applications.

Concerning the use of joint source-target information to estimate HMM, from the experimental results it seems better to use only source data. We must take into account that using joint data increases the vector dimensions. This result contrasts with the previous studies which showed that GMM systems trained with only source data or with joint source-target data had a similar performance [8].

4.2. Perceptual test

To evaluate the proposed systems an ABX test and a preference test has been carried out. In ABX test, A and B represents either the source or target speaker and X the converted speech. The listeners are asked to select if X is closer to A or B. In the preference test, pairs of sentences are presented, and the listeners are asked to select the most natural one for each pair. The following pairs have been chosen to be tested: GMM-sourceHMM and GMM-CART with 20 training sentences, Each listener, 10 in total, evaluates three examples of each pair. All of

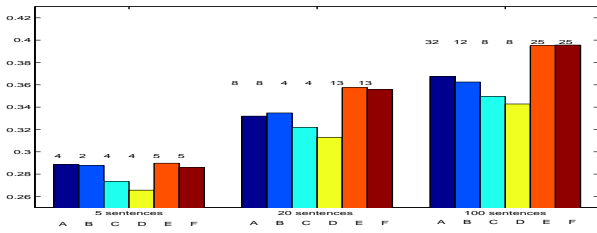


Figure 2: Performance index male → female: A) GMM, B) source HMM, C) joint HMM A, D) joint HMM B, E) CART, F) CART+

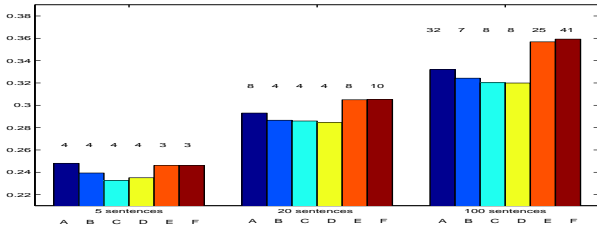


Figure 3: Performance index female → male: A) GMM, B) source HMM, C) joint HMM A, D) joint HMM B, E) CART, F) CART+

them had tests with different speech files and the systems were presented in different order.

To synthesize the test speech data, the transformed LPC filters derived from the transformed LSF are fed with the original residual signal of the target speaker. As we have presented a vocal tract conversion system, our intention is to measure only the effects of the vocal tract, assuming an ideal residual signal transformation.

The listeners reported that all the methods explained in this paper achieve the changing in the speaker identity. When they are asked about GMM and source HMM systems, they couldn't notice any difference. But, when GMM-CART pairs are presented, listeners chose the CART system 71% of the time. These perceptual results are correlated with the objective test.

5. Conclusions

In this paper, two new approaches are presented in order to extend the GMM-based VC systems. The first proposed system uses dynamic information to build the speaker acoustic model. GMM are replaced by HMM to model not only the probability density but also the dynamics of the speaker features. So, the transformation is carried out according to sequences of frames.

The second proposed system introduces phonetic information in the estimation of the mapping function and during the transformation, using CART decision trees.

The objective results have shown that GMM and source HMM systems performs in a similar way, but CART systems improves the performance of the conver-

sion. Also, with the proposed methodology, CART systems do not need any tuning of their parameters. Perceptual results are correlated with objective results.

6. Acknowledgements

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, <http://www.tc-star.org>) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, <http://gps-tsc.upc.es/veu/aliado>).

7. References

- [1] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *International Conference on Acoustics, Speech, and Signal Processing*, 2001.
- [2] D. Sündermann and H. Höge, "VTLN-Based Cross-Language Voice Conversion," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2003, pp. 676–681.
- [3] M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion using bilingual and non-bilingual databases," in *International Conference on Spoken Language Processing*, 2002.
- [4] J. Hosom, A. Kain, T. Mishra, J. van Santen, M. Fried-Oken, and J. Staehely, "Inteligibility of modifications to dysarthric speech," in *International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. 924–927.
- [5] O. Turk and L. Arslan, "Subband based voice conversion," in *International Conference on Spoken Language Processing*, Bogazici University, Istanbul, 2002.
- [6] E. Kim, S. Lee, and Y. Oh, "Hidden Markov model based voice conversion using dynamic characteristics of speaker," in *European Conference On Speech Communication And Technology*, 1997, pp. 1311–1314.
- [7] R. Laroia, N. Phamdo, and N. Farvardin, "Robust efficient quantization of speech LSP parameters using structured vector quantizers," in *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 641–644.
- [8] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing*, 1998.