# Including dynamic information in voice conversion systems*

**H. Duxans, A. Bonafonte**
TALP Research Center
Universitat Politecnica de Catalunya
Barcelona, Spain

**A. Kain, J. van Santen**
Center for Spoken Language Understanding
OGI School of Science Engineering
Oregon Health Science University
Portland, Oregon, USA

**Resumen:** Los sistemas de conversión de voz modifican la voz de un locutor (*locutor fuente*) para que se perciba como si hubiera sido producida por otro locutor (*locutor objetivo*). Muchos trabajos se basan en un modelado mediante mezcla de Gaussianas de las características conjuntas de ambos locutores, realizado asumiendo independencia para cada tramo de voz. En este artículo se estudia la inclusión de información dinámica, tanto del locutor fuente, como del locutor objetivo o de ámbos. Los sistemas propuestos se comparan basándose en medidas objetivas y perceptuales.
**Palabras clave:** conversión de voz, GMM, HMM

**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 performs the conversion in a frame by frame basis. In this paper, the inclusion of dynamic information of the source, target or both joint source-target speakers in the conversion is studied. Objective and perceptual results compare the performance of the proposed systems.
**Keywords:** voice conversion, GMM, HMM

## 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, education, medical aids and entertainment.

Nowadays, high quality TTS are based on the concatenation of acoustic units, i.e. to produce an utterance the most appropriated acoustic units are selected from a single speaker stored database. Then some strategy of spectral continuity is applied to joint the selected units together. In order to have available a wide range of acoustic units, huge amount of pre-recorded labeled data is needed, what makes expensive and time consuming to develop a new speaker voice.

VC can be a fast and a cheap way to build new voices for a TTS. So, it will be able to read e-mails or SMS with their sender's voice, to assign our and ours friends voices to characters when playing on a computer, or to give different voices to different computer applications. Recently (Kawanami et al., 2003), VC has been also applied to emotional speech synthesis, as an aid to prosodic modifications on a neutral sentence.

VC can also be very useful in interpreted telephony, when the translation task requires speaker identification by listeners. For example, in a conference call with three participants is very important to be able to differentiate between speakers by their voices. Also, it can help to learn foreign languages (Mashimo et al., 2001),(Mashimo et al., 2002), especially in pronunciation exercises, when students would listen to their own voices pronouncing foreign sounds properly.

Another field to apply VC is in speaking aids for people with speech impairments, improving the intelligibility of abnormal speech uttered by a speaker who has speech organ problems (Hosom et al., 2003). Also, it can be useful for designing hearing aids appropriated for specific hearing problems. In dive practices, VC can be applied to enhance the helium speech signals of the submariners.

There are several applications of VC in the multimedia entertainment too. One of the most obvious is karaoke, where the singer can be "helped" to success in every kind of songs. Also, some experiments have been done in

* 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).

film dubbing and looping (replacing undesired utterances with the desired ones) (Turk and Arslan, 2002), and in restoring old films.

VC can also be applied to the most classical fields of speech technology, for example in very low bandwidth speech encoding, transmitting the speech without speaker information, and adding it at the decoding step. Moreover, acquiring a high level of knowledge about speaker individuality can help speech or speaker recognition tasks.

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 store several speech databases, one for each speaker. So, the amount of training data is not a problem. Only high quality is required.

The following topics are studied, introducing a new approach to vocal tract conversion:

- The effects of including dynamic characteristics in the acoustic model used to build local vocal tract mapping functions.
- The effects of including joint source and target information during the training of the vocal tract acoustic model.

The outline of this paper is as follows. In section 2 the block architecture of VC systems is presented. Then in section 3 a GMM-based system is explained. In section 4 the inclusion of delta parameters to the acoustic model is proposed, and in section 5 a new approach based on HMM is introduced. Finally, in section 6 the results are discussed and the conclusions can be found in section 7.

## 2. *Voice conversion system architecture*

In figure 1 a generic VC system is presented. A VC system has two different operating modes: the training step and the transforming step. In the training step, all the components of the system are estimated from source and target speech data. So, for each new target speaker, a training must be carried out. Once the conversion function for a set of two speakers has been learned, any source utterance can be converted to sound as if the target speaker had uttered it.

Several features has been used in VC. They can be grouped in:

- Parametric features: formant frequencies and bandwidths, also glottal flow parameters (Narendranath et al., 1995) (Gutierrez-Arriola et al., 1998) (Mori and Kasuya, 2003) (Rentzos et al., 2003).
- LP related features. These kinds of features are based on the source-filter model for speech production. Usually, the polynomial coefficients are derived to other parameters with better interpolation properties, such as: LSF (Kain and Macon, 2001)(Arslan, 1999), lar (Iwahashi and Sagisaka, 1995), reflexion coefficients (Verhelst and Mertens, 1996) or LPC cepstrum.
- Spectral features without assuming any signal model, such as spectral lines (Sündermann and Höge, 2003) or mel frequency cepstrum (Masuko et al., 1997) (Mashimo et al., 2001).

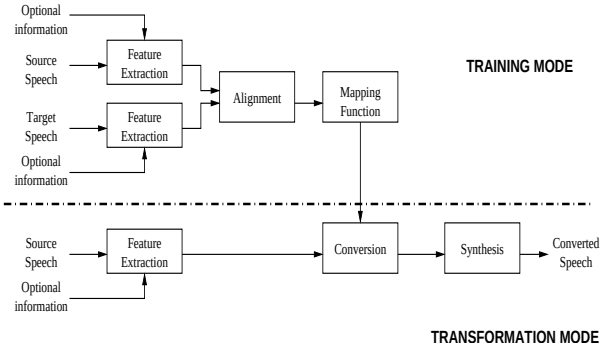


Figure 1: VC system block diagram

In this paper LSF features, extracted pitch synchronous, are used as vocal tract parameters. As it was mentioned, this paper is focused on studying a new vocal tract conversion system approach, so dealing with the residual LP signal remains as a future study and it is out of the scope of this paper.

Once both speaker's training data is parametrized, some kind of alignment is needed at the frame level in order to learn the mapping. Several strategies has been used, from manual alignment to DTW (Abe et al., 1988)(Kain and Macon, 2001), sentence HMM (Arslan, 1999) or source-target class mapping (Sündermann and Ney, 2003) for corpus with different contents for source and target speakers. We use lineal frame alignment, based on phoneme labeling.

Most of the techniques used for the mapping functions come from the fields of speaker recognition and speaker adaptation for au-

tomatic speech recognition systems. Usually, a vocal tract mapping function is trained by estimating the correspondence between spectral features of the source speaker with aligned features of the target. Then, residual adjustments and prosodic modifications are carried out. Several approaches have been used for the spectral mapping, such as mapping codebooks (Abe et al., 1988), Linear Multivariate Regression (LMR) and Dynamic Frequency Warping (DFW) (Valbret, Moulines, and Tubach, 1992), Speaker Transformation Algorithm using Segmental Codebooks (STASC) (Arslan, 1999), speaker interpolation (Iwahashi and Sagisaka, 1995), or Artificial Neural Networks (ANN) (Narendranath et al., 1995). In the next section, one of the standard vocal tract conversion system is presented, as a baseline system for comparisons.

## 3. *Baseline system: GMM-based voice conversion*

A GMM can model the probability distribution of any feature vector $\mathbf{x}$ as a sum of Q multivariate Gaussian functions,

$$p(\mathbf{x}) = \sum_{q=0}^{Q-1} \alpha_q N(\mathbf{x}; \mu_q, \mathbf{\Sigma}_q) \qquad (1)$$

$$\sum_{q=0}^{Q-1} \alpha_q = 1 \qquad \alpha_q \geq 0 \qquad (2)$$

where $N(\mathbf{x}; \mu_q, \mathbf{\Sigma}_q)$ is a normal distribution and $\alpha_q$ is the prior probability of the Gaussian $q$. The parameters $(\alpha_q, \mu_q, \mathbf{\Sigma}_q)$ can be estimated using the Expectation-Maximization (EM) algorithm.

A GMM is highly suitable to model a speaker acoustic space, since it can deal with different acoustics classes. Also, the classification of a frame is smoothed and the transformation function continuous, avoiding spectral jumps in the transformed speech.

The baseline system chosen is based on modeling the joint acoustic space of the source and target features with a GMM, first published in (Kain and Macon, 1998). The GMM is estimated maximizing the likelihood function of the joint source-target probability.

The transformation function can be obtained through the regression on the GMM of the target given the source parameters:

$$F(\mathbf{x}) = E[\mathbf{y}/\mathbf{x}] = \int \mathbf{y} p(\mathbf{y}/\mathbf{x}) dy \qquad (3)$$

$$F(\mathbf{x}) = \sum_{q=0}^{Q-1} h_q(\mathbf{x})[\mu_q^{\mathbf{y}} + \mathbf{\Sigma}_q^{yx} \mathbf{\Sigma}_q^{xx^{-1}} (\mathbf{x} - \mu_q^x)] \qquad (4)$$

where $\mathbf{x}$ and $\mathbf{y}$ are source and target feature vectors, and $h_q(\mathbf{x})$ is the posterior probability of the qth Gaussian.

The mixture of gaussians splits the acoustic space according joint information, and learns a mixture of linear regression functions.

Conversion systems using GMM work with a frame by frame basis. It means that to convert one frame the information about past and future frames isn't relevant. This is a simplification of the real speech production mechanism. Our propose is to include dynamic information in the voice conversion task. Two alternatives are presented: to extend the parameters employed in the estimation of the GMM to include dynamic information, or to extend the acoustic model using HMM to model not only the probability density but also the dynamics of the speaker features.

## 4. *GMM with delta parameters*

The first approach to include some dynamics in the voice conversion task is keeping the same mathematical model for the acoustic space and changing the parameters employed in training. So, in the training of a joint GMM the following parameters are used: source LSF and $\Delta$LSF, and target LSF and $\Delta$LSF. Then, source LSF and $\Delta$LSF are used to estimate only target LSF. Note that the target dynamics are used only in the training step, while source dynamics are used both in the training and transformation steps. The reason to include target $\Delta$LSF in the training is to allocate the class parameters more judiciously.

As a delta parameters, smoothed delta over N=2 periods are used:

$$\Delta\mathbf{x}(n) = \frac{\sum_{i=-N; i\neq 0}^{N} i\mathbf{x}(n+i)}{\sum_{i=1}^{N} 2i^2} \qquad (5)$$

To include $\Delta$LSF in the acoustic model implies to estimate conversion function parameters of twice dimension. So, the amount of

training data will be more critical than working with only LSF.

## 5. *HMM-based voice conversion*

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 $ith$ state and $\pi_i$ the initial probability of the $ith$ 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. LSF parameters has been used as a vocal tract features. In this section, we don't use $\Delta$LSF.

The block diagram of a HMM-based VC system is presented in figure 2.

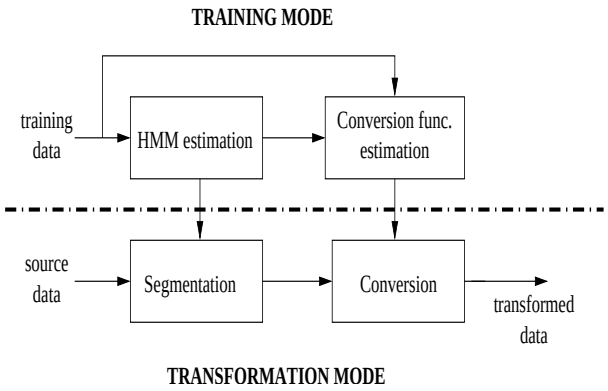


Figure 2: HMM-based VC system block diagram

In the training step, an 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 according the HMM states. Then, each frame is transformed applying the conversion function of its segmentation state.

It should be remarked that choosing appropriated sequences in training a HMM is critical, as our goal is to capture its dynamics. So, only phonetically matching sentences between source and target speakers are used.

### 5.1. *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 propose in (Kim, Lee, and Oh, 1997), 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:

$$F_s(x) = \mu_s^y + \mathbf{\Sigma}_s^{yx}\mathbf{\Sigma}_s^{xx^{-1}}(\mathbf{x} - \mu_s^x) \quad (6)$$

as a conversion function, where $s$ indicates the state. To transform a new sequence, we need to segment it according to the *source* HMM. Then, the conversion function of each state is applied.

### 5.2. *Target* HMM-based system

When *source* HMM-based system is used for VC, the transformed speech follows the dynamics of the source speaker, as it determines the sequence of transformation functions applied. In order to incorporate the dynamics of the target speaker, the HMM can be estimated with *target* data. In this case, the first step of the training is to estimate a HMM with *target* data. Then the same *target* data is segmented according to the *target* HMM. For each state, all the target vectors with their source aligned vectors are collected, and a Gaussian function is estimated to model the *source* data. The conversion function is estimated for each state in the same way than before, it is building and regressing a joint Gaussian for each state.

In the transformation step, the *source* sequences are segmented by Viterbi according the estimated *target* transition probabilities between *target* HMM states, and the emission probabilities of the source speaker vectors estimated by the Gaussians. Then, the

corresponding transformation functions are applied. Although the segmentation is not as accurate as in the previous case, it is expected that the correct dynamics increases the perceptual performance.

### 5.3. 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 task-oriented, 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. On one hand, the new sequence can be segmented according to the optimal state path $s^*$:

$$s^* = \arg\max_s p(\mathbf{x}, \mathbf{s}/\lambda) \tag{7}$$

$$s^* = \arg\max_s p(\mathbf{x}/\mathbf{s}, \lambda) p(s/\lambda) \tag{8}$$

where $\lambda = (a_{ij}, b_i(\mathbf{x}), \pi_i)$ $i = 1...N$, 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 is to include the regression in the search of the optimal path.

$$s^* = \arg\max_s p(\mathbf{y}, \mathbf{x}, \mathbf{s}/\lambda) \tag{9}$$

$$s^* \approx \arg\max_s p(\tilde{\mathbf{y}}, \mathbf{x}, \mathbf{s}/\lambda) \tag{10}$$

$$s^* \approx \arg\max_s p(\tilde{\mathbf{y}}/\mathbf{x}, \mathbf{s}, \lambda) p(\mathbf{x}/\mathbf{s}, \lambda) p(\mathbf{s}/\lambda) \tag{11}$$

where $\tilde{y}$ 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 11 allows to compute it applying the regression function of the state $s$ to the source frame.

Both approaches, called method A and method B respectively, will led to different solutions.

## 6. *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 total corpus size is more than one hour for each speaker (one male and one female), but we use few sentences from each one.

The frame alignment used is lineal inside each phoneme. Only phonetic transcription matching sentences are used. To train HMM's, each sentence without pauses is considered a sequence.

To evaluate the proposed systems objective and perceptual test have been carried out.

### 6.1. Objective Tests

The performance index used for test is:

$$P = 1 - \frac{IHMD(\tilde{\mathbf{y}}, \mathbf{y})}{IHMD(\mathbf{x}, \mathbf{y})} \tag{12}$$

where the distances are Inverse Harmonic Mean Distance (Laroia, Phamdo, and Farvardin, 1991). As it can be seen, the optimal performance is $P = 1$, while a conversion system that doesn't change the source speech will led to $P = 0$. The expression for Inverse Harmonic Mean Distance is:

$$IHMD(\mathbf{x}, \mathbf{y}) = \sum_{p=1}^{P} c(p)(x(p) - y(p))^2 \tag{13}$$

$$c(p) = \frac{1}{w(p) - w(p-1)} + \frac{1}{w(p+1) - w(p)} \tag{14}$$

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.

Two sets of experiments has been carried out: using 20 sentences (about 7.800 aligned vectors), and using 162 sentence (about 68.000 vectors) for the training.

Figures 3 and 4 show the results for the voice conversion task from a male speaker to a female speaker, and vice-versa for the two amounts of data. In each case, after training 4, 8, 12, 16, 20, 32, 64 components of the mixture, the optimal number is shown. That corresponds to 8 components for 20 sentences and 20 components for 162 sentences. The systems tested are: baseline GMM, GMM with delta parameters (dGMM), source HMM (sHMM), target HMM (tHMM) and joint HMM method A and B (jHMMA and jHMMB).

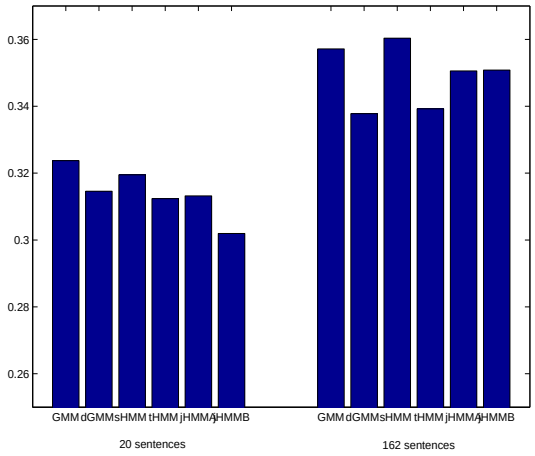


Figure 3: Performance index for male→female conversion.

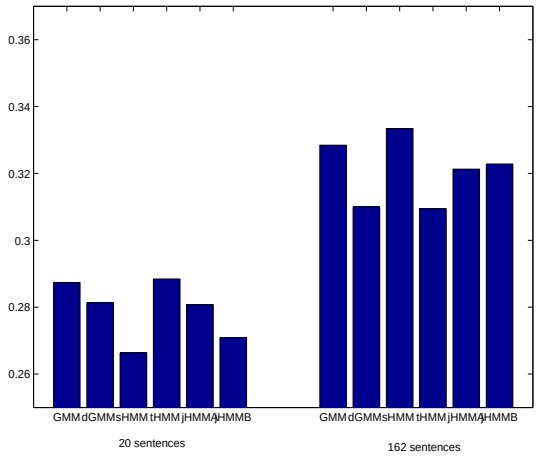


Figure 4: Performance index for female→male conversion.

It can be seen that the performance of the system depends on the speakers involved in the conversion. Moreover, doing the conversion in one direction or in the other changes the performance.

As it was expected, increasing the amount of training data improves the performance. It is more relevant in HMM, since the model has more parameters, so more data is needed to estimate it accurately. For this reason, the GMM-based system performs similar to the HMM-based systems with only 20 training sentences, but *source* HMM-based system has slightly higher performance index value using 162 sentences. However, the differences are minimal, so we expect similar quality perceptual rates.

Concerning the use of joint source-target information, 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.

As a remark, all the HMM have learned a similar topology. Not all the states are likely to be initial states, and the probability of remaining in the actual state is larger than changing of state.

### 6.2. Perceptual test

Two kinds of perceptual test have been carried out: ABX and preference test. 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. ABX test forces the listeners to choose between A or B, although the transformed speech X can neither resemble the source or the target speaker. In the preference test, pairs of sentence 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: GMM20-GMM162, GMM162-sHMM162, and for each test the listener evaluates three examples of each pair. Both tests have been done in male to female and female to male conversions. The number of listeners was 10. All of 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. Also, we have imposed the target prosody (including pitch values) to the source and transformed speech, in order to avoid preferences due to prosodic characteristics.

The listeners reported that all the methods explained in this paper achieve the chang-

ing in the speaker identity. But they reported great difficulties in the preference test, saying that all the speech files have the same quality and naturalnesses. When they are forced to decided between methods, we can only conclude that GMM162 presents higher quality than GMM20. No significant results are observed between GMM162 and sHMM162, and also they are not very reliable because the listeners couldn't distinguish any difference between them. The results of the preference test are showed in figure 5.

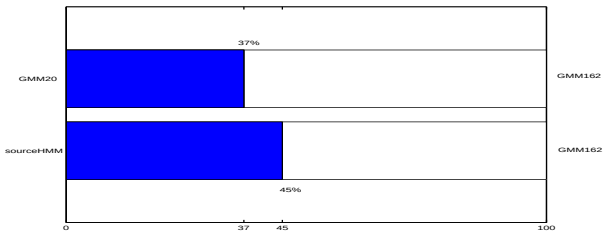


Figure 5: Results of the preference test.

## 7. Conclusions

In this paper a new voice conversion approach is presented. It is based on including dynamic information in the training of the conversion function. So, previous published GMM-based systems that work with a frame by frame basis have been extended. Two alternatives have been presented: to extend the parameters employed in the estimation of the GMM including $\Delta$LSF, or to extend the acoustic model using HMM to model not only the probability density but also the dynamics of the speaker features. In this latter case, both only source or target data and joint source-target data have been used as training data for the acoustic models.

The objective results have shown that the inclusion of delta parameters doesn't improve the performance of a GMM-based system. On the other hand, the performance of the HMM-based systems depend on the amount of training data. When the system were tested using 162 training sentences source HMM presented higher performance index value then the GMM-based system. However, when perceptual tests have been carried out, the listeners reported no perceptual differences between both methods.

As a future work, we are studying the effects of including phonetic information (the actual phoneme and its characteristics such as: point of articulation, voiced and manner) through unsupervised learning to the estimation of the conversion function.

## References


Abe, M., S.Ñakamura, K. Shikano, and H. Kuwabara. 1988. Voice conversion vector quantization. In *International Conference on Acoustics, Speech, and Signal Processing.*

Arslan, L.M. 1999. Speaker Transformation Algorithm using Segmental Codebooks (STASC). *Speech Communication*, 28:211–226.

Gutierrez-Arriola, J.M., Y.S. Hsiao, J.M. Montero, J.M. Pardo, and D.G. Childers. 1998. Voice Conversion based of parameter transformation. In *International Conference on Spoken Language Processing.*

Hosom, J.P., A.B. Kain, T. Mishra, J.P.H. van Santen, M. Fried-Oken, and J. Staehely. 2003. Inteligibility of modifications to dysarthric speech. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 924–927.

Iwahashi, N. and Y. Sagisaka. 1995. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. *Speech Communication*, 16:139–151.

Kain, A. and M. W. Macon. 1998. Spectral voice conversion for text-to-speech synthesis. In *International Conference on Acoustics, Speech, and Signal Processing.*

Kain, A. and M. W. Macon. 2001. 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.*

Kawanami, H., Y. Iwami, T. Toda, H. aruwatari, and K. Shikano. 2003. GMM-based voice conversion applied to emotional speech synthesis. In *European Conference on Speech Communication and Technology*, pages 2401–2404.

Kim, E.K., S. Lee, and Y.H. Oh. 1997. Hidden Markov model based voice conversion using dynamic characteristics of speaker. In *European Conference On Speech Communication And Technology*, pages 1311–1314.

Laroia, R., N. Phamdo, and N. Farvardin. 1991. Robust efficient quantization of speech LSP parameters using structured vector quantizers. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 641–644.

Mashimo, M., T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell. 2002. Evaluation of cross-language voice conversion using bilingual and non-bilingual databases. In *International Conference on Spoken Language Processing*.

Mashimo, M., T. Toda, K. Shikano, and N. Campbell. 2001. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In *European Conference on Speech Communication and Technology*.

Masuko, T., K. Tokuda, T. Kobayashi, and S. Imai. 1997. Voice characteristics conversion for HMM-based speech synthesis system. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1611–1614.

Mori, H. and H. Kasuya. 2003. Speaker conversion in ARX-based source-formant type speech synthesis. In *European Conference on Speech Communication and Technology*, pages 2421–2424.

Narendranath, M., H.A. Murthy, S. Rajendran, and B. Yegnanarayana. 1995. Tranformation of formants for voice conversion using artificial neural networks. *Speech Communication*.

Rentzos, D., S. Vaseghi, Q. Yan, C. Ho, and E. Turajlic. 2003. Probability models of formants parameters for voice conversion. In *European Conference on Speech Communication and Technology*.

Sündermann, D. and H. Höge. 2003. VTLN-Based Cross-Language Voice Conversion. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 676–681.

Sündermann, D. and H.Ñey. 2003. An Automatic Segmentation and Mapping Approach for Voice Conversion Parameter Training. In *International Workshop on Advances in Speech Technology*.

Turk, O. and L.M. Arslan. 2002. Subband based voice conversion. In *International Conference on Spoken Language Processing*, Bogazici University, Istanbul.

Valbret, H., E. Moulines, and J.P. Tubach. 1992. Voice transformation unsing PSOLA technique. *Speech Communication*, 11:175–187.

Verhelst, W. and J. Mertens. 1996. Voice conversion using partitions of spectral feature space. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 365–368.