VOICE CONVERSION USING FRAME SELECTION AND WARPING FUNCTIONS

A. J. Uriz[†] P. D. Agüero[†] J. C. Tulli[†] E. L. González[†] A. Bonafonte Cávez[‡]

[†]*Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina*
{*ajuriz, pdaguero, jctulli, elgonzal*}*@fi.mdp.edu.ar*
[‡]*Universitat Politècnica de Catalunya, Barcelona, Spain*

***Abstract*— The goal of voice conversion systems is to modify the voice of a source speaker to be perceived as if it had been uttered by another specific speaker. Many approaches found in the literature introduce an oversmoothing in the target features. Our proposal is the use of features produced by the target speaker without any smoothing to preserve speaker's identity. The first proposed algorithm combines several techniques used in unit selection for text-to-speech. The second one includes an additional renormalization stage before the frame selection algorithm. Objective results support the later proposed approach.**

***Keywords*— speech synthesis, voice conversion, frame selection**

1 INTRODUCTION

The primary goal of voice conversion systems is to modify the voice of a source speaker in order to be perceived as if it had been uttered by another specific speaker: the target speaker. For this purpose, relevant features of the source speaker are identified and replaced by the corresponding features of the target speaker.

In the area of text-to-speech synthesis (TTS) [7] voice conversion techniques play an important role. Since the output voice of TTS is obtained using a large speech database (at least one hour), the development of a new voice takes many time and resources. Voice conversion techniques may convert the output into a new voice by using just a small amount of data to find out the mapping function, reducing costs and development time.

Several voice conversion techniques have been proposed since the problem was first formulated in 1988. In this year Abe et al. [1] proposed to convert voices through mapping codebooks created from a parallel training corpus. Since then, many authors tried to avoid spectral discontinuities caused by the hard partition of the acoustic space by means of fuzzy classification [2] or frequency axis warping functions [18].

The appearance of statistical methods based on gaussian mixture models (GMM) for spectral envelope transformation was an important breakthrough in voice conversion [9, 13], because the acoustic space of speakers was partitioned into overlapping classes and the weighted contribution of all the classes was considered when transforming acoustic vectors. The spectral envelopes were successfully converted without discontinuities, but in exchange the quality of the converted speech was degraded by over-smoothing. This problem was faced in further works [4, 16, 19], while the usage of GMM-based techniques became almost standard, up to the point that the research was focused on increasing the resolution of GMM-based systems through residual prediction [8, 9, 14] in order to improve both the quality scores and the converted-to-target similarity.

Nevertheless, the problem of creating high-quality voice conversion systems that could be used in real-life applications has not been completely solved. At present, there is still a tradeoff between the similarity of converted voices to target voices and the quality achieved by the different conversion methods.

Erro et al. [6] presented a new voice conversion technique named Weighted Frequency Warping (WFW), which combined the conversion capabilities of GMM-based systems and the quality of frequency-warping transformations. The aim of WFW was to obtain a better balance between similarity and quality scores than previous existing methods. At the same time, other authors tried to improve conventional GMM-based systems by applying frequency-warping functions to residuals [15]. Both kinds of systems resulted in significant quality improvements and a slight decrement in the converted-to-target similarity scores, although they were conceptually different.

Speech synthesis with small databases to accomplish voice conversion without a transfer function was studied in Duxans et al. [8]. Although in this case the output speech waveforms were derived directly from the target training data, the identity of the target speaker could not be obtained. The artifacts introduced during the concatenation process (due to the reduced size of the database) degraded the speech signal and made difficult the identification.

Another interesting approach focused in improving target speaker identity is the frame selection approach proposed by Dutoit et al. [5]. In that paper the authors propose to find the optimal sequence of frame target features in training data reducing the distance between source converted features by GMM and target features by means of dynamic programming (the Viterbi algorithm, which was also used in the work of Salor and Demirekler [12]). Sündermann [15] proposed a similar approach just using the source features without any conversion.

In this paper we propose two systems that go back to Abe's proposal, with continuity constraints to avoid concatenation artifacts in speech. The main goal is to maximize the similarity to target speaker by using features extracted from training data, without any smoothing process. The already mentioned oversmoothing of other techniques in the literature produces target features that may not be uttered by the target speaker. In order to compare the system's performances, we made experiments with other state-of-the-art techniques: GMM, WFW and the method proposed by Dutoit in [5].

This paper is organized as follows. In Section 2, the five voice conversion techniques are explained in detail, emphasising the differences. In Section 3, the results of the objective tests are presented and discussed. Finally, the main conclusions are summarized in Section 4.

2 DESCRIPTIONS OF THE METHODS UNDER STUDY

In this Section we describe five methods to perform voice conversion: GMM, WFW, Dutoit and our proposals, Frame Selection (FS) and a novel method named Frame Selection Warped (FSW).

2.1 Voice Conversion Using GMM

Assuming that a parallel training corpus is available, the acoustic vectors of the source speaker, x_t, and those of the target speaker, y_t, may be aligned in pairs. Then, a joint-density GMM may be estimated from vectors z_t by means of the EM algorithm, where z_t is obtained by concatenating x_t and y_t. The resulting model is given by the weights p_i, the mean vectors μ_i and the covariance matrices Σ_i of its m gaussian components. Individual models for each speaker can be extracted from these parameters, since the mean vectors and covariance matrices can be decomposed into

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \tag{1}$$

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \tag{2}$$

Once the model is trained, it is possible to calculate the probability that a source vector x belongs to the i^{th} acoustic class (each gaussian component represents one of the m overlapping acoustic classes):

$$p_i(x) = \frac{\alpha N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{m} \alpha N(x, \mu_j^x, \Sigma_j^{xx})} \tag{3}$$

where $N(\cdot)$ denotes a gaussian distribution. In conventional GMM-based methods, each gaussian component is assigned a statistical transformation function, so for a given input vector x to be converted, the m probabilities $p_i(x)$ are used as weights for combining the contribution of all classes:

$$F(x) = \sum_{i=1}^{m} p_i(x) |\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx-1} (x - \mu_i^x)| \tag{4}$$

More information about GMMs can be found in [9, 13], with studies about the dimension of the matrices involved in training. In those papers some simplifications are proposed to reduce the number of parameters and the estimation error, such as diagonal covariance matrices.

2.2 Voice Conversion Using WFW

On the other hand, Erro et al. [6] proved that high-quality transformations were obtained if optimal frequency warping functions $W_i(f)$ were calculated for each class. Given an input vector x, the idea was to apply an individual envelope dependent frequency warping function for converting it, assuming that vectors belonging to the same acoustic class probably required similar warping trajectories. A full explanation of the method is shown in Erro's PhD Thesis [6].

2.3 Voice Conversion Using the Method Proposed by Dutoit

A new scheme for voice conversion was presented by Dutoit [5]. The algorithm proposed is a mixture between GMM and a frame selection method (FS). Voice conversion using frame selection uses a similar approach to unit selection, a commonly used technique in text-to-speech synthesis. In this case, the features of each frame are used to resynthesize the target speaker given an excitation signal using a filter, emulating the concatenation process of unit selection. The filter's coefficients are obtained from the features of the frame (for example, an LSF to LPC conversion) [7].

The method is based in two main steps, a GMM stage followed by a FS stage. Initially, the algorithm obtains the LSF parameters for each frame (x). These parameters are transformed through a GMM model, and a new set of parameters ($\hat{y}'$) are obtained. These are the transformed parameters using the GMM approach. The following step takes $\hat{y}'$ and finds the closest target vectors into the training set. Then, a frame selection algorithm is used to find the optimal sequence of target parameter vectors($\hat{y}$). In this process both target and concatenation costs are taken into

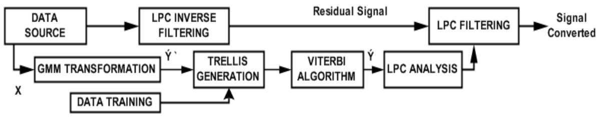

Figure 1: Model developed by Dutoit

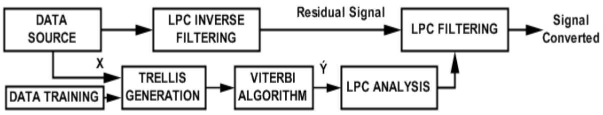

Figure 2: Method based on Frame Selection

account, favouring the selection of consecutive frames to minimize discontinuities. An scheme of the voice conversion method proposed by Dutoit is presented in Fig. 1.

2.4 Voice Conversion Using FS

Voice conversion techniques that do not use frame selection introduce a smoothing in the parameter estimation, and the converted feature vector may not be realizable by the target speaker.

In this paper we propose two voice conversion methods using frame selection to avoid smoothing effects. The first approach assumes that given a sequence of feature vectors of source speaker (x) we may find an optimal sequence of feature vectors of target speaker in training data ($\hat{y}$). The optimal sequence is obtained using the following formulation, which assumes that in the first frame ($i = 0$) the concatenation cost $d(\hat{y}_{i-1}, \hat{y}_i)$ is equal to 0:

$$min_{\hat{y}} \left[\sum_i d(x_i, x_j^{train}) + d(\hat{y}_{i-1}, \hat{y}_i) \right] \tag{5}$$

In this expression, $d(x_i, x_j^{train})$ represents the target cost which measures the distance between the source parameters of i^{th} frame and the source parameters of j^{th} frame in the training set. In this way we find appropiated converted target parameters assuming that closer source parameters imply closer target parameters. In this approach it is necessary to have temporally aligned source-target parameters vectors. The acoustic parameters included into the target cost are LSF, energy, fundamental frequency and phone identity (each phone is divided in three zones: start, medium and end). The phone identity is concatenated with the zone code to preserve the dynamics of phone evolution both for source and target frames.

The concatenation cost $d(\hat{y}_{i-1}, \hat{y}_i)$ minimizes the discontinuities between adjacent frames, and also favours the selection of consecutive frames. The parameters listed above are weighted to normalize their effects, and the weights are manually chosen using a subjective listening.

Therefore, the selection of the optimal sequence of target parameters is based on the sequence of source parameters, and the optimization is performed using the Viterbi algorithm.

A problem of computational load arises with the proposed conversion method: the size of the search space. The amount of frames in the lookup table is around 60.000 for a 15 minutes database using pitch synchronous analysis. As a consequence, we decided to introduce a clustering (k-means) of source feature vectors to reduce the search space. In this way we avoid a high number of available frames for each frame i (see Eq. 5) of source speaker.

We can summarize that the training process of the proposed algorithm consists of extracting the parallel feature vectors (20^{th} order LSF vectors) of source and target speakers. Then, a clustering of source feature vectors is performed to reduce the search space in the voice conversion task.

The steps of the complete frame selection algorithm are: given the source feature vector, the closest centroid with the same phone identity is found. Then, all source feature vectors of the closest centroid and the corresponding target feature vectors are included in the search as candidates.

The voice conversion task uses the Viterbi algorithm to obtain an optimal sequence of target feature vectors given the source feature vectors and the target and concatenation costs. Then, the source speaker excitation obtained by inverse filtering is used to synthesize the converted target voice through the converted LPC filters (obtained through an LPC to LSF conversion).

The fundamental frequency contour of the target speaker is obtained with a renormalization in mean and standard deviation of the source speaker contour in the log-scale. The pitch modification is synthesized using TD-PSOLA [10, 18].

2.5 Voice Conversion Using Frame Selection Warped FSW

The following algorithm tries to overcome some limitations of the method proposed in Section 2.4, such as acoustic discontinuities and abnormal trajectories of line spectral frequency parameters.

A study of the acoustic parameters showed that the use of an euclidean distance may generate problems in the frame selection process. Figure 3 shows the LSF parameters along the time of four aligned sentences. The first three sentences correspond to source, target and voice converted using FS, from top to bottom. In source and target sentences the behavior of the parameters is very similar, but in the FS converted sentence, the behavior is very different.

In this paper we propose to employ a mean and standard deviation normalization for each LSF parameter. The mean and standard deviation parameters are unique for each speaker.

This voice conversion step may be thought as a frequency warping to move the formants of source speaker

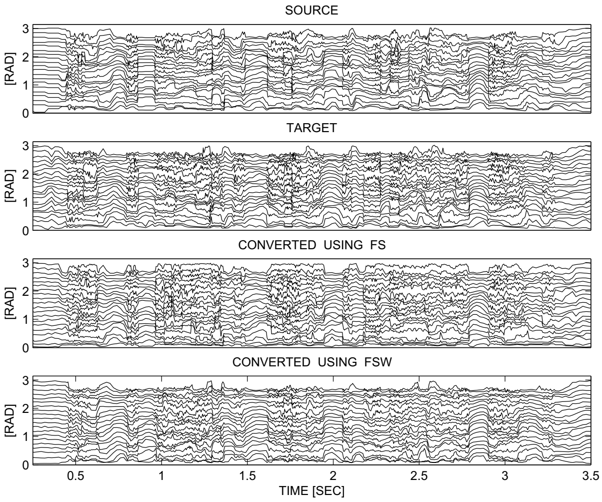

Figure 3: LSF trajectories of Source, Target , FS and FSW converted sentences.

towards the target formant position. For example, Fig. 4 shows the warping function from a Female Speaker into a Male Speaker in continuous line. The warping function that does not make any frequency changes is shown using a dashed line to ease the visualization of the warping process.

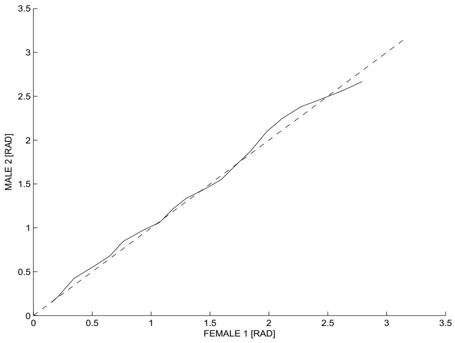

Figure 4: Warping function proposed.

The new proposed conversion scheme consists in performing a renormalization of source LSF parameters before the frame selection algorithm explained in Section 2.4. The statistical parameters for the renormalization (mean and standard deviation) are obtained from all frames in training data (without silent frames).

The result of the new algorithm is shown in the last aligned sentence of Fig. 3. The LSF parameters of the converted sentence have a similar behaviour compared with the trajectories of source and target LSF parameters. Figure 5 shows a zoom of a region where the advantage of the proposed method is evident. The proposed method (in the same way as Dutoit's algorithm) has two stages:

- The first stage makes a transformation using the mean-standard deviation. The LSF parameters of source speaker x are transformed into $\hat{y'}$.
- Then, the converted LSF parameters $\hat{y'}$ are transformed using a second stage based on frame selection to obtain a new set of parameters $\hat{y}$.

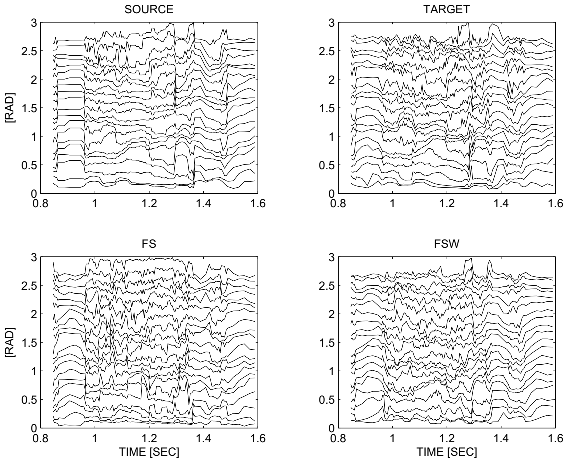

Figure 5: Zoom in a region of sentences when the advantage of proposed algorithm is shown.

Although both methods have a second stage based in frame selection, the number of parameters is reduced, and the estimation of the parameters is more accurate. For example, Dutoit's proposal needs 5120 parameters (assuming 32 gaussians multiplied by 20 LSF parameters multiplied by 2 directions multiplied by 4 vectors, with diagonal covariance matrices) versus 80 parameters of our proposal.

In the proposed approach is not necessary a temporal alignment of source and target frames, compared with the algorithm of Section 2.4. The frame selection is performed using the converted LSF parameters $\hat{y'}$ in the optimization formulation of Eq. 5. In this case, x_i does not correspond to the source LSF parameters, but to their converted counterpart $\hat{y'}$, and x_j^{train} are the training LSF target parameters y_j^{train}.

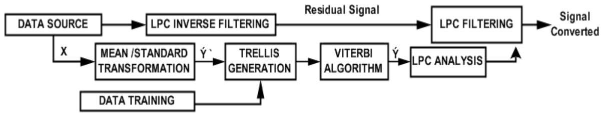

Figure 6: New model proposed

In order to improve the method, we performed an automatic weigth adjustment. We used an approach of text-to-speech synthesis: MultiLinear Regression (MLR) [3].

3 EXPERIMENTS

The audio database used for this experiment contained 200 sentences in Spanish, uttered by two male and two female speakers. The sampling frequency was 16 KHz

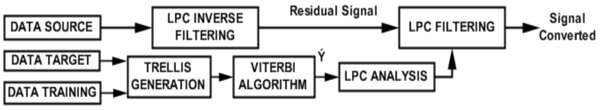

Figure 7: Architecture of FSWopt

and the average duration of the sentences was 4 seconds. 50% of the sentences were used to train the conversion functions, while 30% were kept as evaluation set (to tune model parameters) and 20% were used to perform objective test.

One male and one female speaker were chosen as source, and the other two speakers were used as target, so four different conversion directions were considered: male to male (m2m), female to female (f2f), male to female (m2f) and female to male (f2m). 38 sentences unseen during training were converted and resynthesized for all methods.

A sixth voice conversion method was included in the experiments. It consists of finding the closest feature vector of target speaker in training data to the real feature vector of target speaker. This voice conversion method based on FSW that uses privileged information is named **FSWopt**. It is a measure of the highest achievable quality and identity by the proposed method.

The results will be shown using box-plots [17]. This representation is an useful statistical tool to compare several statistical distributions. In our case we will use it to compare the distribution of the scores of the different systems to study the significance of the differences.

3.1 Objective Results Using P Distance

The P distance was used to measure the closeness of the converted voice to the target voice using the six voice conversion methods included in the experiments. The P distance was already used in several works about voice conversion [8, 13]:

$$P = 1 - \frac{d(y, \hat{y})}{d(x, y)} \tag{6}$$

The closer the converted parameters ($\hat{y}$) to the parameters of the target speaker (y) produces that P approaches to one. The distance between source parameters (x) and target parameters (y) allows to scale the P distance in the virtual path that goes from source to target parameters.

The results in Fig. 8 show that **FSWopt** is not as close as expected to the target voice, due to missing data in the limited training data.

The difference in P parameter between **FSWopt** and **FSW** shows the margin of improvement available using the proposed technique.

The lower score in **GMM**, **WFW** and **DUTOIT** shows identity problems in these techniques introduced by over-smoothing. FSW only scales and translates each LSF parameter using the warping function, without introducing any averaging.

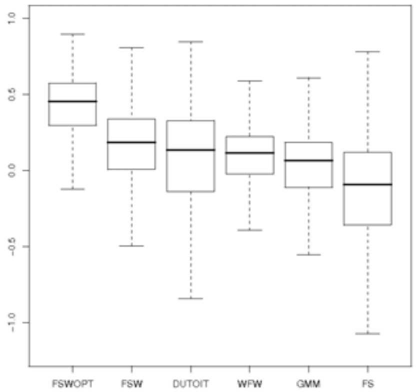

Figure 8: P distance.

FS obtains the worst objective score using the P distance, which shows the impact of the incorrect trajectories in the quality of the chosen LSF parameters.

3.2 Objective Results Using a Small Speaker Verification System

In order to validate the results of the first objective experiment, we made a second experiment using a small speaker verification system based on a GMM model [11]. MFCC coefficients are used to code the voice signal using a framing each 20ms. Two GMM models were trained using evaluation data to build source and target models. Given an utterance of a converted voice, these models may be used to establish the closeness to source and target. The substraction of the log-likelihood of source and target models is an indicator (score) of the performance of the conversion. A positive score indicates a good conversion, while a negative score is an indicator of closeness to source voice model.

The Fig. 9 presents the results of the different algorithms. Real source and target audio scores were also included to show the operation of this performance measurement method. The methods are ordered according to the median of the scores. The results show that the proposed method (FSW) has the higher performance without using privileged information.

4 CONCLUSIONS

In this paper we presented a voice conversion algorithm that avoids the smoothing effects of other proposals in the literature.

Target and concatenation costs are included in the search of the optimal sequence of target feature vectors given the sequence of source speaker's feature vectors (FS method). We also propose to use a sequence of converted source speaker's feature vectors (FSW method) using a warping function. Objective results show that the later proposed technique achieves a higher similarity to the target speaker, as expected due to the better trajectories observed in converted

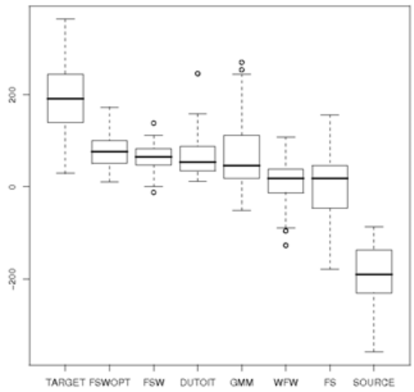

Figure 9: Second objective test.

voices.

Future work will be devoted to improve the quality of the voice conversion introducing a mapping in the excitation. In this paper the excitation extracted from the voice of the source speaker was used to synthesize the voice of the target speaker using converted target LPC parameters. Mismatches between LPC coefficients and excitation contributed to reduce the final quality. We will also study new ways to obtain converted voices using a more complex warping function without introducing smoothing.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 655–658, 1988.

[2] L.M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). In *Speech Communication*, number 28, pages 655–658, 1999.

[3] A. Black and N. Campbell. Optimising selection of unit from speech databases for concatenative synthesis. In *Proceedings Eurospeech*, pages 581–584, 1995.

[4] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu. Voice conversion with smoothed GMM and MAP adaptation. In *Proceedings of the European Conference on Speech Communications and Technology*, pages 2413–2416, 2003.

[5] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou. Towards a voice conversion system based on frame selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[6] D. Erro. Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models. PhD Thesis (Universitat Politècnica de Catalunya). 2008.

[7] X. Huang, A. Acero, and H.W. Hon. Spoken language processing. A Guide of Theory, Algorithm, and System Development. 2001.

[8] H. Duxans i Barrobés. Voice conversion applied to text-to-speech synthesis. In *PhD Thesis (Universitat Politècnica de Catalunya)*, 2006.

[9] A. Kain. High resolution voice transformation. In *PhD thesis, OGI - OHSU*, 2001.

[10] E. Moulines and F. Chanpentier. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Speech Communication*, 1990.

[11] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *Speech Communication 17*, pages 91–108, 1995.

[12] O. Salor and M. Demirekler. Voice transformation using principle component analysis based LSF quantization and dynamic programming approach. In *Proceedings of Interspeech 2005*, pages 1889–1892, 2005.

[13] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. In *Proceedings of ICASSP*, volume 6, pages 131–142, 1998.

[14] D. Sundermann, H. Hoge, A. Bonafonte, and H. Duxans. Residual prediction. In *Proceedings of the IEEE Symposium on Signal Processing and Information Technology*, pages 512–516, 2005.

[15] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan. Text-independent voice conversion based on unit selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[16] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 841–844, 2001.

[17] J.W. Tukey. Exploratory data analysis, addison-wesley. 1970.

[18] H. Valbret, E. Moulines, and J.P. Tubach. Voice transformation using PSOLA technique. In *Speech Communication*, volume 1, pages 145–148, 1992.

[19] H. Ye and S. Young. Quality-enhanced voice morphing using maximum likelihood transformations. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 14, pages 1301–1312, 2006.