Extended Conditional GMM and Covariance Matrix Correction for Real-Time Spectral Voice Conversion

Pierre Lanchantin, Nicolas Obin, Xavier Rodet

IRCAM - CNRS-UMR9912-STMS,
Analysis-Synthesis Team,
1, place Igor-Stravinsky,
75004 Paris, France

lanchant@ircam.fr, nobin@ircam.fr, rod@ircam.fr

Abstract

Gaussian mixture model (GMM)-based spectral voice conversion (VC) can be performed in real-time by applying the conversion method frame by frame. However, this local method can produce inappropriate trajectories of parameters and the converted spectrum can be excessively smoothed due to the statistical approach. In order to address these limitations, we propose an approach based on a new *Extended Conditional GMM* model. Two different features vectors are used for the description of the source characteristics: one is specifically designed for a precise description of the spectral features to be transformed, the other one being designed for the selection of the transformations to be applied. The latter include local descriptors of the trajectories of parameters via *Discrete Cosine Transform* (DCT) coefficients in order to generate local trajectories of parameters. Finally, the effect of over-smoothing is alleviated by a *covariance matrix correction* method. The proposed VC method is evaluated objectively and subjectively, showing a dramatic improvement compared to conventional VC method.

Index Terms: Voice conversion, Extended Conditional GMM, Discrete Cosine Transform.

1. Introduction

The aim of *speaker conversion* - a typical application of *voice conversion* technique (VC) - is to modify the speech signal of a source speaker so as to be perceived as that of a target speaker. The overall methodology for speaker conversion is to define and learn a mapping function of acoustic features of a source speaker to those of a target speaker. Among other statistical approaches described in [1, 2, 3], one of the most popular method proposed by Stylianou and al. [4] is based on a *Gaussian mixture model* (GMM) that defines a continuous mapping between the features of source and target voices. Kain extended Stylianou's work by modelling directly the joint probability density of the source and target speaker's acoustic space [5]. This method allows the system to capture all the existing correlations between the source and target speaker's acoustic features. In most cases, the method is applied frame by frame which make its implementation for real-time conversion straightforward. Although this type of method is relatively efficient, conversion performances are still insufficient regarding speech quality: the frame by frame conversion process induces inappropriate spectral parameters trajectories and the converted spectrum can be excessively smoothed. Toda and al. have proposed in [6] a method based on maximum likelihood estimation of trajectories of parameters, which greatly improves the quality of synthesis by taking into account the dynamic features and the global variance. However, this method require a global optimization which can be a problem for real-time applications.

In order to address these limitations, we propose a novel approach presented in this study. First, we define an *Extended Conditional GMM* (XcGMM) in which the mixture weights depend on an alternative representation of the source characteristics different from the one used for the description of the spectral characteristics to be converted. This modelling allows the use of a high resolution representation of the spectral characteristics to be transformed without necessarily increasing the complexity of the model. At the same time, it allows the inclusion of additional informations for the selection of the transformations to apply. In this way, *Discrete Cosine Transform* (DCT) can be used to stylize the trajectories of the spectral parameters. This additional parameters are taken into account in order to generate local trajectories of parameters. Finally, we propose a *covariance matrix correction* method to overcome the over-smoothing of the transformed spectral characteristics.

The paper is organized as follows: Section 2 presents the proposed approach and the related VC system, its optimization and objective evaluation are described in Section 3; finally subjective evaluation of the proposed approach is presented and discussed in Section 4.

2. Proposed Approach

Let $Z = (X, Y)$ be the joint random process of source-target acoustic spectral features in which $X = \{X_n\}_{n\in\mathcal{N}}$ and $Y = \{Y_n\}_{n\in\mathcal{N}}$ are the source, and target processes respectively, and $\mathcal{N}$ the set of frame indexes. Each X_n and Y_n takes its values in $\mathbb{R}^d$ where d is the dimension of the acoustic feature vector. We will denote $z = (x, y) = \{(x_n, y_n)\}_{n\in\mathcal{N}}$ a realization of this process, in which x_n and y_n are the acoustic features vector at frame n for the source and that for the target, respectively. We assume that Z is an independent and identically distributed process (i.i.d.) such as $p(z) = \prod_{n\in\mathcal{N}} p(z_n)$. We introduce the auxiliary i.i.d. process of mixture components $U = \{U_n\}_{n\in\mathcal{N}}$, each U_n taking its values in $\mathcal{U}$ with cardinal K. The joint probability distribution of the source and target features vectors is then modeled by a Gaussian mixture as follows

$$p(z_n|\phi) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(z_n; \phi_k) \tag{1}$$

This study was supported by FEDER Angelstudio : Générateur d'Avatars personnalisés ; 2009-2011

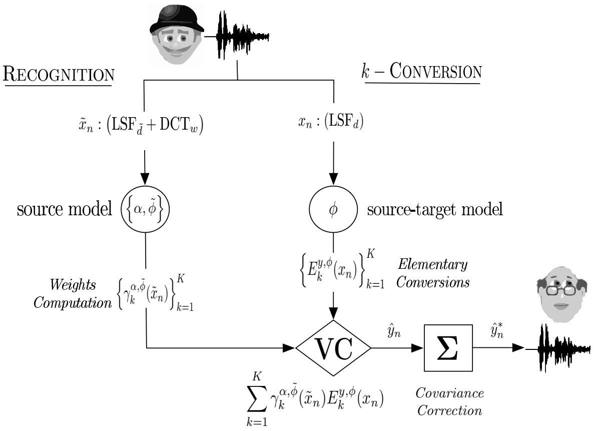
RECOGNITION
k – CONVERSION
$\tilde{x}_n : (\mathrm{LSF}_{\tilde{d}} + \mathrm{DCT}_w)$
$x_n : (\mathrm{LSF}_d)$
source model
$\{\alpha, \tilde{\phi}\}$
ϕ
source-target model
$\left\{E_k^{y,\phi}(x_n)\right\}_{k=1}^{K}$
Elementary Conversions
Weights Computation
$\left\{\gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n)\right\}_{k=1}^{K}$
VC
$\hat{y}_n$
Σ
$\hat{y}_n^*$
$\sum_{k=1}^{K} \gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n) E_k^{y,\phi}(x_n)$
Covariance Correction

Figure 1: *The proposed approach: at each frame $n \in \mathcal{N}$ the global conversion function is a weighted sum of elementary conversion functions $\{E_k^{y,\phi}(.)\}_{k=1}^{K}$, whose weights $\{\gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n)\}_{k=1}^{K}$ depend on an alternative representation $\tilde{x}_n$ of the source characteristics different from x_n which is used for the description of the spectral characteristics to be converted. A covariance matrix correction method is finally applied on the transformed vector.*

with $\mathcal{N}(z_n; \phi_k) = p(z_n | u_n = k; \phi_k)$ and the mixture weight $\alpha_k = p(u_n = k)$ for each $k \in \mathcal{U}$. $\phi_k = \{\mu_k^z, \Sigma_k^z\}$ is the parameters set including the mean vector $\mu_k^z = [\mu_k^x, \mu_k^y]^t$ and the covariance matrix $\Sigma_k^z = \left[[\Sigma_k^{xx}, \Sigma_k^{yx}]^t, [\Sigma_k^{xy}, \Sigma_k^{yy}]^t\right]$ for each mixture components $k \in \mathcal{U}$. μ_k^x and μ_k^y are the mean vector of the k-th mixture component for the source and that for the target, respectively. Σ_k^{xx} and Σ_k^{yy} are the covariance matrix of the k-th mixture component for the source and that for the target, respectively. Σ_k^{xy} and Σ_k^{yx} are the cross-covariance matrix of the k-th mixture component for the source and that for the target, respectively.

2.1. Extended Conditional GMM

We introduce an alternative source i.i.d process $\tilde{X} = \{\tilde{X}_n\}_{n \in \mathcal{N}}$ in which each $\tilde{X}_n$ takes its values in $\mathbb{R}^{\tilde{d}}$. $\tilde{x}_n$ is the alternative acoustic feature vector for the source at frame n which is used exclusively for the computation of the mixture weights of the transformation mixture. We also make the assumptions that for each $n \in \mathcal{N}$, u_n is independent of x_n conditionally on $\tilde{x}_n$ and y_n is independent of $\tilde{x}_n$ conditionally on (x_n, u_n) such that the conditional probability density of y_n given $x_n^+ = (x_n, \tilde{x}_n)$ is given by the following extended conditional Gaussian mixture distribution with $\tilde{\phi}_k = \{\mu_k^{\tilde{x}}, \Sigma_k^{\tilde{x}\tilde{x}}\}$, $\phi^+ = \{\phi, \tilde{\phi}\}$, $\phi = \{\phi_k\}_{k=1}^{K}$ and ditto for $\tilde{\phi}$ and α:

$$p(y_n | x_n^+; \phi^+) = \sum_{k=1}^{K} \gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n) \mathcal{N}\left(y_n; E_k^{y,\phi}(x_n), C_k^{y,\phi}\right) \quad (2)$$

where

$$\left\{ \begin{array}{l} \gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n) = p(u_n = k | \tilde{x}_n; \alpha_k, \tilde{\phi}_k) \propto \alpha_k \mathcal{N}(\tilde{x}_n; \tilde{\phi}_k) \\ \mathcal{N}\left(y_n; E_k^{y,\phi}(x_n), C_k^{y,\phi}\right) = p(y_n | x_n, u_n = k; \phi_k) \end{array} \right. \quad (3)$$

and

$$\left\{ \begin{array}{l} E_k^{y,\phi}(x_n) = \mu_k^y + \Sigma_k^{yx} \left(\Sigma_k^{xx}\right)^{-1} (x_n - \mu_k^x) \\ C_k^{y,\phi} = \Sigma_k^{yy} - \Sigma_k^{yx} \left(\Sigma_k^{xx}\right)^{-1} \Sigma_k^{xy} \end{array} \right. \quad (4)$$

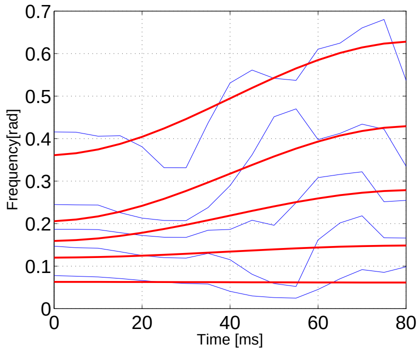

Figure 2: *2-order DCT trajectory stylization (red) of the first 5 LSFs (blue) on a 80ms window size.*

$\{\alpha, \phi^+\}$ are estimated on a parallel speech database (x_{train}^+, y_{train}) in which x_{train}^+ and y_{train} have been temporally aligned. To do so, the parameters $\left\{\alpha_k, \tilde{\phi}_k\right\}_{k=1}^{K}$ of the distribution of $\tilde{X}$ are first estimated by Expectation-Maximization on the speech database $\tilde{x}_{train}$. Then, at the last step of the estimation, ϕ_k can be estimated for each $k \in \mathcal{U}$ in the following way

$$\left\{ \begin{array}{l} \mu_k^z = \frac{\sum_{n=1}^{N} \gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n) z_n}{\sum_{n=1}^{N} \gamma_k^{\tilde{\phi}}(\tilde{x}_n)} \\ \Sigma_k^z = \frac{\sum_{n=1}^{N} \gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n) \left(z_n - \mu_k^z\right)\left(z_n - \mu_k^z\right)^t}{\sum_{n=1}^{N} \gamma_k^{\tilde{\phi}}(\tilde{x}_n)} \end{array} \right. \quad (5)$$

Note that when $x = \tilde{x}$, the proposed model is equivalent to the conventional one. However, the estimation procedure is sensibly different from the conventional approach in which the distribution of X is deduced for the estimation of the joint source-target distribution.

The conversion is finally performed on the basis of the minimum mean-square error (MMSE): the global conversion function is a weighted sum of *elementary conversion functions* - the conditional mean vectors $E_k^{y,\phi}(.)$ - whose weights depend on the source vector to be transformed.

$$\hat{y}_n = E[y_n | x_n^+; \phi^+] = \sum_{k=1}^{K} \gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n) E_k^{y,\phi}(x_n) \quad (6)$$

Each weight $\gamma_k^{\alpha,\tilde{\phi}}(\tilde{x}_n)$ gives the probability of a given source vector to belong to each one of the k-mixture components such that their computation can be interpreted as a *recognition* step, the components being seen as clusters. A schematic representation of the proposed VC system is presented on Figure 1.

2.2. Local Trajectory Modelling

The introduction of $\tilde{X}$ allows to use lower resolution representation of the spectral characteristics for the computation of the weights while keeping a high resolution representation of the spectral characteristics to be converted. Furthermore, it allows to considerate other descriptors which can facilitate the clusters discrimination. In this way, we use the *Discrete Cosine Transform* (DCT [7]) to locally stylize the acoustic features trajectories - Line Spectral Frequencies (LSF) in the following - over various temporal segments. The principle is to decompose acoustic features trajectories on a basis of slowly time-varying functions defined by zero-phase cosine functions $\varphi = (cos(\omega_1), \ldots, cos(\omega_M))$ at discrete frequencies $\omega_m =$

$\frac{\pi}{2T}(2m+1)$. The trajectories can be stylized using different DCT orders and different window sizes w. Compared to derivate and second derivate often used in conventional VC methods, DCT allow a finer description of trajectories using a reduced amount of parameters. The stylization over various temporal segments aims at representing the acoustic features trajectories with more or less details, and to model short and long term dependencies.

2.3. Covariance Matrix Correction

The source-target global covariance matrix can be deduced from the Gaussian mixture given in the equation 1 as follows:

$$\Sigma_G^z = \sum_{k=1}^{K} \alpha_k \left[\Sigma_k^z + (\mu_k^z - \mu_G^z)(\mu_k^z - \mu_G^z)^t\right] \qquad (7)$$

with $\mu_G^z = \sum_{k=1}^{K} \alpha_k \mu_k^z$. The loss of global variance which is usually observed and which result in the over-smoothing of the spectrum is explicitely given by the residual covariance matrix $\Sigma_G^{res} = \Sigma_G^{yx} \left(\Sigma_G^{xx}\right)^{-1} \Sigma_G^{xy}$. This covariance matrix represents the covariance non explained by the Gaussian mixture regression. The idea is to correct the converted frame values for each $n \in \mathcal{N}$ according to Σ_G^{res}. To do so, we use a Cholesky decorrelation/correlation method. The new value $\hat{y}_n^*$ of the converted source vector is given by

$$\hat{y}_n^* = \left[(\hat{y}_n - \mu_G^y)^t L_G^{yy} \left(L_G^{y|x}\right)^{-1}\right]^t + \mu_G^y \qquad (8)$$

where L_G^{yy} and $L_G^{y|x}$ are upper triangular so that

$$\begin{cases} (L_G^{yy})^t \, L_G^{yy} = \Sigma_G^{yy} \\ \left(L_G^{y|x}\right)^t L_G^{y|x} = \Sigma_G^{yy} - \lambda_G \Sigma_G^{res} \end{cases} \qquad (9)$$

where λ_G is a weight governing the amount of correction. Note that when λ_G=0 the resulting covariance matrix is the one of the target given the source vector and the value of $\hat{y}$ remains unchanged. However, when λ_G=1, the resulting covariance matrix become the one of the target.

3. System Optimization

The parameters of the proposed system were optimized on a development set. Both reduced dimension alternative source vector and stylization of the spectral parameters are shown to improve objectively the VC performance.

3.1. Evaluation procedure

Two speech databases have been recorded (32bits, F_s=44.1kHz) by French speakers with different accents (Standard and Hispanic). Each database includes 200 phonetically balanced utterances, which represent approximately 10 minutes of speech depending on speakers speech rate. The speech databases were then aligned by a *Dynamic Time Warping* (DTW) algorithm constrained by phone boundaries extracted using *ircamAlign* [8], an HMM-based alignment system based on the HTK toolbox. The obtained parallel speech database was then splitted into a training set (150 parallel utterances) a development set (30 parallel utterances) and a test set (20 parallel utterances). Line spectral frequencies (LSF) were used as spectral features for the source x_n^+ and target y_n vectors for each $n \in \mathcal{N}$. To do so, the spectral envelope was estimated each 5ms by True Envelope method on a Mel scale (MTE) and coded by Linear Predictive Coding (LPC). The optimal cesptral coefficient order considering the MTELPC [9] estimator was determined equal according to [9]. The order of the LPC was chosen equal to 45 which give a good approximation of the spectral envelope. Analysis and synthesis were performed using phase vocoder [10].

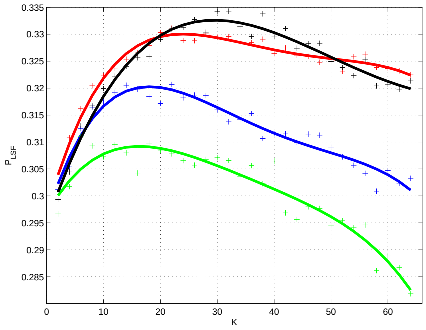

Figure 3: *P_{LSF} with respect to the number of mixtures components in the following cases: in green: the conventional method with d=45, in blue: the proposed method with $x = \tilde{x}$, in red: the proposed method with $\tilde{d}$=10, in black: the proposed method with w=80ms DCT window size.*

The performance index P_{LSF} proposed by Kain [11] was used for the objective evaluation of the VC performance:

$$P_{LSF} = 1 - \sum_{n=1}^{N} \frac{dist(y_n, \hat{y}_n)}{dist(y_n, x_n)} \qquad (10)$$

where $\hat{y}_n$ is the converted target vector and the distance $dist$ is Euclidean. The upper part is the *trans-speaker* error defined as the spectral distance between the converted speech and the target speaker, while the lower part is the *inter-speaker* error defined as the spectral distance between the source and target speaker.

3.2. Estimation procedure

The estimation procedure of the proposed approach is sensibly different from the conventional one. On the Figure 3, we presented in green line the P_{LSF} value with respect to the number of mixture components K obtained with the conventional approach with an LPC order d=45. The blue line corresponds to the P_{LSF} value obtained with the new approach with $x = \tilde{x}$. In this case, both approaches are equivalent excepted for the estimation method. The proposed estimation method outperforms the conventional one with a maximum of P_{LSF}=0.321 for K=22 compared to a maximum of P_{LSF}=0.310 for K=18. This indicates that estimating the mixture weights α only on the source part of the aligned speech database give better performance than the conventional method which consists of estimating the weights on the joint source-target aligned speech database.

3.3. Optimal LPC order of the alternative source vector

We kept x unchanged while decreasing the LPC order $\tilde{d}$ of $\tilde{x}$ ranging from 45 to 5. We observed an improvement of performance for each case where $\tilde{d} < d$ with a maximum performance P_{LSF}=0.331 for K=22 with $\tilde{d}$=10. For clarity, we only presented the results for $\tilde{d}$=10 in red line on the Figure 3. One explanation of this improvement can be the reduction of the complexity of the model which allows the use of more mixture components.

3.4. DCT Window Size

We kept the LPC order of 10 and we chose an order of DCT equal to 2 giving a dimension vector of $\tilde{d}$=30. We then varied the size of the frame centered window on which the DCT is computed at each frame $n \in \mathcal{N}$. The best performance curve, plotted in black on the Figure 3, is obtained for a window size w =80ms with a maximum of P_{LSF}=0.334 for K=32. It is of interest to note that the optimal temporal segment (80ms) is found to be closer to the phoneme one ($\sim$50ms) than to the syllable one ($\sim$200ms). This clearly indicates that the stylization process succeeds in modelling co-articulation.

4. Evaluation

In this section, we aim to evaluate the proposed system according to the *speaker individuality* and to the *quality* of the resulting converted speech.

4.1. Speaker individuality

The objective of the first part of the subjective test was to evaluate the *speaker individuality* after conversion of the source speech using the optimized system and the covariance correction method with λ_G=0.9. We aimed to qualify if the converted speech signal synthesized from $\hat{y}^*$ was perceived closer to the target, to the source or between them. 3 speech utterances of the source converted considering both methods (conventional vs proposed method) have been presented to 22 listeners. At each trial, an utterance was randomly chosen among the 12 utterances and presented to the listener which had to vote on the perceived position relative to the source and target on a scale of 5 values, 0 if perceived as the source and 4 if perceived as the target voice. The results of this first test are presented on the left part of Figure 4. Most listeners couldn't recognize neither the source not the target, which is a well known effect (*third-speaker effect*) reported in the litterature, especially in intra-gender VC task. Nevertheless, the listeners perceived the converted speech using the proposed method as significantly closer to the target speaker (MOS=2.17±0.18) than using the conventional method (MOS=1.75±0.22).

4.2. Speech quality

The second test was focused on the *quality* of the converted speech. A comparison category rating test (CCR) was used to assess the quality of the speech converted by the proposed method in comparison to the speech converted by the conventional method. 6 utterances were chosen to generate the test samples for each methods. Utterances were presented in random order to the listeners for evaluation. They were asked to attribute a score to the quality of the second sample of a pair compared to the quality of the first one on the comparison mean opinion score (CMOS) scale. The ranking of the 2 methods was evaluated by averaging the scores of the CCR test for each method. The listeners perceived the quality of the converted speech using the proposed method as drastically better than using the conventional method (CMOS=+1.00±0.18).

5. Conclusion

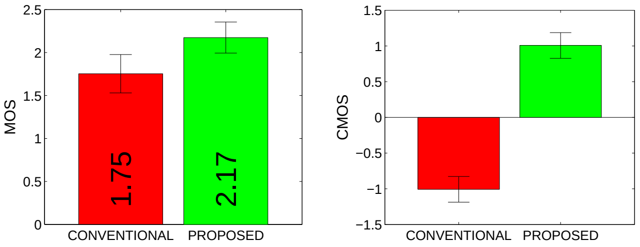

Figure 4: *Left figure: Speaker individuality considering both method (the vote scale has 5 values from 0 to 4). Right figure: Speech quality (CCR boxplots).*

We presented a novel approach for spectral VC based on XcGMM. In this model, the mixture weights depend on an alternative representation of the source characteristics different from the one used for the description of the spectral characteristics to be converted. In this way, the spectral characteristics to be transformed can be represented with a high LPC order without increasing the complexity of the model. Furthermore, this model allows to take into account local trajectories of parameters. Finally, we proposed a covariance correction method to alleviate the well-known over-smoothing of the converted spectral parameters. Objective and subjectives tests revealed that the proposed method dramatically improve the VC performances. Further researches include the study of the impact of the DCT order on the conversion performances. Finally, the covariance matrix correction method will be compared to the global variance method proposed in [6] .

6. References

[1] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

[2] P. Lanchantin and X. Rodet, "Dynamic Model Selection for Spectral Voice Conversion," in *Proc. Interspeech 2010*, Makuhari, Japan, Sept. 2010.

[3] E. Helander, H. Silen, J. Miguez, and M. Gabbouj, "Maximum a posteriori voice conversion using sequential monte carlo methods," in *Proc. Interspeech 2010*, Makuhari, Japan, Sept. 2010.

[4] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," in *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, 1998, pp. 131–142.

[5] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. International Conference on Acoustics Speech and Signal Processing (ICASSP'88)*, 1998, pp. 285–288.

[6] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[7] N. Obin, *Analysis and Modelling of Speech Prosody and Speaking Style*. Ph.D. dissertation, IRCAM, Paris VI University, 2011.

[8] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic phoneme segmentation with relaxed textual constraints," in *International Conference on Language Resources and Evaluation*, Marrakech, Morroco, 2008, pp. 2403–2407.

[9] F. Villavicencio, A. Röbel, and X. Rodet, "Extending efficient spectral envelope modeling to mel-frequency based representation," in *IEEE 2008 International Conference on Acoustics, Speech, and Signal processing (ICASSP'08)*, 2008, pp. 1625–1628.

[10] P. Depalle and G. Poirrot, "Svp: A modular system for analysis, processing and synthesis of sound signals," in *Proceedings of the International Computer Music Conference*, 1991.

[11] A. Kain, *High resolution voice transformation*. Ph.D. dissertation, OGI School of Science and Engineering at Oregon Health and Science University, 2001.