Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface

Thomas Hueber, Gérard Bailly, Bruce Denby

Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface

Thomas Hueber[1], *Gérard Bailly*[1], *Bruce Denby*[2]

[1]GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France
[2]ESPCI ParisTech / Université Pierre et Marie Curie, Paris, France
(thomas.hueber, gerard.bailly)@gipsa-lab.grenoble-inp.fr, denby@ieee.org

Abstract

The article presents an HMM-based mapping approach for converting ultrasound and video images of the vocal tract into an audible speech signal, for a silent speech interface application. The proposed technique is based on the joint modeling of articulatory and spectral features, for each phonetic class, using Hidden Markov Models (HMM) and multivariate Gaussian distributions with full covariance matrices. The articulatory-to-acoustic mapping is achieved in 2 steps: 1) finding the most likely HMM state sequence from the articulatory observations; 2) inferring the spectral trajectories from both the decoded state sequence and the articulatory observations. The proposed technique is compared to our previous approach, in which only the decoded state sequence was used for the inference of the spectral trajectories, independently from the articulatory observations. Both objective and perceptual evaluations show that this new approach leads to a better estimation of the spectral trajectories.

Index Terms: silent speech interface, handicap, HMM-based speech synthesis, audiovisual speech processing

1. Introduction

In the past few years, the design of silent speech interfaces (SSI) has emerged as a new field in the speech research community [1]. SSI may be defined as automatic systems enabling oral communication without the necessity of vocalizing the speech sound. Application areas are in the medical field, as an aid for laryngectomized patients, and in the telecommunication sector, in the form of a "silent telephone", which could be used for confidential or furtive communication, or in very noisy environments. To date, several technologies have been proposed to capture the articulatory activity (or the very low acoustic activity) during silent speech: surface electromyography (sEMG) [2]; tissue-conducted microphone (also called NAM microphone) [3]; and permanent-magnetic articulography (PEMA) [4]. In our approach, articulatory movements are captured by a multimodal imaging system composed of an ultrasound transducer placed beneath the chin and a video camera placed in front of the lips [5].

In this paper, we address the problem of "articulatory-to-acoustic" mapping, *i.e.* the synthesis of an audible speech signal, from (visual) articulatory data only. In our previous work, this problem has been addressed using non-linear regression techniques based respectively on artificial neural networks (ANN) and Gaussian mixture models (GMM). In [6], we proposed an HMM-based approach which allows the introduction of external a priori linguistic information in the mapping process. The mapping was achieved in two steps: 1) a "phonetic decoding" step during which the most likely phonetic sequence was predicted from the articulatory observations; and 2) a "synthesis" step during which spectral trajectories were estimated from the predicted phonetic sequence and the decoded HMM state sequence, using the MLPG algorithm [7]. Unlike GMM and ANN-based approach, the mapping here was achieved not at the frame level, but at the phone level. External linguistic constraints could thus be introduced in the mapping via a limitation on the authorized vocabulary (as in [6]) or by using a statistical language model (as in [8]). A HMM-based approach outperforms ANN-based and GMM-based approaches: the use of linguistic constraints helps to recover missing information in the articulatory data, such as the voicing characteristic of course, but also the position of some articulators like the velum. However, this approach presents a major drawback: the spectral trajectories are estimated only from the decoding phonetic sequence, independently of the articulatory observation. As a consequence, the quality of the synthesis depends exclusively on the accuracy of the decoding phonetic sequence: an error during the decoding stage corrupts necessarily the synthesis.

This paper focuses on this issue and investigates a new approach to estimate the spectral feature trajectories from both the decoded phonetic sequence and the articulatory observations. To do so, we adapted the approach originally proposed by Toda in [9] for GMM-based mapping to the framework of HMM-based mapping. An almost identical approach has been proposed by Zen in [10] for voice conversion and acoustic-to-articulatory mapping. The proposed approach is referred in this paper as the "continuous HMM-based mapping technique".

In this approach, the dependency between the articulatory and the acoustic variables is learned explicitly by jointly modeling sequences of articulatory and spectral features, for each phonetic class, with a "full-covariance" HMM (i.e. HMM for which the emission probability density functions (pdf) are modeled by multivariate Gaussian distributions with full covariance matrices). Spectral trajectories are estimated using a ML-based parameter estimation algorithm, which explicitly adjusts the spectral targets from the articulatory observations.

The article is organized as follows. Section 2 details the theoretical aspects of the continuous HMM-based mapping technique. Section 3 describes the data acquisition protocol, the feature extraction process, and details the practical implementation of the two mapping techniques. Experimental results are presented and discussed in section 4. Conclusions and perspectives are presented in the last section.

2. HMM-based feature mapping

Sequences of articulatory and spectral feature vectors, $\mathbf{x}$ and $\mathbf{y}$, are written as: $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T]$ and $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_t, ..., \mathbf{y}_T]$, where $\mathbf{x}_t$ and $\mathbf{y}_t$, are D_x/D_y dimensional vectors of articulatory/spectral features observed at the time t (T is the sequence length). As usual in HMM-based parameter estimation, spectral features are augmented with their first derivatives, such as $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_T]$ with $\mathbf{Y}_t = [\mathbf{y}_t, \Delta\mathbf{y}_t]$.

2.1. Baseline mapping technique

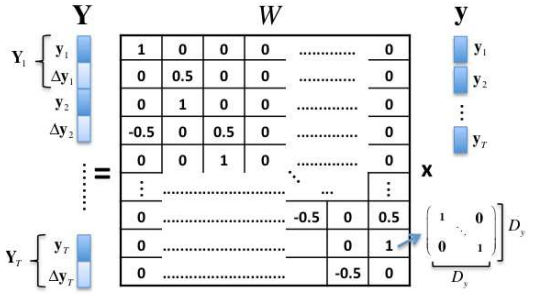
The following section briefly recalls the theoretical aspects of the HMM-based mapping technique introduced in [6], which is referred in this paper as the "baseline technique". In the training stage, streams of articulatory and spectral feature vectors (recorded synchronously) are modeled, for each phonetic class, by a multistream HMM. For each stream, the emission probability density of each state is modeled by a multivariate Gaussian distribution with diagonal covariance matrix. In the mapping stage, the sequence of spectral feature vectors $\hat{\mathbf{y}}$ is estimated from the sequence of articulatory feature vectors $\mathbf{x}$ such as $\hat{\mathbf{y}} = \arg\max_y \{p(\mathbf{y} \mid \mathbf{x})\}$ with:

$$p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y} \mid \lambda, q) \cdot P(\lambda, q \mid \mathbf{x}) \quad (1)$$

where λ is the parameters set of the HMM and q the HMM state sequence. In our implementation, $\hat{\mathbf{y}}$ is obtained by maximizing separately the two conditional probability terms of Equation 1: (1) by estimating $(\hat{\lambda}, \hat{q})$ with $(\hat{\lambda}, \hat{q}) = \arg\max_{\lambda, q} \{P(\lambda, q \mid \mathbf{x})\}$ using the Viterbi algorithm (phonetic decoding stage); and (2), by estimating $\hat{\mathbf{y}}$ such as $\hat{\mathbf{y}} = \arg\max_y \{P(\mathbf{y} \mid \hat{\lambda}, \hat{q})\}$, using the MLPG algorithm [7] (synthesis stage). This algorithm estimates the feature trajectories by solving the following equation:

$$\hat{\mathbf{y}} = \left(W^T \Sigma_{\hat{q}}^{-1} W\right)^{-1} W^T \Sigma_{\hat{q}}^{-1} M_{\hat{q}} \quad (2)$$
$$\text{with } M_{\hat{q}} = [\mu_{\hat{q}_1}, ..., \mu_{\hat{q}_T}] \text{ and } \Sigma_{\hat{q}}^{-1} = diag[\Sigma_{\hat{q}_1}^{-1}, ..., \Sigma_{\hat{q}_T}^{-1}]$$

where $\hat{q} = [\hat{q}_1, ..., \hat{q}_T]$ is the decoded HMM state sequence, μ_k and Σ_k are respectively the mean and the diagonal covariance matrix of the Gaussian emission probability density associated with state k. W is a $[2D_xT\text{-by-}D_yT]$ matrix representing the relationship between static and dynamic feature vectors:



(3)

2.2. Continuous HMM-based mapping

The new approach aims at modeling more explicitly the local correlations between articulatory and spectral features. For that purpose, sequences of articulatory and spectral features are modeled jointly, for each phonetic class, by a single-stream "full-covariance" HMM: the joint probability density function (pdf) of articulatory and spectral observations is modeled, for each HMM state q, by a single Gaussian, with:

$$p_q(\mathbf{z}) = N(\mathbf{z}, \mu_q, \Sigma_q) \text{ with } \mathbf{z} = [\mathbf{x}, \mathbf{Y}]$$
$$\Sigma_q = \begin{bmatrix} \Sigma_q^{\mathbf{xx}} & \Sigma_q^{\mathbf{xY}} \\ \Sigma_q^{\mathbf{Yx}} & \Sigma_q^{\mathbf{YY}} \end{bmatrix} \text{ and } \mu_q = \begin{bmatrix} \mu_q^{\mathbf{x}} \\ \mu_q^{\mathbf{Y}} \end{bmatrix} \quad (4)$$

where $N(., \mu, \Sigma)$ is a normal distribution with mean μ and covariance matrix Σ. Similarly to the baseline technique, the mapping starts with the phonetic decoding stage, which determines the most likely phonetic sequence, and the corresponding HMM state sequence $(\hat{\lambda}, \hat{q})$, from the articulatory observations $\mathbf{x}$. Unlike the baseline technique, the sequence of spectral feature vectors is estimated by taking into account, not only the decoded HMM states, but also the articulatory observations, such as $\hat{\mathbf{y}} = \arg\max_y \{p(\mathbf{y} \mid \mathbf{x}, \hat{\lambda}, \hat{q})\}$. For each frame t, a conditional pdf $p(\mathbf{Y}_t \mid \mathbf{x}_t, \hat{q}_t, \hat{\lambda})$ is derived from the joint pdf $p_{\hat{q}_t}(\mathbf{x}_t, \mathbf{Y}_t)$, estimated during training:

$$p(\mathbf{Y}_t \mid \mathbf{x}_t, \hat{q}_t, \hat{\lambda}) = N(\mathbf{Y}_t, E_{\hat{q}_t, t}^{\mathbf{Y}}, D_{\hat{q}_t}^{\mathbf{Y}})$$
$$\text{with } \begin{cases} E_{\hat{q}_t, t}^{\mathbf{Y}} = \mu_{\hat{q}_t}^{\mathbf{Y}} + \Sigma_{\hat{q}_t}^{\mathbf{Yx}} \Sigma_{\hat{q}_t}^{\mathbf{xx}^{-1}} (\mathbf{x}_t - \mu_{\hat{q}_t}^{\mathbf{x}}) \\ D_{\hat{q}_t}^{\mathbf{Y}} = \Sigma_{\hat{q}_t}^{\mathbf{YY}} - \Sigma_{\hat{q}_t}^{\mathbf{Yx}} \Sigma_{\hat{q}_t}^{\mathbf{xx}^{-1}} \Sigma_{\hat{q}_t}^{\mathbf{xY}} \end{cases} \quad (5)$$

(the mathematical basis of this derivation can be found in [11], p.337). As shown in Equation 5, the target vector of spectral features $E_{\hat{q}_t, t}^{\mathbf{Y}}$, is expressed as a linear function of the articulatory observation $\mathbf{x}_t$, and is based on the "local" correlations between the articulatory and the spectral features for state $\hat{q}_t$, estimated during training. Spectral trajectories $\hat{\mathbf{y}}$ are finally estimated by solving the following equation:

$$\hat{\mathbf{y}} = \left(W^T D_{\hat{q}}^{-1} W\right)^{-1} W^T D_{\hat{q}}^{-1} E_{\hat{q}} \quad (6)$$
$$\text{with } E_{\hat{q}} = [E_{\hat{q}_1 1}, ..., E_{\hat{q}_T T}] \text{ and } D_{\hat{q}}^{-1} = diag[D_{\hat{q}_1}^{-1}, ..., D_{\hat{q}_T}^{-1}]$$

which can be seen as an adaptation of Equation 2, to the problem of HMM-based feature mapping. As with the MLPG algorithm, this method determines the vector sequence that maximizes the likelihood of the model with respect to a continuity constraint on the predicted feature trajectories.

3. Experimental protocol

3.1. Data acquisition

The two mapping techniques described in section 2 are evaluated on a continuous speech database consisting of one-hour of high-speed ultrasound and video sequences, recorded synchronously with the audio signal. Data were acquired using the *Ultraspeech* acquisition system (http://www.ultraspeech.com) [11]. Ultrasound and video streams were both recorded at a frame rate of 60 frames per second, the audio signal was recorded at 16 kHz (16 bits). A female native English speaker was asked to pronounce the 1132 sentences of CMU ARCTIC corpus [13]. Acquisition was split into 10 sessions, spaced in time. An inter-session re-calibration mechanism (detailed in [11]), was used to maintain the positioning accuracy of the sensors across all sessions. A typical pair of ultrasound and video images, extracted from the recorded database, is shown in figure 1.

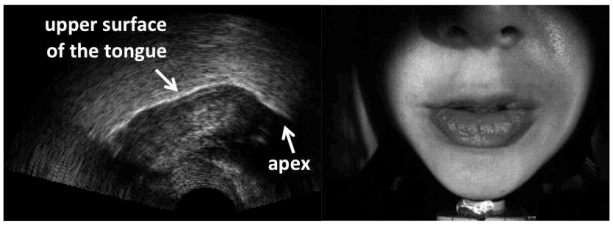


Figure 1: *Ultrasound and video images recorded with the Ultraspeech system.*

3.2. Feature extraction

The EigenTongues/EigenLips decomposition technique [14] was used to encode each ultrasound/video frame. First, regions of interest (ROI), selected in ultrasound and video images, were resized to 64x64 pixels. Sets of EigenTongues/EigenLips were calculated by performing a Principal Component Analysis on a phonetically balanced subset of frames. Each ultrasound/video image was projected onto the set of EigenTongues/EigenLips. The number of projections used for coding was determined by keeping the eigenvectors carrying at least 80% of the variance of the training set; 30 coefficients were used as static features for each visual stream. In order to be compatible with the speech analysis rate, the EigenTongues/EigenLips feature sequences were oversampled from 60 Hz to 100 Hz. Finally, they were concatenated with their first derivatives in a single articulatory feature vector (120-dimensional vector). In order to make tractable the training of full-covariance HMMs, the dimensionality of articulatory feature vectors was reduced to 30, using Locality Preserving Projection technique (LPP) [15]. Dimensionality reduction was performed only for the continuous mapping technique, since it did not lead to any improvement for the baseline technique.

The spectral content of the audio speech signal was parameterized by 25 mel-cepstrum coefficients (Blackman window, 25ms frame length, 10 ms frame shift). Static spectral features were concatenated with their first derivatives in a single spectral feature vector (50-dimensional vector). Silence frames were removed automatically using a threshold-based silence detection method, at the beginning and end of each recorded sentence.

3.3. Training

For the baseline technique, sequences of articulatory and acoustic feature vectors were modeled by a multistream HMM, for each of the 40 phonetic classes (with diagonal covariance matrix). Two streams were dedicated to the modeling of the visual features (ultrasound/video), one stream was used to model the spectral features. HMMs were first trained separately, using the Baum-Welch algorithm and then processed simultaneously, using an embedded training strategy. Context-dependency was then introduced in the modeling to take into account context effects such as co-articulation and anticipation (triphone modeling). A tree-based state-tying strategy was used to address the problem of data sparsity. Each resulting multistream HMM was then split into two distinct HMMs: a 2-streams "articulatory HMM" (ultrasound/video), used for the recognition stage, and a 1-stream "acoustic HMM", used for the synthesis stage. Articulatory HMMs were finally refined by increasing incrementally the number of Gaussian mixture components.

For the continuous HMM-based mapping technique, sequences of articulatory and acoustic feature vectors were modeled, for each of the 40 phonetic classes, by a single-stream "full-covariance" HMM. Due to the lack of training data, the training of context-dependent full-covariance HMMs on this database was found to be not feasible. As a consequence, we use the context-dependent HMMs, trained for the baseline technique, for the phonetic decoding stage; the context-independent full-covariance HMMs being used only for the synthesis stage (the target sequence of HMM states was obtained using the results of the phonetic decoding stage, and a forced-alignment procedure).

4. Results & Discussion

In the two HMM-based mapping techniques, linguistic constraints can be introduced to help the phonetic decoding. With that in mind, we implemented two decoding scenarios. In the first, considered "unconstrained", the structure of the decoding network was a simple loop in which all phones loop back to each other. In the second, or "constrained" scenario, the decoding network allows all possible word combinations which can be built from a 3k word dictionary. No statistical language model was used in the present study. The first 1110 sentences of the recorded database were divided into 37 lists of 30 sentences. A K-fold validation (leave-one-out) technique was employed for evaluation: each list was used once as the test set while the other 34 lists composed the training set. Two test lists were excluded from this procedure to be used as a validation set for the determination of two hyperparameters: (1) the optimal number of Gaussians for the articulatory HMM used for the decoding stage (which was found to be 4); and (2), the model insertion penalty (which was found to be respectively -20 and -150 for the unconstrained and constrained scenario). The performance of the decoding stage was measured by evaluating the *recognition accuracy* defined as $Acc = 100.(N - D - S - I) / N$, where N is the total number of phones in the test set, S, D and I are respectively the number of substitution, deletion, and insertion errors. The recognition accuracy was found to be 68.4% for the unconstrained scenario and 78.3 % for the constrained scenario.

The quality of the estimated spectral trajectories was first evaluated by calculating the Mel-cepstral distance (MCD) between the target and the predicted mel-cepstrum coefficients,

defined as: $MCD_s[dB] = (10 / \ln 10)\sqrt{2.\sum_{d=s}^{24}(\hat{m}_d - m_d)^2}$.

If *s=0*, the distance includes the 0th cepstral dimension which corresponds to overall signal power. In this paper, we focus on the value of MCD_1 since we are interested more in the shape of the target spectral envelope, than in the intensity variation of the synthetic speech sound. Results are presented in Table 1.

Table 1. *Objective performance evaluation (MCD_s[dB]).*

Scenario	Baseline HMM-based mapping	Continuous HMM-based mapping
Unconstrained (Acc=68.4%)	**MCD_1 = 6.01** (MCD_0 = 8.35)	**MCD_1 = 5.68** (MCD_0 = 7.8)
Constrained (Acc=78.3%)	**MCD_1 = 5.97** (MCD_0 = 8.30)	**MCD_1 = 5.60** (MCD_0 = 7.76)
Forced-alignment (Acc=100%)	**MCD_1 = 5.76** (MCD_0 = 7.86)	**MCD_1 = 5.46** (MCD_0 = 7.4)

The continuous HMM-based mapping technique leads to an average improvement of 0.33 dB for MCD_1 (and 0.51 dB for MCD_0). Paired-sample *t*-tests showed that this improvements was statistically significant, for each of the two decoding scenarios (and for both MCD_0 and MCD_1, with $p<0.001$). The continuous HMM-based mapping is also slightly less sensitive to decoding errors: the degradation of the performance between the "forced-alignment" scenario (for which the phonetic target is given, i.e. Acc=100%) and the unconstrained scenario (Acc=68.4%) is 3.9% for the continuous mapping technique, whereas it is 4.2% for the baseline technique.

In order to confirm the objective evaluation conclusions, a perceptual comparison of the two mapping techniques was performed using a XAB listening test. 15 sentences were randomly selected from the test corpus. For each sentence, 3 audio stimuli (named X, A and B) were synthesized using the STRAIGHT vocoder [16]. The target speech sound X was built by analyzing and (re)-synthesizing the original audio signal. The spectral content of stimuli A and B was estimated from the articulatory observations using either the baseline, or the continuous mapping technique. The constrained scenario (Acc=78.3%) was used for the phonetic decoding stage. In order to evaluate only the accuracy of the derived spectral trajectories, excitation characteristics of the target sound X (pitch, aperiodic component and energy) were used for the synthesis of A and B (so that A,B and X share the same prosodic content). 10 listeners were asked to say which of the sounds A or B was the most similar to X (A and B were presented in a random order). In 80% (σ=9%) of the cases, the listeners chose the stimuli synthesized with the continuous HMM-based mapping technique (inter-listener agreement (Fleiss' Kappa coefficient)=0.53±0.02).

5. Conclusions and Perspectives

The article introduces a new approach to estimate spectral feature trajectories from ultrasound and video articulatory data, for a silent speech interface application. We describe a parameter generation algorithm which explicitly takes into account the local dependencies between the articulatory and the spectral features, modeled by a set of full-covariance HMM. Both objective and perceptual evaluations shows that this technique outperforms our previous approach, in which the parameter generation were driven only by the decoded HMM state sequence, independently from the articulatory observations.

Future work will focus on the real-time implementation of the continuous HMM-based mapping technique. The adaption of low-delay feature mapping techniques [17] will be investigated.

6. Acknowledgements

This work is supported by the 6th Christian Benoit Award (ISCA/AVISA/AFCP/ACB). The authors would like to acknowledge useful discussions with Laurent Beneroya, Atef Ben-Youssef, Pierre Badin, Gérard Chollet, and Tomoki Toda.

7. References

[1] Denby, B., Schultz, T., Honda, K., Hueber, T., et al., "Silent Speech Interfaces," Speech Communication, vol. 52, no. 4, pp. 270-287, 2010.

[2] Schultz, T., Wand, M., "Modeling coarticulation in EMG-based continuous speech recognition", Speech Communication, vol. 52, no. 4, pp. 341-353, 2010.

[3] Tran, V.-A., Bailly, G., Loevenbruck, H., Toda, T., "Improvement to a NAM-captured whisper-to-speech system", Speech Communication, vol. 52, no. 4, pp. 314-326, 2010.

[4] Gilbert, JM., Rybchenko, SI., Hofe, R., et. al. "Isolated word recognition of silent speech using magnetic implants and sensors", Med. Eng. and Physics, vol. 32, no. 10, pp. 1189-1197, 2010.

[5] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, 52(4), pp. 288-300, 2010.

[6] Hueber, T., Benaroya, E.L, Denby, B., Chollet, G., "Statistical Mapping between Articulatory and Acoustic Data for an Ultrasound-based Silent Speech Interface", in Proc. of Interspeech, pp. 593-596, Firenze, Italia, 2011.

[7] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", in Proc. of ICASSP, pp. 1315-1318, 2000.

[8] Cai, J., Hueber, T. Denby, B., et al. "A Visual Speech Recognition System for an Ultrasound-based Silent Speech Interface", in Proc. of ICPhS, pp. 384-387, 2011.

[9] Toda, T., Black, A.W., Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model", Speech Comm., vol. 50, no. 3, pp. 215-227, 2008.

[10] Zen, H., Nankaku, Y., Tokuda, K., "Continuous Stochastic Feature Mapping Based on Trajectory HMMs", IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 2, pp. 417-430, 2011.

[11] M. Kay., S, "Fundamentals of Statistical Signal Processing: Estimation Theory", Prentice Hall, 1993.

[12] Hueber, T., Chollet, G., Denby, B., Stone, M., "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application", in Proc. of ISSP, pp. 365-369, 2008.

[13] Kominek, J. and Black, A., "The CMU Arctic speech databases", in Proc. of the 5th Speech Synthesis Workshop, pp 223-224, 2004.

[14] Hueber, T., Aversano, G., Chollet, G., Denby, B., et al. "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface", in Proc. of ICASSP, pp. I1245-I1248, 2007.

[15] Feng, G., Hu, F., Zhou, Z., "A direct locality preserving projections (DLPP) algorithm for image recognition", Neural Processing Letters, vol. 27, no. 3, pp.1370-4621, 2008.

[16] Kawahara, H., Morise, M., et al., "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation", in Proc. of ICASSP, pp. 3933-3936, 2008.

[17] Muramatsu, T., Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory", in Proc. of Interspeech, pp. 1076-1079, 2008.