Cross-speaker Acoustic-to-Articulatory Inversion using Phone-based Trajectory HMM for Pronunciation Training

Thomas Hueber[1], *Atef Ben-Youssef*[1], *Gérard Bailly*[1], *Pierre Badin*[1], *Frederic Elisei*[1]

[1]GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France

(firstname.lastname)@gipsa-lab.grenoble-inp.fr

Abstract

The article presents a statistical mapping approach for cross-speaker acoustic-to-articulatory inversion. The goal is to estimate the most likely articulatory trajectories for a *reference speaker* from the speech audio signal of another speaker. This approach is developed in the framework of our system of *visual articulatory feedback* developed for computer-assisted pronunciation training applications (CAPT). The proposed technique is based on the joint modeling of articulatory and acoustic features, for each phonetic class, using full-covariance trajectory HMM. The acoustic-to-articulatory inversion is achieved in 2 steps: 1) finding the most likely HMM state sequence from the acoustic observations; 2) inferring the articulatory trajectories from both the decoded state sequence and the acoustic observations. The problem of speaker adaptation is addressed using a voice conversion approach, based on trajectory GMM.

Index Terms: acoustic-to-articulatory inversion, intelligent tutoring systems, pronunciation training, trajectory HMM, voice conversion, talking head

1. Introduction

Several studies tend to show that a visual feedback of the articulatory movements facilitates pronunciation training [1]. Different approaches have been proposed in the literature to provide this visual feedback. In [2], tongue-palate contact points are monitored using electro-palatography; derived visual patterns help the patient to improve the positioning of his/her tongue in velar and alveolar regions. In [3] and [4] (Ultrax project), ultrasound imaging is used to provide a real-time visual feedback of tongue movements. Another approach consists in using an “augmented” talking head, i.e. a talking head displaying all speech articulators including usually hidden ones like the tongue. In [6], Massaro et. al. proposed to use an augmented talking head as a language tutor: the talking head displays pre-calculated animations which help the user to visualize the articulatory movements that correspond to a specific speech sound. However, this approach does not provide the learner with a real feedback on his/her own articulation. In [5], Engvall proposed to use a wizard-of-Oz approach to provide this feedback: a human listener evaluates the user's pronunciation and animates the talking head from a set of pre-generated sequences. In our previous work [7], we describe a system of visual articulatory feedback, based on the 3D augmented talking head developed at GIPSA-lab [8]. This system aims to provide any speaker with a real feedback on his/her own articulation. In our approach, the talking head is animated *automatically* from the audio speech signal, using acoustic-to-articulatory inversion.

The problem of acoustic-to-articulatory inversion has been addressed in many studies, using either codebook-based approaches, as in [9], or statistical regression techniques, as in [10], [11], [12], [13] (based respectively on ANN, SVM, GMM and HMM). However, only a few studies ([15], [16]) address the problem of cross-speaker acoustic-articulatory inversion, which consists in recovering the most likely articulatory trajectories of a *reference speaker* from the speech audio signal of another speaker. This is a critical issue for the design of a multi-speaker system of visual articulatory feedback. In [7], we proposed an HMM-based approach to address this issue. In the training stage, sequences of acoustic and articulatory data of the reference speaker were modeled, for each phonetic class, by a 2-stream context-dependent HMM. The acoustic-to-articulatory mapping was achieved in two steps: 1) a “phonetic decoding” step during which the most likely phonetic sequence was predicted from the acoustic observations; and 2) a “synthesis” step during which the articulatory trajectories were estimated from the decoded phonetic target and the corresponding HMM state sequence, using the MLPG algorithm [14]. The speaker adaptation problem was addressed by adapting the *acoustic stream* of each HMM, using the MLLR technique and a small amount on (audio-only) adaptation data. This technique, called in this paper the *baseline technique,* gave encouraging results but presents a major drawback: during the synthesis stage, the articulatory trajectories are estimated only from the decoding phonetic sequence, independently of the acoustic observations. As a consequence, the estimated articulatory trajectories depend almost exclusively on the decoded phonetic label. They do not reflect the acoustic variability that exists within each phonetic class (nevertheless, in the baseline technique, this variability is partly captured by introducing context-dependency in the modeling). This ability to take explicitly into account the acoustic observations during the generation of the articulatory trajectories is a critical step for pronunciation training applications.

This paper addresses this specific issue and investigates a new approach to estimate the articulatory trajectories from both the decoded phonetic sequence and the acoustic observations. To do so, we adapted the GMM-based mapping approach proposed by Toda in [12] to the problem of HMM-based feature mapping. An almost identical adaptation has been proposed by Zen in [17]. The proposed approach is called in this paper the *continuous mapping technique*. Speaker adaptation is here addressed using a voice conversion approach, based on trajectory GMM [18].

The article is organized as follows. Section 2 details the theoretical aspects of both the baseline and the proposed techniques. Section 3 describes the data acquisition protocol and details the practical implementation of the 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 acoustic (spectral) and articulatory 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 acoustic/articulatory features observed at the time t (T is the sequence length). As usual in the framework of trajectory HMM, target 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 baseline mapping technique [7]. In the training stage, streams of acoustic and articulatory 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 articulatory feature vectors $\hat{\mathbf{y}}$ is estimated from the sequence of acoustic feature vectors $\mathbf{x}$ such as $\hat{\mathbf{y}} = \arg\max_y \{p(\mathbf{y} \mid \mathbf{x})\}$ with:

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

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 [14] (synthesis stage).

2.2. Continuous mapping technique

The continuous mapping technique aims at modeling more explicitly the articulatory-acoustic local dependencies. For that purpose, sequences of acoustic and articulatory features are modeled jointly, for each phonetic class, by a single-stream "full-covariance" HMM. For each HMM state q, the joint probability density function (pdf) of acoustic and articulatory features is modeled by:

$$p_q(\mathbf{z}) = N(\mathbf{z}, \mu_q, \Sigma_q) \text{ with } \mathbf{z} = [\mathbf{x}, \mathbf{Y}]$$
$$\Sigma_q = \begin{bmatrix} \Sigma^{\mathbf{xx}}_q & \Sigma^{\mathbf{xY}}_q \\ \Sigma^{\mathbf{Yx}}_q & \Sigma^{\mathbf{YY}}_q \end{bmatrix} \text{ and } \mu_q = \begin{bmatrix} \mu^{\mathbf{x}}_q \\ \mu^{\mathbf{Y}}_q \end{bmatrix} \qquad (2)$$

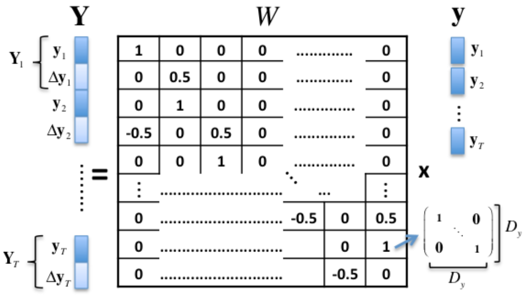
where $N(., \mu, \Sigma)$ is a normal distribution with mean μ and covariance matrix Σ. Similarly to the baseline technique, the mapping starts with the phonetic decoding stage, which determines the most likely phonetic sequence, and the corresponding HMM state sequence $(\hat{\lambda}, \hat{q})$, from the acoustic observations $\mathbf{x}$. Unlike the baseline technique, the articulatory trajectories are estimated by taking into account, not only the decoded HMM states, but also the acoustic 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:

$$p(\mathbf{Y}_t \mid \mathbf{x}_t, \hat{q}_t, \hat{\lambda}) = N(\mathbf{Y}_t, E^{\mathbf{Y}}_{\hat{q}_t,t}, D^{\mathbf{Y}}_{\hat{q}_t})$$
$$\text{with} \begin{cases} E^{\mathbf{Y}}_{\hat{q}_t,t} = \mu^{\mathbf{Y}}_{\hat{q}_t} + \Sigma^{\mathbf{Yx}}_{\hat{q}_t} \Sigma^{\mathbf{xx}^{-1}}_{\hat{q}_t} (\mathbf{x}_t - \mu^{\mathbf{x}}_{\hat{q}_t}) \\ D^{\mathbf{Y}}_{\hat{q}_t} = \Sigma^{\mathbf{YY}}_{\hat{q}_t} - \Sigma^{\mathbf{Yx}}_{\hat{q}_t} \Sigma^{\mathbf{xx}^{-1}}_{\hat{q}_t} \Sigma^{\mathbf{xY}}_{\hat{q}_t} \end{cases} \qquad (3)$$

(the mathematical basis of this derivation can be found in [19], p.337). As shown in Equation 3, the target vector of articulatory features $E^{\mathbf{Y}}_{\hat{q}_t,t}$ is expressed as a linear function of the acoustic observation $\mathbf{x}_t$, and is based on the local correlations between the articulatory and the acoustic features for state $\hat{q}_t$, estimated during training. Spectral trajectories $\hat{\mathbf{y}}$ are finally estimated by solving the following equation:

$$\hat{\mathbf{y}} = \left(W^T D^{-1}_{\hat{q}} W\right)^{-1} W^T D^{-1}_{\hat{q}} E_{\hat{q}}$$
$$\text{with } E_{\hat{q}} = [E_{\hat{q}_1 1}, ..., E_{\hat{q}_T T}] \text{ and } D^{-1}_{\hat{q}} = diag[D^{-1}_{\hat{q}_1}, ..., D^{-1}_{\hat{q}_T}] \qquad (4)$$

where $\hat{q} = [\hat{q}_1, .., \hat{q}_T]$ is the decoded HMM state sequence and W is a *[2D$_x$T-by-D$_y$T]* matrix representing the relationship between static and dynamic feature vectors:



(5)

Like the MLPG algorithm, this method determines the sequence of feature vectors that maximizes the likelihood of the model with respect to a continuity constraint on the predicted feature trajectories.

2.3. Speaker adaptation

In the baseline technique, speaker adaptation is achieved by adjusting the parameters of the acoustic stream of each HMM, using the MLLR technique [20]. MLLR estimates linear transformations for models parameters to maximize the likelihood of the adaptation data. However, this approach can not be used for the continuous mapping technique: since it is based on full-covariance HMM, it would require the recording of both audio and articulatory adaptation data, which is not acceptable in the envisioned application. To overcome this issue, we propose to adapt the acoustic observations rather than the model parameters. In that purpose, we investigate the use of voice conversion as a speaker adaptation technique. The goal is to modify the acoustic (spectral) observations of the *source speaker* (i.e. the system user) so that it sounds as if it had been pronounced by the *reference speaker*. The resulting cross-speaker acoustic-to-articulatory inversion technique can be formulated as follows:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}}\{p(\mathbf{y} \mid \mathbf{x}_{source})\} \text{ with}$$
$$p(\mathbf{y} \mid \mathbf{x}_{source}) = p(\mathbf{y} \mid \mathbf{x}_{ref}) \cdot p(\mathbf{x}_{ref} \mid \mathbf{x}_{source}) \qquad (7)$$

where $\mathbf{x}_{source}$ is a sequence of acoustic feature vectors of the source speaker and $\mathbf{x}_{ref}$ the corresponding sequence of converted feature vectors. In our implementation, the "inversion term" $p(\mathbf{y} \mid \mathbf{x}_{ref})$ and the "adaptation term" $p(\mathbf{x}_{ref} \mid \mathbf{x}_{target})$ are maximized separately. For the voice conversion step, we use the approach proposed by Toda in [18] which is based on trajectory GMM. In the adaptation stage, the joint pdf of (time-aligned) acoustic observations for the source and the reference speakers is modeled with:

$$p(\mathbf{z}) = \sum_{q=1}^{Q} N(\mathbf{z},\mu_q,\Sigma_q) \text{ with } \mathbf{z} = [\mathbf{x}_{ref}\mathbf{x}_{source}] \qquad (8)$$

using similar notations as in Equation 2. In the mapping stage, the suboptimum sequence of mixture component $\hat{q}$ defined as $\hat{q} = \arg\max_{q}\{P(q \mid \mathbf{x}_{source},\eta\}$, is determined using the Viterbi algorithm (η is the GMM parameter set). The sequence of converted feature vectors is finally estimated from $(\hat{q},\eta)$, using the inference technique described by Equations 3 and 4.

3. Experimental protocol

3.1. Data acquisition and Feature Extraction

Articulatory data of the reference speaker "PB" were recorded synchronously with the audio signal using the Carstens 2D EMA system (AG200). Six coils were glued on the tongue tip, blade, and dorsum, as well as on the upper lip, the lower lip and the jaw. The recorded database consists of two repetitions of 224 VCVs, two repetitions of 109 pairs of CVC real French words, and 88 sentences. The database consisted of approximately 17 minutes of speech, long pauses being excluded. In order to evaluate the speaker adaptation technique, a second database of audio data only, was recorded using a second speaker named TH. This speaker was asked to pronounce the same text material as described above.

Sequence of articulatory features (i.e. x and y coordinates of each EMA coils) were downsampled from 200 Hz to 100 Hz and low-pass filtered at 20 Hz. The audio speech signal was parameterized by 25 mel-cepstrum coefficients (Blackman window, 25 frame length, 10 ms frame shift).

3.2. Training

For the baseline technique, sequences of articulatory and acoustic feature vectors were modeled by a set of 2-stream context-dependent HMM (right biphone), using first the maximum-likelihood criterion (ML) and then the minimum generation error criterion (MGE), as reported in [7]). 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 (one-stream): an "acoustic HMM" used for the phonetic decoding stage, and an "articulatory HMM", used for the synthesis stage. Acoustic HMMs were finally refined by increasing incrementally the number of Gaussian mixture components. For the cross-experiments, acoustic HMMs were adapted by (1) aligning the audio adaptation data (recorded by speaker TH) at the phonetic level, and (2), updating the model parameters using the MLLR technique in order to maximize the likelihood on the adaptation data.

For the continuous HMM-based mapping technique, sequences of articulatory and acoustic feature vectors were modeled, for each of the 30 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). Also, MGE criterion was not used for to train the full covariance HMMs, since it did not lead to any improvement (compared to the ML criterion).

3.3. Evaluation

The articulatory-acoustic database (recorded by the reference speaker PB) was divided into 5 partitions of equal size. A 5-fold cross-validation technique was employed for evaluation: each list was used once as the test set while the other 4 lists composed the training set. A few utterances were excluded from the training set and were used as a validation subset for the determination of some hyperparameters: (1) the optimal number of Gaussians for the acoustic HMM used for the decoding step (which was found to be 8); (2), the model insertion penalty (which was found to be -20); and (3), the optimal number of mixture components for the GMM-based voice conversion step (which was found to be 64). For the cross-speaker experiments, 1/5 of the database (~3mn) recorded by speaker TH was used as adaptation dataset, for both the MLLR-based approach and the GMM-based approach (in that case, sequences of acoustic feature vectors for source (TH) and target speakers (PB), were time-aligned using dynamic time warping).

For the phonetic stage, the structure of the decoding network was a simple loop in which all phones loop back to each other. The performance of the decoding stage was measured by evaluating the *recognition accuracy* defined as $Acc_{audio} = 100 \cdot (N - S - D - I) / N$, 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.

For the synthesis stage, the accuracy of the estimated articulatory trajectories was measured by calculating, for each partition, the root mean square error between the measured and the estimated EMA parameters, such as:

$$\mu RMS = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_{d,t} - y_{d,t})^2} \qquad (5)$$

where *T* is the number of frames in the test set, *D* is the number of EMA parameters (*D=12*), $y_{d,t}$ and $\hat{y}_{d,t}$ are respectively the estimated and the measured position of the *d*th EMA parameters at time *t*. Since no articulatory data was available for the speaker TH, μRMS could not be calculated for the cross-speaker experiment. Therefore, the “articulatory recognition” paradigm, introduced in [7], was used: an HMM-based phonetic decoder trained on the articulatory data of the reference speaker PB (using a standard training procedure similar to the one described at section 3.2), was used to decode the synthetic articulatory trajectory at the phonetic level. The obtained recognition accuracy, referred as Acc_{art} , was considered as a measure of the accuracy of the synthetic trajectory.

4. Results & Discussion

For the inversion experiment on the reference speaker PB, the baseline and the continuous mapping techniques gave almost identical results with: Acc_{audio}=84%, μRMS=1.48mm, and Acc_{art}=80.2%. Thus, a similar performance can be obtained with two distinct mapping strategies: (1) partitioning very finely the articulatory-acoustic space and performing the mapping at the class-level (context-dependant phonetic class), like in the baseline technique; or (2), partitioning less precisely the articulatory-acoustic space (30 phonetic class), but learning local regression functions, like in the continuous mapping technique.

For the cross-speaker inversion experiment, the MLLR adaptation approach slightly outperforms the GMM-based approach for the phonetic decoding stage, with Acc_{audio}=80% (MLLR) vs. Acc_{audio}=73% (GMM). As a consequence, the accuracy of the estimated articulatory trajectories is lower for the continuous mapping technique, with Acc_{art}= 67% vs. 77% for the baseline technique. However, when combining MLLR adaptation (for the phonetic decoding step) and GMM-based speaker adaptation (to adapt the acoustic observations for the inversion step), the performance of the continuous mapping technique is almost similar to the performance obtained with the baseline technique (i.e. Acc_{art}= 73% vs. Acc_{art}= 77%).

5. Conclusions and Perspectives

The article introduces a new approach for estimating the most likely articulatory trajectories of a *reference speaker*, given the acoustic signal of another speaker (cross-speaker acoustic-to-articulatory inversion). This approach is based on the explicit modeling of the articulatory-acoustic local correlations, using phone-based full-covariance HMMs. Speaker adaptation is performed using a voice conversion approach, based on trajectory GMM. Results obtained with the proposed approach are similar to those obtained with our previous technique, in which acoustic and articulatory observations were modeled independently, using context-dependant diagonal-covariance HMM. In future work, we intend to investigate the use of the continuous HMM-based mapping technique for both cross-speaker and cross-language acoustic-to-articulatory inversion. The objective will be to test capacity of the system to deal with speech produced by a foreign speaker and also with pathological speech.

6. Acknowledgements

This work was supported by the French National Research Agency (ANR) under contract numbers ANR-08-EMER-001-02 ARTIS and Joseph Fourier University (Vizart3D project). The authors would like to acknowledge Tomoki Toda for useful discussions and Christophe Savariaux for his help in the recording of EMA data.

7. References

[1] Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., and Hueber, T., "Visual articulatory feedback for phonetic correction in second language learning", in Proc. of SLATE workshop, P1-10, 2010.

[2] Wrench, A., Gibbon, F., McNeill, A.M., Wood, S., “An EPG therapy protocol for remediation and assessment of articulation disorders”, in Proc. of ICSLP, Denver, USA, pp. 965-968, 2002

[3] Bernhardt, B.M., Gick, B., Bacsfalvi, P., Adler-Bock, M. “Ultrasound in speech therapy with adolescents and adults”, Clinical Linguistics & Phonetics, vol. 19, pp. 605-617, 2005.

[4] Cleland, J., Scobbie, J.M. & Wrench, A., “Visual Feedback for Children with Speech Sound Disorders”, Poster presented at the 3rd Colloquium of British Association of Clinical Linguistics, 2011.

[5] Engwall, O., “Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes”, in Proc. of Interspeech, pp. 2631-2634, 2008.

[6] Massaro, D. W., Liu, Y., Chen, T. H., Perfetti, C. A. “A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learning”, in Proc. of Interspeech, Pittsburg, USA, pp. 825-828, 2006.

[7] Ben Youssef A., Hueber T., Badin P., Bailly G., "Toward a multi-speaker visual articulatory feedback system", in Proc. of Interspeech, Firenze, Italia, pp. 489-492, 2011.

[8] Badin, P., Elisei, F., Bailly, G., Tarabalka, Y., “An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data”, in 5th Conf. on Articulated Motion and Deformable Objects, Eds.: F.J. Perales & R.B. Fisher, Berlin, Heidelberg, pp. 132-143, 2008.

[9] Ouni, S., Laprie, Y., “Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion”, J. Acoustical Society of America, vol. 118, pp. 444-460, 2005.

[10] Richmond, K., “Estimating Articulatory Parameters from the Acoustic Speech Signal”, PhD thesis, CSTR Edinburgh, 2002.

[11] Toutios, A., Margaritis, K. “A support vector approach to the acoustic-to-articulatory mapping”, in Proc. of Interspeech, pp. 3221-3224, 2005.

[12] 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, 2007.

[13] Hiroya, S., Honda, M., “Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model”, IEEE Transactions on Speech and Audio Processing, vol. 12, no. 2, pp. 175-185, 2004.

[14] 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.

[15] Hiroya, S. and M. Honda, “Speaker Adaptation Method for Acoustic-to-Articulatory Inversion using an HMM-Based Speech Production Model”, IEICE Transactions On Information And Systems, E87-D(5), pp. 1071-1078, 2004.

[16] Ghosh, P., Narayanan, S., “A subject-independent acoustic-to-articulatory inversion”, in Proc. of ICASSP, pp. 4624-4627, 2011.

[17] Zen, H., Nankaku, Y., Tokuda, K., “Continuous Stochastic Feature Mapping Based on Trajectory HMMs”, IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 19, no. 2, pp. 417- 430, 2011.

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

[19] M. Kay., S, “Fundamentals of Statistical Signal Processing: Estimation Theory”, Prentice Hall, 1993.

[20] Leggetter, C. and Woodland, P., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer, Speech and Language, vol. 9, pp. 171-185, 1995.