EVALUATION OF LINEAR REGRESSION FOR SPEAKER ADAPTATION IN HMM-BASED ARTICULATORY MOVEMENTS ESTIMATION

Hao Li, Jianhua Tao, Yang Wang

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{hli, jhtao, yangwang}@nlpr.ia.ac.cn

ABSTRACT

Acoustic-to-articulatory inversion problem is usually studied in speaker-specific manner because both articulatory data and acoustic features contain speaker-specific components. This paper presents our work on speaker-adaptation training for this problem. We implement speaker adaptation in HMM-based acoustic-to-articulatory inversion mapping, and evaluate different combinatorial structures of the articulatory data and acoustic features. The HMM-based inversion mapping models are built with single-stream and multi-stream, independent clustering and shared clustering structures. The speaker adaptation is implemented in stream-independent structure and shared adaptation structure. The constrained maximum likelihood linear regression method is used for the speaker-adaptive transformation. The experimental results show that the sharing of the speaker-adaptive transformation of the articulatory feature stream and acoustic feature stream can improve the estimation accuracy in inversion mapping. The multi-stream system with shared clustering and shared adaptive transformation has the best result among all the tested structures.

Index Terms— speaker adaptation, acoustic-to-articulatory inversion, maximum likelihood linear regression

1. INTRODUCTION

Acoustic-to-articulatory inversion is considered a difficult and ill-posed problem due to its high nonlinearity and one-to-many nature. Moreover, both the speech acoustic features and the articulatory movements contain speaker-specific components, which depend on the shapes of articulators, speaking style of speakers and their gender, age etc. Many inversion methods have been proposed in both speaker-dependent manner and speaker-adaptation manner. In the speaker-dependent manner, there are methods such as codebook mapping [1], mixture density network [2], Gaussian mixture model (GMM) based mapping [3], and hidden Markov model (HMM) based methods [4-7]. In the study of the speaker-adaptation for the inversion problem, Hueber et al. [8] use an approach that merges the voice conversion step and the acoustic-articulatory inversion step; Hiroya and Honda [9] use a speaker adaptation method in HMM-based speech production model; Hiroya and Mochida [10] proposed a multi-speaker articulatory trajectory formation based on speaker–independent articulatory HMMs using adaptive training.

Among all the inversion methods, HMM-based methods have many advantages. Firstly, context information can be taken into account in this method; secondly, once the HMMs are trained, the articulatory movements can be synthesized with only the HMM state sequence, which makes it possible to synthesize articulatory movement from text without speech signal. Such methods have been realized in [11]. Moreover, the speaker adaptation method based on HMMs has been verified to be very effective, which has made it possible to adapt initial HMMs to a new speaker with a small amount of adaptation data. One important speaker adaptation technique is maximum likelihood linear regression (MLLR) [12-14], it estimates the speaker-adaptive transformation for each HMM. This method has been used in the adaptation of the articulatory movements such as in the work by Hiroya and Mochida [10]. However, the speaker adaptive training in the acoustic domain and the articulatory domain are independent in most of the previous research. The speaker-specific components of both acoustic features and articulatory features are not considered together. Therefore, we implement different model with different combinatorial structures to find out if the combination of the two feature streams can improve the performance of inversion mapping using speaker adaptation. The acoustic excitation and spectrum features and the articulatory data are used as different streams, the HMMs are built with single-stream structure and multi-stream structure. For the multi-stream structure, we use both stream-independent clustering and stream-shared clustering method. The adaptation is implemented with stream-independent adaptation manner and shared adaptation manner.

2. HMM-BASED ARTICULATORY MODEL

2.1. Single-stream structure

The single-stream model is the HMM with only a stream of articulatory data. The framework of the HMMs training for articulatory features is shown in Figure 1. Let $X=[x_1^{\mathsf{T}},x_2^{\mathsf{T}},...,x_N^{\mathsf{T}}]^{\mathsf{T}}$ denote the observed articulatory feature sequence, N is the length of the sequence. The observation feature vector for each frame consists of static articulatory features x_{St} and their velocity Δx_{St} and acceleration $\Delta^2 x_{St}$, which can be written as $x_t=[x_{St}^{\mathsf{T}},\Delta x_{St}^{\mathsf{T}},\Delta^2 x_{St}^{\mathsf{T}}]^{\mathsf{T}}$, A set of context-dependent HMMs λ are estimated to maximize the likelihood $P(X\mid\lambda)$. The training includes several stages as in the case of training an HMM-based speech synthesis system (HTS) [15]. The parameters of context-dependent HMMs are estimated by the Baum-Welch algorithm. After the initial context-dependent HMM training, a decision tree is trained using the minimum description length (MDL) [16] criterion to cluster the probability density function of all

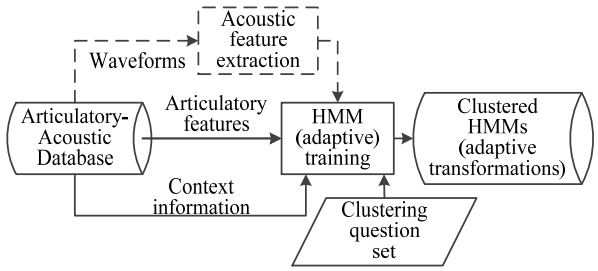

Figure 1: Flowchart of the articulatory HMMs training. The dashed part is used for multi-stream HMM training. The speaker-adaptation training gives the clustered HMMs and speaker-adaptive transformations.

HMM states. The clustering is to alleviate the data sparsity problem.

2.2. Multi-stream structure

In the multi-stream structure, acoustic features are used to train HMMs along with the articulatory features. The flowchart is also shown in Figure 1. Let $Y=[y_1^{\mathsf{T}}, y_2^{\mathsf{T}}, \cdots, y_N^{\mathsf{T}}]^{\mathsf{T}}$ denote the observed acoustic feature sequence. The observation feature vector for each frame consists of static acoustic features y_{St} and their velocity Δy_{St} and acceleration $\Delta^2 y_{St}$, which can be written as: $y_t=[y_{St}^{\mathsf{T}}, \Delta y_{St}^{\mathsf{T}}, \Delta^2 y_{St}^{\mathsf{T}}]^{\mathsf{T}}$. HMM set λ for the combined acoustic and articulatory streams are estimated to maximize the likelihood function $P(X,Y \mid \lambda)$. The two streams are independent in the observation probabilities, therefore, we can write the likelihood function as

$$\begin{aligned} P(X,Y \mid \lambda) &= \sum_{\forall q} P(X,Y,q \mid \lambda) \\ &= \sum_{\forall q} \pi_{q_1} b_{q_1}(x_1,y_1) \prod_{t=2}^{N} a_{q_{t-1}q_t} b_{q_t}(x_t,y_t) \end{aligned} \tag{1}$$

$$b_j(x_t,y_t) = b_{xj}(x_t) b_{yj}(y_t) \tag{2}$$

where $q=\{q_1,q_2,\cdots,q_N\}$ denotes the state sequence shared by the two streams; π_i denotes the initial state probability of state i and a_{ij} denotes the state transition probability from state i to j. $b_{xj}(x)$ and $b_{yj}(y)$ denote the state observation probability density function for state j of articulatory features and acoustic features, respectively, which are single Gaussian distributions with diagonal covariance matrices. The parameter estimation procedure of context-dependent HMMs for the multi-stream HMMs is the same as that of the single-stream.

2.3. Clustering structures

The decision tree is trained to cluster the probability density function of all HMM states after the initial context-dependent HMM training. We consider two structures for model clustering as described by Ling et al [11]: (1) independent clustering, cluster the acoustic model and articulatory model independently; (2) shared clustering, combine the acoustic spectrum stream and articulatory stream and build a shared decision tree to cluster their HMM states.

Note that the acoustic features include spectral parameters and excitation, specifically, the fundamental frequency (F0). Due to the nature of the two kinds of parameters, the F0 is modeled as multi-space probability distributions (MSD) [17] while the spectral parameters are modeled as single-Gaussian distributions in each HMM state. Therefore, the F0 stream is always a separate stream during the clustering whether the spectral parameter and acoustic features are shared clustering or not.

3. SPEAKER ADAPTATION

3.1. Single-stream speaker adaptation

The speaker adaptation is based on constrained maximum likelihood linear regression (MLLR) [12, 14], it computes a set of linear transformations to reduce the mismatch between the initial model set and the adaptation data. In the single-stream HMM system, the constrained MLLR transformation of the mean vector of state m is

$$\begin{bmatrix} \hat{\mu}_{X_S}^m \\ \hat{\mu}_{\Delta X_S}^m \\ \hat{\mu}_{\Delta^2 X_S}^m \end{bmatrix} = \begin{bmatrix} A_{X_S}^m & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{\Delta X_S}^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_{\Delta^2 X_S}^m \end{bmatrix} \begin{bmatrix} \mu_{X_S}^m \\ \mu_{\Delta X_S}^m \\ \mu_{\Delta^2 X_S}^m \end{bmatrix} - \begin{bmatrix} b_{X_S}^m \\ b_{\Delta X_S}^m \\ b_{\Delta^2 X_S}^m \end{bmatrix} \tag{3}$$

where μ and $\hat{\mu}$ denote the mean vector of observations before and after the transformation. The subscripts X_S, ΔX_S and $\Delta^2 X_S$ denote the static articulatory features and their velocity and acceleration, respectively. $A_{X_S}^m$, $A_{\Delta X_S}^m$ and $A_{\Delta^2 X_S}^m$ are $n_{X_S} \times n_{X_S}$ adaptive matrices for static articulatory features and their velocity and acceleration, respectively (n_{X_S} is the dimension of the static articulatory feature vector); $b_{X_S}^m$, $b_{\Delta X_S}^m$ and $b_{\Delta^2 X_S}^m$ represents $n_{X_S} \times 1$ bias vectors. The transformation of the variance is corresponding to that applied to the mean. The transform is trained by the method described in [12]. To improve the flexibility of the adaptation process, a decision tree is used to group the states in the model set.

3.2. Stream-independent adaptation

There are two structures for the speaker-adaptive transformation in the multi-stream HMM system. The first one is stream-independent adaptation. In this structure, the constrained MLLR strategy is implemented for the acoustic features and the articulatory features independently. The speaker-adaptive transformation for the articulatory features is the same as (3) and the transformation for acoustic features is also estimated by the same procedure. Two feature streams have independent speaker-adaptive transformations.

3.3. Shared adaptation

The second structure for multi-stream speaker adaptation is the shared adaptation structure. In this structure, the acoustic feature stream and the articulatory features share the same transformation matrix and bias vector sets, which have the following representation of transformation of the mean vector of state m:

$$\begin{bmatrix} \hat{\mu}_{X_S}^m \\ \hat{\mu}_{Y_S}^m \\ \hat{\mu}_{\Delta X_S}^m \\ \hat{\mu}_{\Delta Y_S}^m \\ \hat{\mu}_{\Delta^2 X_S}^m \\ \hat{\mu}_{\Delta^2 Y_S}^m \end{bmatrix} = \begin{bmatrix} A_S^m & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{\Delta S}^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_{\Delta^2 S}^m \end{bmatrix} \begin{bmatrix} \mu_{X_S}^m \\ \mu_{Y_S}^m \\ \mu_{\Delta X_S}^m \\ \mu_{\Delta Y_S}^m \\ \mu_{\Delta^2 X_S}^m \\ \mu_{\Delta^2 Y_S}^m \end{bmatrix} - \begin{bmatrix} b_S^m \\ b_{\Delta S}^m \\ b_{\Delta^2 S}^m \end{bmatrix} \tag{4}$$

where the subscripts Y_S, ΔY_S and $\Delta^2 Y_S$ represent the static acoustic features and their velocity and acceleration, respectively. A_S^m, $A_{\Delta S}^m$ and $A_{\Delta^2 S}^m$ represent $n_S \times n_S$ shared transformation matrices for static features and their velocity and acceleration, respectively. The dimension of the static feature vector is $n_S = n_{X_S} + n_{Y_S}$, n_{Y_S} is the dimension of the static acoustic feature vector. b_S^m, $b_{\Delta S}^m$ and $b_{\Delta^2 S}^m$ represents $n_S \times 1$ shared bias vectors. The transformation matrices are full matrices so that the dependency between features is considered.

In the stream-independent adaptation structure each feature stream has its own speaker-adaptive transformation, the transform of each component of one stream is related to other components of this

stream but independent from the components in another stream. In the shared adaptation, the speaker-adaptive transformation for each component consists of the linear combination of the two streams. Note that due to the MSD nature of the acoustic F0, the F0 stream is always independent in both of the structures.

4. EXPERIMENTS

4.1. Data

In this paper we use MOCHA-TIMIT database [18], it contains two speakers' data, one female (fsew0) and one male (mask0). The 460 sentences of British TIMIT are uttered by each speaker. Only the waveforms and the electromagnetic articulography (EMA) data from this database are used. The EMA data is the position trajectories in 3D space of seven sensors attached to the tongue dorsum, tongue body, tongue tip, lower incisor, upper lip, lower lip and velum. The EMA data are sampled at 500Hz and the speech waveform is recorded at a sampling rate of 16 kHz. We downsample the EMA data to 200Hz, the normalization process described in [19] is performed to reduce the noise. STRAIGHT analysis [20] is used to extract the line spectrum pairs (LSP) of order 24, the logarithm of gain and the logarithm of fundamental frequency (LF0). The frame length and frame shift are set to 25ms and 5ms respectively. Quinphone labels are used as context information, which has been proven to be effective in [11]. Our experiments use HTS tool-kit version 2.2 [15, 22]. The hidden semi-Markov model (HSMM) [23] based force alignment tool HSMMAlign is used for the force aligns in our experiment. The female speaker is the initial speaker and the male speaker is the test speaker whose data is used for adaptation training and test.

4.2. Model structures

The HMMs are set to be 5 states left to right structure. Four adaptation structures are evaluated in our experiments. The training procedures for the four structures are shown in Figure 2.

a) SS: single-stream structure. This method use only the EMA stream. It is our baseline method;
b) IC: independent clustering structure. It has three feature streams, EMA, LSP-E (LSP plus logarithm of gain) and LF0, they are independent during the HMM states clustering and the learning of speaker-adaptive transformations;
c) SC: shared clustering structure. It has EMA, LSP-E and LF0 stream, the EMA and LSP-E stream share one decision tree during the clustering but the speaker-adaptive transformations are stream-independent;
d) SA: shared adaptation structure. It has EMA, LSP-E and LF0 stream, the EMA and LSP-E stream share one decision tree during the HMM states clustering and the speaker-adaptive transformations are also shared as described in Section 3.3.

The adaptation systems use all the 460 utterances of the initial speaker to train the initial HMMs; 46 utterances of the test speaker are adaptation data, and another 46 utterances are test data. In order to compare the speaker-adaptation training and speaker-dependent training, for the SS, IC and SC, we also implement speaker-dependent training using the rest utterances of test speaker.

The maximum likelihood parameter generation (MLPG) algorithm [24] is applied to generate the optimal EMA trajectories using dynamic features. The optimal state sequence for the parameter generation is obtained by state alignment of the acoustic features with standalone acoustic HMMs, which are trained by speaker-adaptation training with acoustic features of the training data and adaptation data. This can be regarded as a stand-alone auto speech recognition (ASR) system.

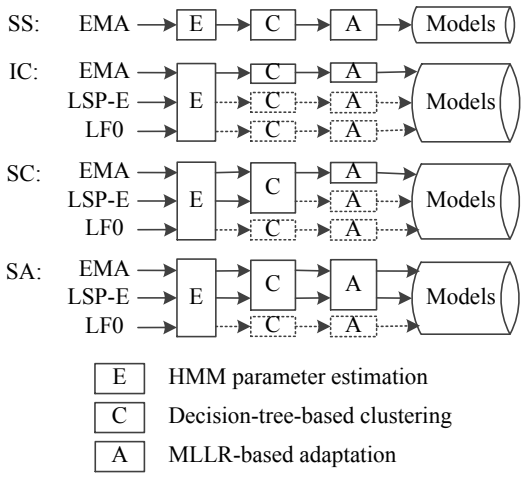

Figure 2: Structures for adaptation training, the models consist of clustered HMMs and speaker-adaptive transformations. The dashed procedures are not necessary since we only estimate the EMA trajectories.

The structures are evaluated by the average of the root mean square error (RMSE) and Pearson product-moment correlation coefficient between the estimated 14 EMA components and their ground truth. All silence and breath sections are removed in the evaluation.

4.3. Independent clustering vs. shared clustering

We first evaluate the independent-clustering structure and shared-clustering structure using both speaker-adaptation training and speaker-dependent training. The number of leaf nodes in the tree-based clustering will affect the systems' performances. Therefore, in order to compare the structures with different leaf node numbers, we control the leaf node number by modifying the MDL. The default MDL factor is 1 for IC and SC. For SC, several MDL factor values are tested and the value 0.7 leads to the closest leaf node numbers to that of IC system. We show the result of the shared clustering structure with the MDL factor set to 1, 0.7 and 0.5, they are denoted as SC1, SC0.7 and SC0.5, respectively. The evaluation results of for SS, IC, SC1, SC0.7 and SC0.5 are shown in Figure 3, and the total leaf node numbers of HMM states after the clustering of the adaptation systems are listed in Table 1.

The experimental results show that the IC and all SC adaptation structures have lower RMSE and higher correlation than SS, which indicate that the acoustic features are helpful in training the HMMs of EMA data regardless of the clustering structure. The IC structure slightly extends the leaf node number of EMA compared with the SS structure and it yields the lowest average RMSE among the tested structures. In the SC1, the combination of EMA with LSP-E reduce the leaf node number to almost half of that of SS, and its performance for speaker-dependent system shows no significant advantage over SS. When we extend the leaf node number to close to the SS structure by modifying the MDL factor to 0.7, the average RMSE become close to that of IC, and the average correlation coefficient slightly surpasses that of IC. The leaf node number of SC0.5 is more than twice that of SS, and the performance of systems are not improving along with the increasing of the leaf node number. The MDL factor 0.7 leads to the best shared clustering adaptation structure, but is still not better than IC. From this result we can infer that even combining

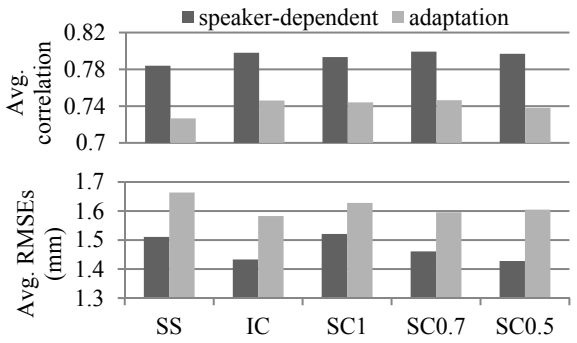

Figure 3: Average RMSE and correlation for speaker-dependent system and speaker-adaptation system

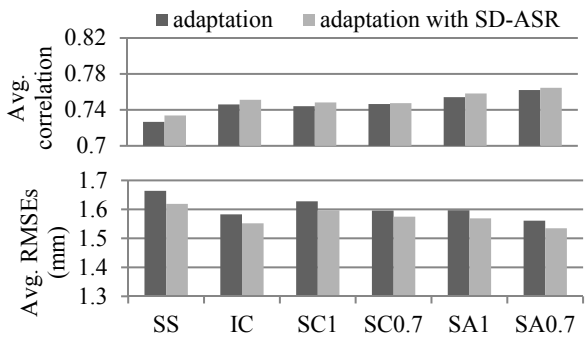

Figure 4: Average RMSE and correlation for adaptation training.

Table 1: The total number of leaf nodes of after states clustering in speaker-adaptation systems.

	SS	IC	SC1/ SA1	SC0.7 /SA0.7	SC0.5
LSP-E	-	2157	3401	6917	14432
EMA	6501	7239			

the EMA with acoustic features for cluster is helpful in speaker-dependent system, but only combine the two stream in clustering stage is not helpful in speaker-adaptation system.

4.4. Independent adaptation vs. shared adaptation

Figure 4 shows the evaluation results for all the adaptation structures. SA1 and SA0.7 denotes the shared-adaptation structure with MDL factor set to 1 and 0.7, respectively. State durations are obtained from stand-alone ASR trained by adaptation training with the acoustic features of training data and adaptation data. Moreover, we also use the state durations obtained from the ASR trained by speaker-dependent training with all the test speakers' data (note that it is not available in real application), these experiments are denoted by "adaptation with SD-ASR" in the figure. The experiments show that the state durations obtained by speaker-adaptive ASR will cause around 0.3mm higher average RMSEs than that by the speaker-dependent ASR.

The average RMSE of SA1 is lower than SC1 while that of SA0.7 is lower than SC0.7, which indicates that shared speaker-adaptive transformation can improve the estimation accuracy of the adaptation structure. Even though the SC0.7 does not show any advantage over IC, but with the shared speaker-adaptive transformation, the SA0.7 yields lower average RMSE than IC and the correlation performance surpasses that of IC. SA0.7 is the best adaptation structure among all the tested structures. It yields an average RMSE of 1.561mm and an average correlation coefficient of 0.762 with 46 utterances of adaptation data from the test speaker.

Figure 5 shows the evaluation for the SC0.7 and SA0.7 structures with different amount of adaptation data. The SA0.7 shows advantage over the SC0.7 regardless of the amount of adaptation data. The 414 adaptation utterances in the last two experiments are the same data that used for training of speaker-dependent HMMs, in other words, those two experiments use not only the same data of the test speaker as in the speaker-dependent system but also take advantage of more data of the initial speaker. The SA0.7 structure yields almost the same results with speaker-dependent system.

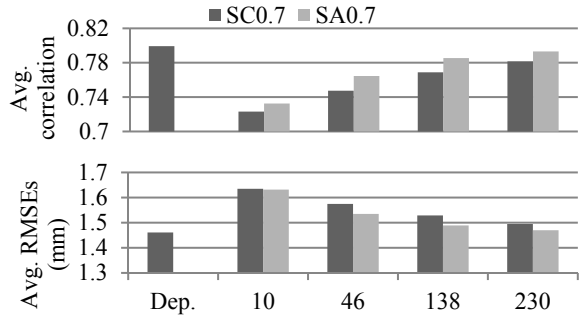

Figure 5: Average RMSE and correlation for adaptation training with different amount of adaptation data. Dep. denotes speaker-dependent training. All the systems use the same state durations.

5. CONCLUSIONS

The experimental results prove that the speaker-adaptation training of HMMs is an effective method for articulatory inversion when the speakers' data is insufficient. Combining with acoustic data, the speaker-adaptation training can get a better performance than use the articulatory data alone. We can improve the performance of speaker adaptation training of articulatory HMMs by sharing the MLLR-based speaker-adaptive transformations. In speaker-adaptation system, sharing both the cluster tree and the adaptive transformation shows advantages over sharing only cluster tree. The multi-stream HMM structure with shared clustering and shared adaptation transformation is the best structure for EMA trajectory prediction among all the speaker-adaptation systems, and it is the suggested structure for this problem. Additional optimal methods can be adopted upon this structure for a better performance. For example, the additional maximum-a-posteriori (MAP) linear regression after the MLLR transformation can be used to improve the prediction accuracy. That is beyond the discussion of the combinatorial structure of the two streams.

6. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No. 61425017 and No.61403386), and the Major Program for the National Social Science Fund of China (13&ZD189).

7. REFERENCES

[1] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America,* vol. 118, p. 444, 2005.

[2] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. ICSLP*, Pittsburgh,USA, pp. 577–580, 2006.

[3] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proc. ICSLP*, Jeju, Korea, pp. 1129–1132, 2004.

[4] Z. Le and S. Renals, "Acoustic-Articulatory Modeling With the Trajectory HMM," *Signal Processing Letters, IEEE,* vol. 15, pp. 245-248, 2008.

[5] S. Hiroya and M. Honda, "Determination of articulatory movements from speech acoustics using an HMM-based speech production model," in *Proc. ICASSP*, Orlando, U.S.A, pp. 437-440, 2002.

[6] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *Speech and audio processing, ieee transactions on,* vol. 12, pp. 175-185, 2004.

[7] A. B. Youssef, P. Badin, G. Bailly, and P. Heracleous, "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models," *Interspeech 2009,* pp. 2255-2258, 2009.

[8] T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded Gaussian mixture regressions," in *Proc. Interspeech*, 2013.

[9] S. Hiroya and M. Honda, "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model," *IEICE TRANSACTIONS on Information and Systems,* vol. 87, pp. 1071-1078, 2004.

[10] S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs," *Speech Communication,* vol. 48, pp. 1677-1690, 2006.

[11] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An Analysis of HMM-based prediction of articulatory movements," *Speech Communication,* vol. 52, pp. 834-846, 2010.

[12] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language,* vol. 12, pp. 75-98, 1998.

[13] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech & language,* vol. 9, pp. 171-185, 1995.

[14] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *Speech and audio processing, ieee transactions on,* vol. 3, pp. 357-366, 1995.

[15]H. Zen*, et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. Sixth ISCA Workshop on Speech Synthesis*, pp. 294-299, 2007.

[16] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Journal of Acoustic Society of Japan (E),* vol. 21, pp. 79-86, 2000.

[17] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE TRANSACTIONS on Information and Systems,* vol. 85, pp. 455-464, 2002.

[18] A. Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/artic/mocha.html, Queen Margaret University College. 1999.

[19] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," PhD thesis, The Centre for Speech Technology Research, Edinburgh University, 2002.

[20] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication,* vol. 27, pp. 187-207, 1999.

[21] S. Young*, et al.*, "The HTK book (for HTK version 3.4)," *Cambridge university engineering department,* 2006.

[22] K. Tokuda*, et al.*, "The HMM-based speech synthesis system (HTS)," http://hts.ics.nitech.ac.jp. 2011.

[23] O. Keiichiro*, et al.*, "A fully consistent hidden semi-Markov model-based speech recognition system," *IEICE TRANSACTIONS on Information and Systems,* vol. 91, pp. 2693-2700, 2008.

[24] K. Tokuda*, et al.*, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, pp. 1315-1318 vol.3, 2000.