



# Acoustic-to-Articulatory Inversion based on Local Regression

Samer Al Moubayed, G. Ananthakrishnan\*

Centre for Speech Technology, KTH (Royal Institute of Technology), Stockholm, Sweden

sameram@kth.se, agopal@kth.se

\* names in alphabetical order

## Abstract

This paper presents an Acoustic-to-Articulatory inversion method based on local regression. Two types of local regression, a non-parametric and a local linear regression have been applied on a corpus containing simultaneous recordings of positions of articulators and the corresponding acoustics. A maximum likelihood trajectory smoothing using the estimated dynamics of the articulators is also applied on the regression estimates. The average root mean square error in estimating articulatory positions, given the acoustics, is 1.56 mm for the non-parametric regression and 1.52 mm for the local linear regression. The local linear regression is found to perform significantly better than regression using Gaussian Mixture Models using the same acoustic and articulatory features.

**Index Terms:** Acoustic-to-Articulatory Inversion, Local Regression, K-Nearest Neighbours

## 1. Introduction

Acoustic-to-Articulatory (A-to-A) inversion has been of special interest to researchers in speech production, due to the theoretical benefits of establishing the relationship between acoustics and articulation. Some of the possible applications include low bit-rate speech coding, driving oro-facial avatars by the acoustic signal, better articulatory models of speech production and the possibility of improving speech recognition.

Inversion by synthesis has been the most important Acoustic-to-articulatory inversion method. Inversion is performed by first using an articulatory synthesizer (e.g. [1]) to create a code-book of different allowed articulator positions and their corresponding acoustics [2]. The inversion is then performed by looking up the input sound in the code-book. The biggest disadvantage of this method, is that the inversion is restricted by the quality of the voice source in the synthesizer which in most cases is not realistic.

Recently, collection of large databases with simultaneous recordings of acoustic and articulatory data, using X-ray microbeam or Electromagnetic Articulography (EMA) has been possible. By using this data to perform statistical regression, the study of acoustic-to-articulatory mapping has thus been made more realistic. Different types of machine learning algorithms have been employed for regression, e.g., Linear Regression [3], Gaussian Mixture Model Regression [4], Artificial Neural Network Regression [5] and HMM regression [6]. Toutios and Margaritis [7] have reviewed the various methods in detail.

Most methods described above are model based methods where it is assumed that the mapping between acoustics and articulatory positions is a single global function. The function is known to be highly non-linear and probably non-unique too [8]. Therefore, model based methods try to include these properties in their model parameters (e.g., mixture density neural networks [9]). Other studies have shown that the important parts of an articulator trajectory are often extrema and are therefore outliers in the data [10].

Memory based methods, on the other hand, do not make any assumptions about the mapping function at a global level. Based on the neighborhood of a specific test sample (in this case

acoustic features) a local model is created to perform a ‘local regression’. Using this model, the corresponding articulatory features of the particular test sample are predicted. There are several advantages of local regression [11] especially for outliers and sparse data. McGowan and Berger [12] have shown that it is possible to perform an acoustic-to-articulatory inversion based on locally weighted regression often called Locally Weighted Scatter plot Smoothing (lowess). They restricted their study to vowels and showed interesting relationships between the articulatory and acoustic features (formants). This paper tries to extend their work to other types of articulations and acoustic features while comparing it with one of the state-of-the-art model based regressions, namely Gaussian Mixture Model Regression (GMMR). We also try to unravel the effects of using dynamic articulatory features resulting in temporal smoothing that allows better estimation of the articulation from the acoustic data.

## 2. Method

The local neighborhood of a test sample are the points in the training data which have the minimum Euclidian distance to the test sample. The local neighborhood may contain only one neighbor or several. This paper discusses two methods of local regression modeling, non-parametric and locally linear (parametric).

### 2.1. Non-parametric Regression

Consider the acoustic space  $Y \in \mathcal{R}^D$  consisting of individual points  $y_n (n : 1 \leq n \leq N)$  which map onto the articulatory space  $X \in \mathcal{R}^\delta$  consisting of  $x_n (n : 1 \leq n \leq N)$  and they are related to each other by the following equation

$$x_n = f^n(y_n) \tag{1}$$

The immediate neighborhood of every sample is considered a perturbation of the sample itself. For a test sample, given the acoustic features  $y_t$  the estimate of the articulatory parameters  $\hat{x}_t$  is calculated from its local neighborhood of  $K$  nearest neighbors  $Y^t = \{y_1^t, \dots, y_k^t, \dots, y_K^t\}$  as follows

$$\hat{x}_t = \sum_{k=1}^K x_k^t \omega_k \tag{2}$$

where  $x_k^t = f^k(y_k^t)$  is as observed in the data and  $\omega_k$  is the weighting given to each neighbor. Many weighting functions have been suggested in the literature (e.g. Gaussian kernels or inverse distance), in this work we considered the inverse distance weighting where  $\omega_k$  is inversely proportional to the distance between  $y_t$  and  $y_k^t$  and  $\sum_{k=1}^K \omega_k = 1$ . The advantage of inverse distance is that it does not require any parameter optimization. Note that  $x_k^t$  are not necessarily the immediate neighbors of the true test sample  $x_t$  but are the corresponding articulatory measurements of the neighbors of the test sample. The estimate  $\hat{x}_t$  is said to minimize the mean square error between the true value and function of the neighborhood. So we call it the Minimum Mean Square Error (MMSE) estimate.

## 2.2. Local Linear Regression

Here, the function defining the relationship between the acoustic features and the articulatory features is considered as a linear function in the local neighborhood. So Equation 2, used to calculate the MMSE estimate  $\hat{x}_t$  is replaced by

$$\hat{x}_t = y_t * \beta^t + \gamma^t \quad (3)$$

Here  $\beta^t$  and  $\gamma^t$  are the parameters of the linear regression and they are estimated from the local neighborhood as the least square's solution to Equation 3. If  $\overline{Y^t} = Y^t - \sum_{k=1}^K y_k^t$  is the mean subtracted neighborhood in the acoustic space, then the solution is

$$\beta^t = \left( \overline{Y^t}^T \Omega \overline{Y^t} \right)^{-1} \overline{Y^t}^T \Omega \left( X^t - \sum_{k=1}^K x_k^t \right) \quad (4)$$

and

$$\gamma^t = \sum_{k=1}^K x_k^t - \beta^t * \left( \sum_{k=1}^K y_k^t \right) \quad (5)$$

where  $\Omega$  is a diagonal matrix consisting of the inverse distance weightings,  $\omega_k$ , for each neighbor.

## 2.3. Maximum Likelihood Trajectory Estimate

The estimates  $\hat{x}_t$  do not make use of the information available in the form of continuity in a trajectory, i.e. dynamic features of the articulators. The  $K$  neighbors are likely to be not only the adjacent frames in one utterance, but also similar sounding frames in many other utterances. However, estimated positions of the articulators do not guarantee that the temporal relation between these articulators (trajectories) are similar to those in the data. In order to provide this information, we use a Maximum Likelihood Trajectory Estimate (MLTE)[13]. In this method, the articulatory parameters are augmented by adding the estimated velocity, acceleration or further delta components of each test data sample, which provide information about how the shape of the trajectory is expected to be. This is done by multiplying  $x_t$  by a matrix  $W_t^{2*d \times d}$  which would give the augmented matrix  $\hat{x}_t = [x_t \ dx_t]^T$ .  $dx_n$  is the first difference or delta of  $x_n$ . Equation 6 shows an example where only the velocity component is added for an utterance consisting of  $\tau$  frames.

$$\begin{bmatrix} x_1 \\ dx_1 \\ x_2 \\ dx_2 \\ \vdots \\ x_{\tau-1} \\ dx_{\tau-1} \\ x_\tau \\ dx_\tau \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 \dots & 0 & 0 \\ 0 & -1 & 1 \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & -1 & 1 \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & -1 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{\tau-1} \\ x_\tau \end{bmatrix} \quad (6)$$

which can be written as  $\widehat{X}_\tau = W_\tau * X_\tau$  where  $X_\tau$  is the articulatory parameters of the utterance. Knowing the matrix  $W_\tau$ , which is a block diagonal matrix of  $W_t : \{1 \leq t \leq \tau\}$  and the MMSE estimate of the augmented articulatory parameters  $\widehat{X}_\tau$ , we can find the weighted least-squares solution to the following equation

$$\widehat{X}_\tau = W_\tau * \widehat{X}_\tau \quad (7)$$

The weights are inversely proportional to the standard deviation of the neighbors giving more importance to neighborhoods which are densely populated. If  $S_t^K$  is the inverse of the standard deviation of the  $K$  nearest neighbors  $\widehat{X}^t$  for frame  $t$ , then  $S_\tau^K$  is a diagonal matrix made from vector  $[S_1^K \dots S_\tau^K]$ .

Then the estimated trajectory  $\widehat{X}_\tau$ , which is the minimum least-squares solution to Equation 7 is given by

$$\widehat{X}_\tau = \left( W_\tau^T S_\tau^K W_\tau \right)^{-1} \left( W_\tau^T S_\tau^K \widehat{X}_\tau \right) \quad (8)$$

The MLTE is expected to be a smoothed version of the MMSE because it also models the dynamics of the trajectory.

## 3. Data

The inversion experiments were conducted using the simultaneously recorded Acoustic-EMA data from the MOCHA database [14] consisting of 460 TIMIT sentences spoken by one female speaker. The sentences had a total number of 45 phonemes including silence. The 14 articulatory channels consisted of the X- and Y-axis trajectories of 7 EMA coils placed on the Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (VE). The trajectories were processed to remove the drift as described in [9]. The EMA data was low-pass filtered and down-sampled to 100 Hz, in order to correspond to the acoustic frame shift rate. 18 Mel Frequency Cepstral Coefficients coefficients (including the 0<sup>th</sup>) were calculated, and 11 adjacent acoustic frames each of duration 25 ms (at a frame rate of 100 Hz) were considered. The features were reduced using Principal Component Analysis (PCA) such that all components that contributed to less than 3% of the variation were removed. We then had 49 acoustic features which contained information from 125 ms of the signal. The delta features (e.g., velocity and acceleration) for the articulatory measurements were also computed for the MLTE estimation. The articulatory trajectory vectors of the training data were normalized to zero mean with a Standard Deviation (SD) of 1. A ten-fold cross-validation was performed where 314 sentences per fold were used for training and 46 sentences were used for testing the method's performance.

The baseline system using GMMR [15] was also implemented using the same features. However, the MLTE from the GMMR was calculated only with velocity coefficients. The number of Gaussians that were used were 64 as recommended by Toda *et. al.* [15].

The articulatory estimates were finally filtered using the cut-off frequencies suggested in [15] in order to smooth the trajectories and achieve a better performance.

## 4. Experiments and Results

The MMSE estimate has only one parameter to optimize, namely  $K$ , the number of neighbors considered. The MLTE has a second parameter,  $ND$ , the number of delta coefficients (dynamics information). We estimated the Root Mean Square Error ( $RMSE$ ) over the utterance trajectories of each articulator and then the mean across all the  $\delta$  number of articulators is  $mRMSE$ . We optimized the  $mRMSE$  over the ten-fold cross-validation for the values of  $K = 1, 2 \dots 20, 50, 100, 200, 500, 700, 1000$  and  $ND = 1, 2, 3, 4, 5$ . The second standard evaluation criterion we adopted was the Correlation Coefficient ( $CC_d$ ) between the measured and the estimated trajectory for every articulator  $d$ . The mean Correlation Coefficient  $mCC$  is calculated by averaging over all articulatory trajectories.

The effect of  $K$  on the performance of the local non-parametric and local linear regression is illustrated in Figure 1. The optimum  $K$  for non-parametric regression is 20 and for local linear regression is 1000. For values of  $K$  between 10 and 60, the local regression estimates show very high error, because most of the estimates for  $\beta^t$  were ill-conditioned matrices, making the linear regression very unstable. Figure 2 shows the comparison between the MMSE estimates of the two local regression methods with optimized  $K$  against the MMSE estimate using GMMR with 64 Gaussians. The non-parametric method performs slightly better than the local regression, although the

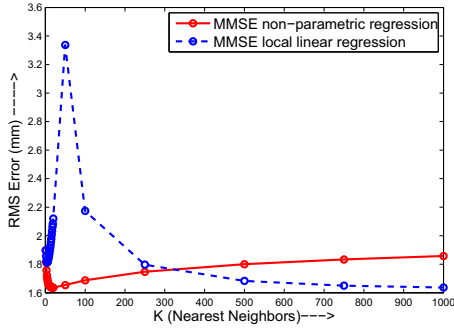


Figure 1: Graph illustrating the effect of  $K$  on the  $mRMSE$  over the MMSE estimates for the two local regression methods. The minimum for non-parametric regression is for  $K = 20$  and for local linear regression is  $K = 1000$

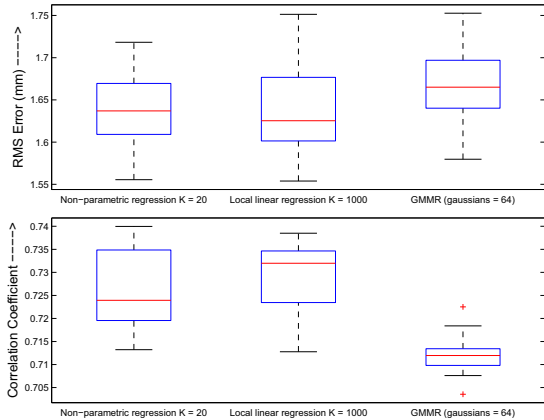


Figure 2: Comparative performance between the two local regression methods against GMMR when no dynamics are considered (MMSE estimate) over 10 fold cross-validation. Top: mean RMS error ( $mRMSE$ )(mm), the lower the better. Bottom: mean Correlation Coefficient ( $mCC$ ), the higher the better. All the results are for the optimum configurations of the respective methods.

error is not significantly reduced. While the improvement of the local regressions over GMMR is not significant in terms of the  $mRMSE$ , the improvement for both the methods is statistically significant ( $p < 0.01$ ) with respect to  $mCC$ .

Figure 3 illustrates the effect of  $K$  and  $ND$  on the performance of the two methods of local regression for the MLTE. The optimum parameters are  $K = 500$  and  $ND = 2$  for local linear regression and  $K = 4$  and  $ND = 1$  for non-parametric local regression. The optimum  $K$ s are different from the MMSE case. The parameter  $ND$  does not affect the performance as much as  $K$ . Information about the dynamics for  $ND > 2$  does not improve the performance. In this case, there is a significant improvement of local linear regression over local non-parametric regression indicating that trajectory estimate using dynamics has a greater effect on local linear regression than on local non-parametric regression. One reason could be because the non-parametric method, which involves finding the mean over the local neighborhood, already performs a smoothing on the estimated trajectories. The smoothing effect provided by the dynamic features does not add to the performance. Figure 4 shows the comparison between the MLTEs using the two local regressions and GMMR with  $ND = 1$ , in order to make a fair comparison. While the improvement of the non-parametric local regression over GMMR is not significant, the improvement for local linear regression is statistically significant ( $p < 0.1$ ) for  $mRMSE$  and for  $mCC$  ( $p < 0.05$ ).

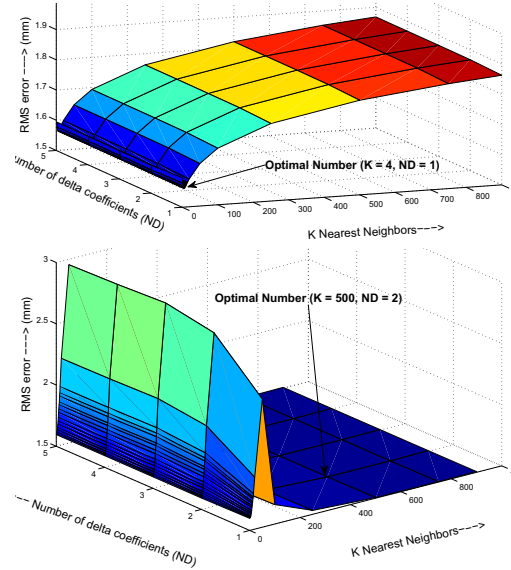


Figure 3: Graph illustrating the effect of  $K$  and  $ND$  on the  $mRMSE$  over the MLTEs for the two local regression methods. Top: Non-parametric local regression. Bottom: Local Linear Regression

Figure 5 shows the estimated and measured trajectories of one of the articulatory parameters (tongue tip) for one utterance. It is clear that although the MLTEs are smoother than the MMSE estimates, they are not better at every frame. The MMSE estimates sometimes model the extreme positions better than the MLTEs.

It is interesting to see that while local linear regression performed significantly better than local non-parametric regression when dynamic information was available, there was a vast difference in the size of the neighborhood that was required to give the optimal results. While local linear regression required a much larger neighborhood ( $K > 200$ ) to provide reliable linear approximation, the non-parametric regression required only a small neighborhood of 4 neighbors. This indicates the differences in approaches of these methods. Another reason could be the sensitivity of the linear regression to noise, thus requiring sufficient number of neighbors to estimate the parameters robustly. The non-parametric regression, on the other hand, utilizes the averaging function to perform smoothing over noisy data.

Richmond [5] has illustrated the effect of adding dynamic information from the acoustics in the form of 11 consecutive acoustic frames and information about dynamics in the articulatory space in the form of delta coefficients (for  $ND > 1$ ). The results presented in this paper reiterate how important the information about dynamics in the articulatory trajectory is. Using the information from a single acoustic neighbor along with articulatory dynamics, MLTE performs almost as well as modeling the entire acoustic-articulatory space using a GMM distribution. Thus it can be said that when information about the articulatory neighbors (or their estimates) in the particular trajectory is present, the information about the neighbors in the acoustic space can be restricted to a small neighborhood.

The main drawback of this method is that it is much slower than other global model based methods (for example local regression is around 100 times slower than GMMR for this database). However, it may be noted that the local regression methods do not have any training time which is quite large for the methods like GMMR. Local regression may be preferred when accuracy of regression is the more important performance parameter and is used for off-line applications.

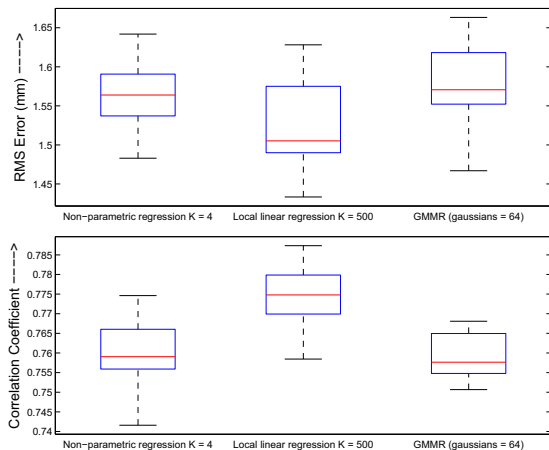


Figure 4: Comparative performance between the two local regression methods against GMMR when  $ND = 1$  (MLTEs) and optimized  $K$  over 10 fold cross-validation. Top: mean RMS error ( $mRMSE$ )(mm), the lower the better Bottom: mean Correlation Coefficient ( $mCC$ ), the higher the better

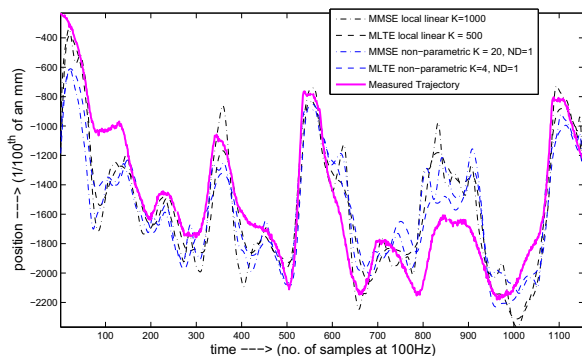


Figure 5: Figure showing the measured and estimated tongue tip (TT) trajectories along the up-down axis for the sentence 'Jane may earn more money by working hard'. Each of the estimates is using the optimum set of parameters.

## 5. Conclusion and Future Work

A system for performing acoustic-to-articulatory inversion using local regression has been described and experiments have been conducted on acoustic-EMA data. Two methods of local regression have been discussed. The local non-parametric regression has an optimum performance of 1.56 mm  $mRMSE$  (0.64 of the standard deviation of the data) and an  $mCC$  of 0.76 and the local linear regression has an optimum performance of 1.52 mm  $mRMSE$  (0.63 of the standard deviation of the data) and an  $mCC$  of 0.78. The effect of trajectory smoothing using dynamic features has a higher effect on local linear regression as compared to local non-parametric regression.

Even though the results presented in this paper show a performance better than GMMR for the same set of features, the method does not perform favorably against other methods reported in the literature on the same data. Richmond [5] reported an  $mRMSE$  of 1.4 mm using trajectory multiple density neural networks and Toda *et. al.* [15] reported an  $mRMSE$  of 1.45 mm using GMMR with dynamic features. However, the training and testing samples for these two studies is not known and the results fall within the variation observed for local linear regression with MLTE.

Future work is targeted towards trying to apply this method of regression to other speakers and databases. An important aspect to study is the relative importance of the dynamic articula-

tory features against acoustic neighborhood. This study gives an indication that dynamic features for articulation are rather important for the inversion, but it is not clear how important they are. Finding an algorithm which could improve the time taken for inversion is another important direction for future work. The largest amount of time taken by local regression methods is to find the nearest neighbors of a test sample. Several methods like pruning, convex clustering and hierarchical trees [11] provide improved performance in searching for the neighbors. Finally, it would be interesting to see whether this method of regression can help improve speech recognition by providing articulatory knowledge.

## 6. Acknowledgements

This work has been partially supported by the Swedish National Graduate School of Language Technology (GSLT) and the Swedish Research Council project 80449001, Computer-Animated Language Teachers.

## 7. References

- [1] Maeda, S., "Improved articulatory models," J. Acoust. Soc. Am., 84(S1):S146, 1988.
- [2] Atal, S., Chang, J., Mathews, J., and Tukey, W., "Inversion of articulatory-to-acoustic information in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am., 63:1535–1555, 1978.
- [3] Yehia, H., Rubin, P., and Vatikiotis-Bateson, E., "Quantitative association of vocal-tract and facial behavior," Speech Communication, 26(1-2):23–43, 1998.
- [4] Toda, T., Black, A., and Tokuda, K., "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in Proc. Int. Conf. on Spoken Language Processing. ISCA, 1129–1132, 2004.
- [5] Richmond, K., "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in Proc. Interspeech. Citeseer, 577–580, 2006.
- [6] Hiroya, S. and Honda, M., "Estimation of articulatory movements from speech acoustics," in IEEE Trans. Speech and Audio Processing, 12(2):175–185, 2004.
- [7] Toutios, A. and Margaritis, K., "A rough guide to the acoustic-to-articulatory inversion of speech," in 6th Hellenic European Conference of Computer Mathematics and its Applications, 1–4, 2003.
- [8] Neiberg, D., Ananthkrishnan, G., and Engwall, O., "The Acoustic to Articulation Mapping: Non-linear or Non-unique?," in Proc. Interspeech, 1485–1488, 2008.
- [9] Richmond, K., *Estimating articulatory parameters from the speech signal*, Ph.D. thesis, The Center for Speech Technology Research, Edinburgh, 2002.
- [10] Ananthkrishnan, G. and Engwall, O., "Important Regions in the Articulator Trajectory," in Proc. of International Seminar on Speech Production, 305–308, 2008.
- [11] Atkeson, C., Moore, A., and Schaal, S., "Locally weighted learning," Artificial intelligence review, 11(1):11–73, 1997.
- [12] McGowan, R. S. and Berger, M. A., "Acoustic-articulatory mapping in vowels by locally weighted regression," J. Acoust. Soc. Am., 126(4):2011–2032, 2009.
- [13] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. International Conference on Acoustics, Speech, and Signal Processing. Citeseer, 1315–1318, 2000.
- [14] Wrench, A., "The MOCHA-TIMIT articulatory database," Queen Margaret University College, Tech. Rep, 1999.
- [15] Toda, T., Black, A., and Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Communication, 50(3):215–227, 2008.