The Acoustic to Articulation Mapping: Non-linear or Non-unique?

Daniel Neiberg, G. Ananthakrishnan, Olov Engwall

Centre for Speech Technology, CSC, KTH (Royal Institute of technology), Stockholm, Sweden
neiberg@speech.kth.se, agopal@kth.se, olov@speech.kth.se

Abstract

This paper studies the hypothesis that the acoustic-to-articulatory mapping is non-unique, statistically. The distributions of the acoustic and articulatory spaces are obtained by fitting the data into a Gaussian Mixture Model. The kurtosis is used to measure the non-Gaussianity of the distributions and the Bhattacharyya distance is used to find the difference between distributions of the acoustic vectors producing non-unique articulator configurations. It is found that stop consonants and alveolar fricatives are generally not only non-linear but also non-unique, while dental fricatives are found to be highly non-linear but fairly unique. Two more investigations are also discussed: the first is on how well the best possible piecewise linear regression is likely to perform, the second is on whether the dynamic constraints improve the ability to predict different articulatory regions corresponding to the same region in the acoustic space. By non-uniqueness, it is meant that multiple vocal tract configurations can produce almost the same acoustic features. In real continuous speech, however, the possibility of finding two data points with exactly the same acoustic parameters is abysmally poor. In order to simplify the problem, the acoustic space is quantized. If two data points within this quantization range fall sufficiently apart in the articulatory space, then the mapping is said to be non-unique. However, a result obtained in this manner can be quite misleading, since this kind of an effect could be caused due to insufficient resolution of quantization and data sparseness. In an attempt to provide answers to these issues a new model based paradigm is proposed, which tries to quantify non-uniqueness.

Researchers have used articulator dynamic constraints in the hope of solving the problem of non-uniqueness by charting the optimum course from among the various possible articulatory trajectories. However, there are very few studies to show whether applying articulatory constraints will actually effect the disambiguation of the mapping, sufficiently.

Three experiments and their results are presented in the following sections. The first one tries to find the best possible piecewise linear regression mapping. The second studies the effect of dynamic constraints and the third studies the non-linearity and non-uniqueness of the A-to-A mapping.

1. Introduction

The acoustic-to-articulatory (A-to-A) mapping, also known as articulatory inversion has been of special interest to speech researchers for quite some time. It deals with estimating or recovering vocal tract shapes from the acoustics of an utterance. It remains one of the fundamental problems in understanding speech production. The inversion has several applications, namely low bit-rate encoding, training visual agents like talking heads, improving articulatory speech synthesis, and robust speech recognition. Research in this topic has shown substantial progress in terms of using several machine learning techniques to minimize the error between the true vocal tract shape and the shape estimated using knowledge of the acoustics. Atal et al. [1] and Roweis [2] have performed inversion using codebooks and dynamic programming while Richmond [3] has done extensive research on performing the mapping using mixture density Neural Network (NN) regression. Hiroya [4] has used Hidden Markov Models (HMM) with one or more states per phoneme, in order to perform the inversion. Toda et al. [5] have used a Gaussian mixture model (GMM) along with Maximum Likelihood Estimation (MLE) smoothing for dynamic features. These methods have been extremely successful at minimizing the average root mean square error (RMSE) over all articulatory channels and maximizing the Pearson’s Correlation Coefficient (PCC) between the actual vocal tract configuration and the predicted ones.

The errors in the mapping are often attributed to the non-uniqueness of the inversion known as ‘fibers’ in the articulatory space. By non-uniqueness, it is meant that multiple vocal tract configurations can produce almost the same acoustic features. Early research presented some interesting evidence corroborating non-uniqueness. Bite-block experiments showed that the speakers were capable of producing sounds perceptually close to the intended sounds even though the jaw was fixed in an unnatural position. The lossless tube models of the vocal tract also indicate the possibility of this non-uniqueness [6]. A similar result was obtained by Schroeder [7] using perturbation analysis. Qin et al. [8] performed, possibly, the first empirical investigation into the non-uniqueness problem. They quantized the acoustic space using the perceptual Itakura distance on LPC features. The articulatory space was modeled using a non-parametric Gaussian density Neural Network (NN) regression. Hiroya [4] has used Hidden Markov Models (HMM) with one or more states per phoneme, in order to perform the inversion. Toda et al. [5] have used a Gaussian mixture model (GMM) along with Maximum Likelihood Estimation (MLE) smoothing for dynamic features. These methods have been extremely successful at minimizing the average root mean square error (RMSE) over all articulatory channels and maximizing the Pearson’s Correlation Coefficient (PCC) between the actual vocal tract configuration and the predicted ones.

The errors in the mapping are often attributed to the non-uniqueness of the inversion known as ‘fibers’ in the articulatory space. By non-uniqueness, it is meant that multiple vocal tract configurations can produce almost the same acoustic features. Early research presented some interesting evidence corroborating non-uniqueness. Bite-block experiments showed that the speakers were capable of producing sounds perceptually close to the intended sounds even though the jaw was fixed in an unnatural position. The lossless tube models of the vocal tract also indicate the possibility of this non-uniqueness [6]. A similar result was obtained by Schroeder [7] using perturbation analysis. Qin et al. [8] performed, possibly, the first empirical investigation into the non-uniqueness problem. They quantized the acoustic space using the perceptual Itakura distance on LPC features. The articulatory space was modeled using a non-parametric Gaussian density Neural Network (NN) regression. Hiroya [4] has used Hidden Markov Models (HMM) with one or more states per phoneme, in order to perform the inversion. Toda et al. [5] have used a Gaussian mixture model (GMM) along with Maximum Likelihood Estimation (MLE) smoothing for dynamic features. These methods have been extremely successful at minimizing the average root mean square error (RMSE) over all articulatory channels and maximizing the Pearson’s Correlation Coefficient (PCC) between the actual vocal tract configuration and the predicted ones.

The errors in the mapping are often attributed to the non-uniqueness of the inversion known as ‘fibers’ in the articulatory space. By non-uniqueness, it is meant that multiple vocal tract configurations can produce almost the same acoustic features. Early research presented some interesting evidence corroborating non-uniqueness. Bite-block experiments showed that the speakers were capable of producing sounds perceptually close to the intended sounds even though the jaw was fixed in an unnatural position. The lossless tube models of the vocal tract also indicate the possibility of this non-uniqueness [6]. A similar result was obtained by Schroeder [7] using perturbation analysis. Qin et al. [8] performed, possibly, the first empirical investigation into the non-uniqueness problem. They quantized the acoustic space using the perceptual Itakura distance on LPC features. The articulatory space was modeled using a non-parametric Gaussian density Neural Network (NN) regression. Hiroya [4] has used Hidden Markov Models (HMM) with one or more states per phoneme, in order to perform the inversion. Toda et al. [5] have used a Gaussian mixture model (GMM) along with Maximum Likelihood Estimation (MLE) smoothing for dynamic features. These methods have been extremely successful at minimizing the average root mean square error (RMSE) over all articulatory channels and maximizing the Pearson’s Correlation Coefficient (PCC) between the actual vocal tract configuration and the predicted ones.
knife principle to rotate the 4 training parts and 1 testing part of the data.

3. The best piece-wise linear mapping

Many algorithms using the HMM framework as used by Hiroya [4] or the GMM such as Toda et al. [5] assume the inversion mapping to be piecewise linear. They separate the acoustic space into many clusters and use linear regression within each of the clusters. This is true also for other code-book based methods like those used by Roweis [2]. Considerable effort is spent on identifying which of the clusters in the articulatory space corresponds to the current cluster in the acoustic space. Toda et al. use an MLE with dynamic features, Roweis uses a self organizing HMM (SOHMM) and Hiroya uses smoothing with dynamic feature constraints. The errors, that these methods incur, are on account of two reasons. Firstly, the error due to incorrect selection of the articulatory cluster and secondly due to the error in the linear mapping from the acoustic to articulatory cluster.

The experiment carried out in this section assumes that there exists a method which tells us the correct articulatory cluster for every acoustic cluster with 100% accuracy. The only error that remains is due to the linear regression estimate. The acoustic clusters are obtained by K-means clustering, for different values of K. For data points belonging to the \( k \)-th cluster in the acoustic space, the articulatory parameters are clustered further. The average error, \( E_r \), for the \( k \)-th acoustic cluster is calculated by

\[
E_r = \frac{\sum_{r=1}^{R_k} \sum_{j=1}^{N_k} (Y_r^j - \hat{Y}_r^j)^T (Y_r^j - \hat{Y}_r^j)}{\sum_{r=1}^{R_k} N_r^k}
\]

where \( Y_r^j = \{y_j \in \mathbb{R}^d : 1 \leq j \leq N_r^k \} \) are the normalized articulatory features of the data points belonging to the \( r \)-th articulatory cluster of the \( k \)-th acoustic cluster, while \( \hat{Y}_r^j \) is its linearly predicted counterpart using one linear predictor per cluster. The number of articulatory clusters, \( R_k \), are increased until the average error, \( E_{rs} \), falls below the threshold \( \tau \).

Table 1 gives us the mean RMSE over all the articulatory channels on the testing set for different values of \( K \) and \( \tau \). We can observe that the performance is best when we have smaller number of acoustic clusters with a larger number of articulatory clusters for each acoustic cluster. For small values of \( \tau \) with a larger number of acoustic clusters, the testing error is high because of over fitting. We can see that the best performance by a piecewise linear regression mapping is a mean RMSE of 0.91 mm for 8 acoustic clusters and around 45 articulatory clusters per acoustic cluster.

4. The best prediction using dynamic constraints

The experiment in this section tries to model just the error in determining the correct cluster using dynamic constraints. It is assumed that the articulatory position for the previous time frame is known. The current articulatory cluster is determined by using dynamic smoothing constraints. The correct articulatory cluster is selected by minimizing either the distance or the difference in velocity between the previous position and the current possible positions. The prediction error within the cluster is assumed to be zero. The mean RMSE is calculated for different number of acoustic and articulatory clusters as shown in Tables 2, 3 and 4.

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( K )</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1.06 (256)</td>
<td>6.47 (251)</td>
<td>45.26 (155)</td>
<td>21.2 (77)</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>1.06 (256)</td>
<td>5.36 (176)</td>
<td>1.17 (80)</td>
<td>1.01 (43)</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1.06 (201)</td>
<td>0.97 (88)</td>
<td>0.91 (45)</td>
<td>0.97 (25)</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>1.05 (141)</td>
<td>0.97 (64)</td>
<td>0.96 (39)</td>
<td>1.01 (18)</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.08 (48)</td>
<td>1.06 (25)</td>
<td>1.09 (13)</td>
<td>1.18 (9)</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>1.34 (14)</td>
<td>1.29 (8)</td>
<td>1.34 (6)</td>
<td>1.52 (3)</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>1.53 (6)</td>
<td>1.59 (3)</td>
<td>1.70 (2)</td>
<td>1.71 (2)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The mean RMSE for the ideal prediction without using dynamic constraints, for the female speaker. \( K \) corresponds to the number of acoustic clusters and \( R_k \) to the number of articulatory clusters per acoustic cluster.

\[
R_k \quad K \quad 2 \quad 4 \quad 8 \quad 16 \quad 32 \quad 64 \quad 128
\]

\[
2 \quad 0.51 \quad 0.49 \quad 0.48 \quad 0.48 \quad 0.48 \quad 0.47 \quad 0.46
\]

\[
4 \quad 1.86 \quad 1.83 \quad 1.80 \quad 1.80 \quad 1.76 \quad 1.74 \quad 1.72
\]

\[
8 \quad 2.74 \quad 2.70 \quad 2.50 \quad 2.50 \quad 2.41 \quad 2.36 \quad 2.29
\]

\[
16 \quad 3.04 \quad 2.95 \quad 2.76 \quad 2.65 \quad 2.50 \quad 2.46 \quad 2.37
\]

\[
32 \quad 2.72 \quad 3.03 \quad 2.54 \quad 2.58 \quad 2.56 \quad 2.47 \quad 2.45
\]

\[
64 \quad 3.24 \quad 3.22 \quad 2.70 \quad 2.77 \quad 2.62 \quad 2.59 \quad 2.56
\]

\[
128 \quad 3.50 \quad 3.39 \quad 3.09 \quad 2.95 \quad 2.72 \quad 2.68 \quad 2.63
\]

We can see that applying dynamic constraints is useful in predicting the correct articulatory cluster. The distance minimizing constraint performs better than the one minimizing velocity difference. It is no surprise the error is lower for a really small number of articulatory clusters, since we are assuming a zero within-cluster error. But the tables indicate a tendency for better performances with a lower number of acoustic clusters.

For a very high number of acoustic clusters, the results show a slight improvement. Decreasing the number of articulatory clusters improves the prediction when dynamic constraints are applied. Thus an optimum choice for piecewise linear regression with dynamic constraints would be 2 acoustic clusters and around 32 to 64 articulatory clusters per acoustic cluster.

5. Gaussian Mixture Modeling

The experiment in this section uses a model based approach which fits the given data into a GMM, in such a way that every Gaussian represents a cluster in the acoustic space. This is achieved using the Expectation Maximization (EM) algorithm [10] to obtain parameters that are Maximum Likelihood (ML) estimates. The Schwartz or Bayesian Information Criterion (BIC) [10], is minimized, in order to find the optimum number of clusters in the acoustic space. Every phoneme (\( \phi \)) in the corpus is modeled by a Gaussian mixture model, \( \lambda_\theta \), containing \( K_\theta \) clusters. \( K_\theta \) is chosen by minimizing BIC.

5.1. Non-linearity

For the data points belonging to the \( k \)-th acoustic Gaussian, \( X_k \), the corresponding articulatory subspace is modeled by an optimal number of Gaussians, \( R_k \), using the same method. The articulatory vectors belonging to the \( r \)-th such Gaussian are given
by \( Y^k_{p}(r) \).

If there exists only one Gaussian in the articulatory space for the \( k \)th Gaussian in the acoustic space (i.e. \( R^k_{p} = 1 \)), then the distribution of the articulatory vectors can be predicted by performing a linear transform on the Gaussian distribution of the acoustic vectors. However, if \( R^k_{p} > 1 \), then it means that this sort of a linear transform cannot be performed. The less normal the articulatory space is, the more non-linear the mapping. The non-linearity is calculated by using the ‘Mardia’s multi-variate kurtosis’ [11] for goodness of fit to a normal distribution.

\[
NL^k_{p} = \frac{1}{N^k_{p}} \sum_{i=1}^{N^k_{p}} \left( Y^k_{p}(i) - \mu^k_{p} \right)^T \Sigma^k_{p}^{-1} \left( Y^k_{p}(i) - \mu^k_{p} \right)^2 - d(d+2)
\]  

(2)

\( N^k_{p} \) is the number of data points and \( \mu^k_{p} \) and \( \Sigma^k_{p} \) are the mean and covariance for \( Y^k_{p}, \) the articulatory subspace corresponding to the \( k \)th Gaussian in the acoustic space of the phoneme \( p. \) \( NL \) is the proposed measure of non-linearity. It is important to note here, that observing a multi-modality in the distribution of the articulatory features corresponding to a single mode Gaussian of acoustic features does not necessarily imply non-uniqueness, but it necessarily implies non-linearity. The authors want to stress this point here, because it is easy to confuse the multi-modality with non-uniqueness. A more stringent measure is necessary to imply non-uniqueness.

### 5.2. Non-uniqueness

Consider the Gaussian acoustic space, \( X^k_{p}, X^k_{p}(r) \) is a subset which corresponds to one mode of the articulatory space, \( Y^k_{p}(r) \). There are two possibilities. The first possibility is that, this subset does not have a Gaussian distribution. In such a case, it may be possible to find a non-linear mapping between each of these sets of data to the corresponding mode in the articulatory space. But if this part has the same Gaussian distribution as the whole single Gaussian, i.e., if \( \rho(X^k_{p}(r)) \) and \( \rho(X^k_{p}) \) have exactly the same parameters, then it connotes that the data points with exactly the same distribution in the acoustic space can actually produce articulatory features with different distributions. This is the necessary and sufficient condition to imply non-uniqueness. In order to find out the similarity between the distributions \( \rho(X^k_{p}(r)) \) and \( \rho(X^k_{p}) \), the Bhattacharya distance is used. However, there is no accurate method of calculating this distance for unknown distributions. Non-parametric distribution estimation techniques would suffer from a data sparseness problem.

So, we use the Bhattacharya distance assuming a Gaussian distribution but weigh it by the Gaussianity of the data. The Bhattacharya distance for Gaussian iid data is given by

\[
D_{bh}(r) = \frac{1}{8} (\mu_r - \mu_k)^T \left( \frac{\Sigma_r + \Sigma_k}{2} \right)^{-1} (\mu_r - \mu_k) + \frac{1}{2} \ln \left( \frac{\Sigma_r + \Sigma_k}{\Sigma_r \Sigma_k} \right)
\]

(3)

where \( \mu_r \) and \( \Sigma_r \) are the mean and covariance estimates of \( X^k_{p}(r) \). \( \mu_k \) and \( \Sigma_k \) are the mean and covariance estimates of \( X^k_{p} \). Non-Gaussianity is determined by the kurtosis of the data points. We bias it so that it takes the value of 1 for a perfect Gaussian distribution and is higher for non-Gaussians. Thus non-uniqueness, \( (NU^k_{p}(r)) \), can be defined as the inverse of the Bhattacharya distance weighted by the measure of its Gaussian nature.

\[
NU^k_{p}(r) = \frac{1}{\left( K_m(X^k_{p}(r)) + 1 \right) D_{bh}(r)}
\]

(4)

\( K_m(.) \) denotes the multi-variate kurtosis of the data points. Thus, \( NU \) is lower for clusters with unique mapping.

### 5.3. Analysis of non-linearity and non-uniqueness for different phonemes

Figure 1 shows a comparative study of the non-linearity (\( NL \)) and non-uniqueness (\( NU \)) for a few phonemes in the database, for the male and female speaker. The mean \( NL_{p} \) and \( NU_{p} \) for phoneme \( p \) are calculated by weighing them with the weight factor of the respective Gaussians. We can see that most consonants exhibit non-uniqueness while most vowels have a small degree of non-linearity. The stop consonants /l/, /d/, /l/ and /p/ and fricatives like /s/, /z/ have a higher degree of non-uniqueness. The fricatives /s/ and /z/ are highly non-linear but rather unique, while liquids such as /l/, /l/ and /l/ are found to be rather non-linear, but unique by the method used in this paper. While the non-linearity and non-uniqueness of stop consonants is expected due to the silence region, the reason why alveolar fricatives show high non-uniqueness could be because the EMA data may not have adequate detail to measure the exact location of the tongue tip. There are considerable variations in the levels of non-uniqueness and non-linearity shown for the male and female speakers, but the trends are more or less similar. Figure 2 and 3 show corresponding data points in the articulatory and the acoustic subspaces.

### 6. Conclusion and Future work

Three experiments have been conducted on the MOCHA database. The first one shows that the performance of the best possible piecewise linear prediction mapping cannot improve the accuracy of the mapping to better than 0.91 mm
mean RMSE in the database for the female speaker. The second experiment shows the role of dynamic constraints in the disambiguation of the mapping between acoustic and articulatory vectors. The experiment shows that applying dynamic constraints improves the performance, but does not completely disambiguate the multiple-mapping. The third experiment proposes to distinguish between non-linearity and non-uniqueness. It suggests measures to quantify the same, and analyzes the non-linearity and non-uniqueness of different phonemes in the database for two speakers. Phonemes such as /t/ and /p/ are found to be non-unique, while other phonemes like /n/ and /s/ are found to be non-linear, but rather unique, from our studies. Future work can be directed in trying to estimate the best possible non-linear estimators for clusters with high non-linearity and further constraints or allowances to tackle the non-uniqueness of the mapping. Work must be done on defining a non-uniqueness criterion for any general distribution. Knowing which articulators contribute more to the non-uniqueness could be another direction for research. Studies with methods than EMA, must be carried out to validate the results obtained from this paper.

7. Acknowledgements

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324”.

8. References