In search of Non-uniqueness in the Acoustic-to-Articulatory Mapping

G. Ananthakrishnan, Daniel Neiberg, Olov Engwall

Department of Speech Music and Hearing (TMH), CSC, KTH, Stockholm, Sweden
agopal@kth.se, neiberg@speech.kth.se, engwall@kth.se

Abstract

This paper explores the possibility and extent of non-uniqueness in the acoustic-to-articulatory inversion of speech, from a statistical point of view. It proposes a technique to estimate the non-uniqueness, based on finding peaks in the conditional probability function of the articulatory space. The paper corroborates the existence of non-uniqueness in a statistical sense, especially in stop consonants, nasals and fricatives. The relationship between the importance of the articulator position and non-uniqueness at each instance is also explored.

Index Terms: acoustic-to-articulatory inversion, non-uniqueness, gaussian mixture modeling.

1. Introduction

Speech production is a physical process which depends on the configuration of the articulators and the voice source. Thus acoustic-to-articulatory inversion, which tries to predict the articulatory configuration given the acoustics, could be expected to be a one-to-one mapping. However, bite-block experiments have shown that speakers are capable of producing sounds perceptually close to the intended sounds even though the jaw is fixed in an unnatural position [1]. Mermelstein’s [2] lossless tubes model of the vocal tract for speech production also indicates the possibility of non-uniqueness, showing that the inverse mapping from the acoustics is to a class of area functions, rather than a unique configuration. A similar result was obtained by Schroeder [3] using perturbation analysis, where he showed that the solutions depended on the vocal tract length.

Statistically based inversion methods rely on copious amounts of simultaneously collected acoustics and articulatory data, gathered with e.g., X-ray microbeam or Electromagnetic Articulography (EMA) [4, 5]. The problem of inversion is tackled using statistical regression methods. It is assumed that the articulatory configuration, given the acoustics, is a random variable with as many dimensions as the number of measured articulator positions.

The concept of non-uniqueness in the statistical sense is different from the deterministic sense. In the deterministic case, one can say that if the same acoustic parameters are produced by more than one articulatory configuration, then the particular mapping is considered to be non-unique. However, such a hypothesis cannot be proved in a statistical sense, using real recorded data, unless two articulatory configurations produce exactly the same acoustic parameters. The probability of finding such instances of exactly the same acoustics is extremely small for a limited database. Hence, the notion needs to be re-defined so that it can be accommodated in the statistical sense.

Non-uniqueness in the statistical sense is not easily definable for several reasons. The first among them is the confusion between variance and non-uniqueness. Since the mapping is considered to be a random phenomenon, it may not be possible to predict the exact value of the articulatory configuration given the acoustic parameters. However, it may be possible to estimate parameters of the probability distribution of the articulatory features, given the acoustic parameters.

With an effort to mitigate this problem, a study made by Qin and Carreira-Perpinán [6] proposed that the mapping is non-unique if, for a particular acoustic cluster, the corresponding articulatory mapping may be found in more than one cluster. They could show the evidence of non-uniqueness in certain acoustic clusters for phonemes like /p/, /b/ and /v/. The problem with such a definition for non-uniqueness is that one does not know what is the optimal method and quantization level for clustering the acoustic and articulatory spaces. A later study by Neiberg et al. [7] postulated that the different articulatory clusters should not only map onto a single acoustic cluster but should also map onto acoustic distributions with the same parameters, for it to be called non-unique. They found that phonemes like /p,t,k,s,z/ are highly non-unique. However, the method made Gaussian assumptions about the distribution of the clusters, and would underestimate the non-uniqueness for non-Gaussian distributions. Besides, the study had inconsistent scales of the measurement. We have therefore suggested an alternative method based on the conditional probability distribution on the joint acoustic-articulatory space.

2. Non-uniqueness as a Function of the Conditional Distribution

The proposed method deals with estimating the conditional probability function of the inversion, i.e., $p(X|Y)$ where ‘X’ is the articulatory space and the ‘Y’ is the acoustic vector at time instant ‘t’ in the acoustic space ‘Y’. By definition,

$$p_{X|Y}(x|y) = \frac{p_{XY}(xy)}{p_{Y}(y)} \tag{1}$$

In order to estimate this quantity, one needs to know the distribution of the joint articulo-audio data $(p_{XY})$ as well as the acoustic data, $p_{Y}$. These two spaces are approximated by Gaussian Mixture Models (GMMs). Thus Equation 1 can be expanded as

$$p_{X|Y}(x|y) = \sum_{m=1}^{M} C_m e^{(1)} \frac{1}{\sqrt{(\Sigma_m)^{-1}}} [(y - \mu_m)^T (\Sigma_m)^{-1} (y - \mu_m)] \tag{2}$$

where $\mu$, $\Sigma$, $e$ are parameters of the GMM. $M$ and $N$ are the number of mixtures for the two distributions respectively. Thus it is possible to estimate the conditional probability function of the articulatory space, given a particular acoustic data frame.

Since we consider the probability of the articulatory con-
configuration, given an acoustic data frame, it is obvious that there is variance in the mapping. This variance is a property of the statistical modeling itself as well as being due to artifacts in the measurement and the representation of the acoustics. We define non-uniqueness in the inversion of an instance of acoustic features as, the conditional probability function having more than one peak. This means that, if there is a high probability for more than one articulatory configuration, given an acoustic vector, then the mapping for that acoustic vector is non-unique. This definition, while relaxing the strictness of the definition of non-uniqueness in the deterministic sense, still provides an intuitively justifiable understanding of non-uniqueness in the statistical sense without making assumptions about the level of quantization and about the distribution of the data. At the same time, this definition of non-uniqueness distinguishes itself from the variance in the mapping.

The problem of finding all the modes in a GMM was addressed by Carreira-Perpinán [8]. The paper distinguished peak/bump (used interchangeably) finding from mode finding, conjecturing that the peaks could be called modes if the distribution was symmetrical around the peak. Several generalized “bump hunting” algorithms for high-dimensional functions were proposed by Friedman and Fisher [9]. Our implementation of the peak finding algorithm is a variation of the “Box Induction” strategy that they proposed. This is done by a top-down peeling search algorithm to find the total number of local maxima in the function $p_{X|Y}(X|y)$. Since such a search is computationally expensive, it was necessary to define initial search space boundaries and a nominal step size $x_{step}$ for the first approximation in the box search. The algorithm guarantees that it would find all the peaks which are $x_{step}$ away from each other, but can also find peaks closer than $x_{step}$, although no guarantee is given.

There could be one or more peaks in the conditional probability function. So we define the measure of non-uniqueness as the average spread of the peaks from the mean location of the peaks. For the articulator space $X^{Ar}$, the $Q$ local maxima are located at $\{M_q : 1 \leq q \leq Q\}$. Then the measure of non-uniqueness ($NU^{Ar}_i$) for articulator $X^{Ar}$ at time instant $t$, is proposed to be

$$NU^{Ar}_i = \frac{1}{Q} \sum_{q=1}^{Q} P_q (M_q - \mu^{Ar}_i)^2$$

(3)

where

$$P_q = \frac{p_{X|Y}(X^{Ar} = M_q|y)}{\sum_{q=1}^{Q} p_{X|Y}(X^{Ar} = M_q|y)}$$

(4)

are the normalized probability values and $\mu^{Ar}_i = \sum_{q=1}^{Q} P_q M_q$ is the mean location of the peaks. We can observe that the non-uniqueness would be zero if there is only one peak. The non-uniqueness does not necessarily increase with the number of peaks, but in fact with the spread of the peaks along the articulatory space. The units of this measure is the same as the units of the measurement of $X^{Ar}$. Thus the proposed measure of non-uniqueness has an intuitive scale. However, when one would like to find the non-uniqueness over several acoustic frames, the scaling is not even, because the relative importance of the distance between the different modes for different time instances depends on the variance in the predictability of the inversion for that time instance. If the variance is large, then a larger distance between the different modes is required for the non-uniqueness to be significant. The same is the case for different phonemes and different articulators. So that the non-uniqueness can be measured in perspective of its importance, a unit-less measure, Normalized Non-Uniqueness ($NNU^{Ar}_i$) is also proposed as

$$NNU^{Ar}_i = \frac{1}{Q} \sum_{q=1}^{Q} P_q (M_q - \mu^{Ar}_i)^T (\Sigma^{Ar}_i)^{-1} (M_q - \mu^{Ar}_i)$$

(5)

where $\Sigma^{Ar}_i$ is the variance of the conditional articulator space. The variance is estimated over the values in the $X^{Ar}$ space used for the the first approximation to obtain the peaks.

3. Experiments

The MOCHA-TIMIT [10] database with simultaneous acoustic and articulatory measurement for two speakers (male and female) has been used to make this study. The acoustic space was represented by 18 MFCCs of 11 acoustic windows of 25 ms each shifted by 10 ms. Thus each acoustic instance parameterized 125 ms of the acoustics. A larger length in time was taken for the acoustic space representation, in order to reduce the effect of non-uniqueness which can be resolved using information about the context. Principal Component Analysis (PCA) was performed on the 198 dimension acoustic features and 62 principal components were taken, which represented 98% of the variance in the features. The articulatory data consisted of 14 channels, which included the X and Y-axis trajectories of the 7 articulators, the Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (V). The articulatory data was low-pass filtered and down-sampled to 100 Hz, in order to correspond to the acoustic frame shift rate. Each articulatory data frame corresponded to the central time instant among the 11 acoustic frames. A GMM based inversion method suggested by Toda et al. [11] was trained on 80% of the data and used to predict the articulation from the acoustics for the remaining 20% of the data. The number of mixtures for which minimum error was obtained was selected as the optimum number of Guassians in the GMM. In this case it was selected to be 64 mixtures. Non-uniqueness was calculated for the entire database of the two speakers in the two dimensional space (along the X and Y directions of the mid-sagittal plane) for each of the 7 articulators. Some instances of the NNU (19 instances out of 228,037) were found to have extremely large values. This was so because the estimated conditional variance was extremely low for certain dimensions, resulting in ill-conditioned covariance matrices (the inverse of which is used in the calculation). These instances were discarded from the results. $x_{step}$ was nominally taken as 0.6% of the entire range of the articulator space.

4. Results

Fig.1 shows the conditional distribution of the acoustic-to-articulatory mapping for a single articulatory frame, which belongs to the phoneme /æ/. We can see a strong indication of non-uniqueness (presence of more than one peak) for the tongue tip, and tongue body. We can also see more than one peak in the distribution of the lower lip, lower jaw and tongue dorsum.

A comparison of the measure of NNU for different phonemes uttered by the male and female speaker are shown according to the different phoneme categories in Fig. 2. At first glance it is obvious that the consonants have a much higher non-uniqueness than vowels. This is specially valid for the dorsal stop consonants /k,g/ which can be explained by the presence of the soft palate above the tongue dorsum. This was similar to the findings in [7]. One can see that fricatives like /f,v,T/ and /θ/ also have high non-uniqueness. This maybe a slightly dif-
different finding in [7] where it was found that the fricatives /θ, ð/ have high non-linearity, but not very high non-uniqueness. The semivowels such as /ø, w, J/ are rather unique, but nasals, especially the velar nasal /ŋ/, have a higher non-uniqueness. Vowels and diphthongs in general have low non-uniqueness. The estimates of the non-uniqueness vary for the speakers, but trends are similar for most phonemes. Notable exceptions are the diphthongs where the male speaker shows a penchant for producing a higher number of non-unique articulator configurations for similar acoustics.

Table 1: Table showing the average absolute non-uniqueness in mm for different phoneme categories, estimated only for those instances where more than one peak has been observed in the conditional probability distribution.

<table>
<thead>
<tr>
<th></th>
<th>LI</th>
<th>UP</th>
<th>LP</th>
<th>TI</th>
<th>TB</th>
<th>TD</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>0.42</td>
<td>0.53</td>
<td>0.86</td>
<td>1.50</td>
<td>1.75</td>
<td>1.73</td>
<td>0.40</td>
</tr>
<tr>
<td>Consonants</td>
<td>0.47</td>
<td>0.38</td>
<td>0.91</td>
<td>1.72</td>
<td>1.71</td>
<td>1.29</td>
<td>0.40</td>
</tr>
<tr>
<td>Semi-vowels</td>
<td>0.56</td>
<td>0.52</td>
<td>0.95</td>
<td>1.44</td>
<td>1.59</td>
<td>1.54</td>
<td>0.46</td>
</tr>
<tr>
<td>Fricatives</td>
<td>0.32</td>
<td>0.41</td>
<td>0.80</td>
<td>1.15</td>
<td>1.10</td>
<td>1.08</td>
<td>0.31</td>
</tr>
<tr>
<td>Diphthongs</td>
<td>0.50</td>
<td>0.38</td>
<td>1.08</td>
<td>1.63</td>
<td>1.94</td>
<td>1.89</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 1 shows the mean absolute non-uniqueness estimates for different phoneme categories, estimated only for those instances where there is evidence of non-uniqueness, in the form of more than one peak in the conditional probability distribution. It can be seen that average non-uniqueness is in the same range for all categories and the difference is only between the different articulators. This shows that the major difference between the different phoneme categories in terms of non-uniqueness is the number of instances of non-unique articulator configurations found in the data. It follows that vowels do not have many sounds which can be produced by non-unique articulator configurations, while consonants have more of such instances.

When different articulators are considered, Fig. 2 shows that the tongue-tip contributes to most of the non-uniqueness of the vowels and diphthongs and phonemes such as /θ, ð/. The velum seems to be highly unique for almost all classes of phonemes, except for the nasals like /m, n/ and /ŋ/. The upper lip and tongue dorsum contribute to most of the non-uniqueness of the phonemes /p, b/ and /k, g/ respectively. These findings are interesting because the non-uniqueness, in most cases, seems to be higher for those articulators which are important for the production of the sound. This is contrary to the intuitive notion that non-uniqueness would be caused by co-articulation of unimportant articulators. In order to understand this effect, a further study has been carried out with respect to the connection between the important regions in the articulator trajectories and non-unique instances in the data.

5. Relation with Critical points

It is well known that certain articulators are more important than others for producing particular sounds. Articulators move between consecutive target positions, which could be a few phonemes apart. During the pronunciation of a certain sound, the non-critical articulators are in the process of moving towards their next target position. One may hypothesize that non-uniqueness may occur for these non-critical articulators, because the transition path depends on the locations of the targets and would not affect the acoustics critically.

Ananthakrishnan and Engwall [12] suggested that articulatory targets are those locations in the articulatory trajectory...
which correspond to minimum velocity in motion or maximum change in angle, called critical points. The relationship between the distance of an instance from the nearest critical point, in terms of number of frames (time), and the NNU was calculated for the entire database of the two speakers. The results, which are an average value over all the occurrences, are displayed in Fig. 3. The hypothesis was that the non-uniqueness would increase with the distance from the critical points. We used a weighted least squares linear regression with weights proportional to the number of points used to estimate the slope. The hypothesis thus translates to the slope of the regression having a positive value. The results of the fit are displayed in Table 2. The null hypothesis could not be rejected since the p-value was greater than 0.5 for most of the articulators for both the speakers (when it must be less than 0.05 in order to prove it). In fact we can see that the estimated slope is negative for all the articulators except the upper lip for the male speaker. We see similar trends for both the absolute non-uniqueness measure as well the number of peaks per data frame. This shows that the non-uniqueness probably has a tendency to increase for instances which are closer to the critical point rather than the other way round. This study infers that the non-uniqueness may not be due to the co-articulation effects of unimportant articulators. This observation leads to an un-intuitive inference, since one would definitely expect higher non-uniqueness for unimportant articulator positions. The reason for this observation needs to be investigated further. A starting point would be to understand that given a mean and variance, the distribution with the maximum entropy for all real valued random variables is a Gaussian distribution. So in fact, having multiple peaks in the distribution makes those articulatory configurations more predictable than having a single symmetric peak for the same variance. Viewing our finding with this perspective, one needs to explore whether this observation is a result of the statistical modeling of the acoustic-to-articulatory mapping, an artifact of the definition of critical points or a true attribute of the speech production mechanism.

6. Conclusions and Future Work

This paper shows a novel way to estimate the non-uniqueness of the acoustic-to-articulatory mapping based on finding peaks in the conditional probability function. The experiments showed that non-uniqueness is higher for stop consonants, fricatives and nasals as compared to vowels, liquids and diphthongs. It was also found that non-uniqueness is not higher for the unimportant articulators. Future work needs to be done on assessing the impact of the type of modeling on the results obtained. The knowledge of the existence of the non-uniqueness could be used to make better estimates for the inversion. It seems counter-intuitive as to why unimportant articulators do not show higher non-uniqueness. This observation must be investigated further, to see its implications on our understanding of the production of speech.

7. Acknowledgements

This work is supported by the Swedish Research Council project 80449001, Computer-Animated Language Teachers.

8. References