Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

Exploring the Predictability of Non-Unique Acoustic-to-Articulatory Mappings

G. Ananthakrishnan, Olov Engwall, Daniel Neiberg
KTH Royal Institute of Technology, Stockholm, Sweden

This paper explores statistical tools that help analyze the predictability in the acoustic-to-articulatory inversion of speech, using an Electromagnetic Articulography database of simultaneously recorded acoustic and articulatory data. Since it has been shown that speech acoustics can be mapped to non-unique articulatory modes, the variance of the articulatory parameters is not sufficient to understand the predictability of the inverse mapping. We, therefore, estimate an upper bound to the conditional entropy of the articulatory distribution. This provides a probabilistic estimate of the range of articulatory values (either over a continuum or over discrete non-unique regions) for a given acoustic vector in the database. The analysis is performed for different British/Scottish English consonants with respect to which articulators (lips, jaws or the tongue) are important for producing the phoneme. The paper shows that acoustic-articulatory mappings for the important articulators have a low upper bound on the entropy, but can still have discrete non-unique configurations.

Index Terms: Acoustic-to-articulatory inversion, entropy of GMM (Gaussian mixture model), many-to-one-mapping.

I. Introduction

Models approximating the vocal tract with a lossless tube of varying cross-sectional area [1], [2] have indicated the possibility of non-uniqueness in the inverse mapping between acoustic and articulatory parameters, showing that the inverse mapping of the acoustics of a vowel is to a class of vocal tract area functions. Atal *et al.* [3] further demonstrated the nonuniqueness in the inverse mapping (or inversion)where an entire region in the articulatory parameter space, which they referred to as ‘fibers’, could correspond to a single point in the acoustic parameter space. This was, in particular, possible because the dimensionality of the acoustic parameter space that they considered (the first three formants) was smaller than that of the articulatory parameter space. The mapping they computed was from an under-determined set of equations and therefore ambiguous. It was also pointed out that not all these non-unique articulatory configurations were viable in a real physical system and could be avoided during acoustic-to-articulatory inversion by placing certain constraints on the articulatory parameter space. The theoretical basis for non-uniqueness was confirmed with bite-block experiments, which showed that speakers are capable of producing sounds perceptually close to the intended sounds even though the jaw is fixed in an unnatural position [4], by compensating with some other articulators (typically the tongue tip).

In order to study acoustic-to-articulatory non-uniqueness in a more realistic scenario, large single speaker databases of simultaneously collected acoustics and articulatory measurements, based on methods like X-ray microbeam, Ultra Sound (US), Electromagnetic Articulography (EMA) or Electropalatography (EPG) were necessary. Using such techniques, the problem of compensation and tongue-jaw synergy was further investigated in [5], where the phonemes /*i, a, l*/ and /*s*/ were elicited in different phonemic contexts and recorded using EPG and US. The contexts were assumed to be some sort of natural constraints, similar to a bite-block. The authors found that the jaw opening and the tongue constriction compensated for each other while keeping the acoustic phonetic identity (larger jaw openings corresponded with raising the tongue to a higher degree). Further, even though certain articulators were not very important (low specification articulators) for production of the sound, they were constrained by trying to compensate for the important articulator (high specification articulators), which was perturbed due to the co-articulation with neighboring phonemes. Kroos *et al.* studied the same phenomenon with two forms of natural perturbation, namely co-articulation and loud speech [6]. They measured the positions of the articulators for the post-alveolar consonants /*t, d, s, l*/ and */n/*, using EMA coils on the tongue and jaws of 5 subjects. In loud speech, the

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

jaw is considered to be constrained to open larger than normal. However, this was not true for all phonemes, especially not the sibilant */s/*. The consonants */l/* and */n/*, showed an inverse correlation between jaw opening and tongue constriction.

Guenther *et al.* [7] and Nieto-Castanon *et al.* [8] performed studies on the production of American English phoneme */r/*. They found two modes of production (bunched and retroflexed) where the vocal tract shape appeared different for different productions, while keeping the acoustic correlate corresponding to the phoneme more or less constant. It was concluded that this was not a case of compensatory behavior, but of different vocal tract shapes corresponding to the two modes of production. Thus one can claim that there are two main means for the occurrence of non-uniqueness in the acoustic-to-articulatory mapping. The first is due to compensatory behavior between the different articulators (also called 'synergy') with positions of individual articulators varying over a continuum [5], [6]. The second occurs when discrete articulator positions can be mapped to similar acoustic correlates [7], [8].

While the above studies were based on specially elicited responses, studies made on continuous read speech (without any further constraints) need to make use of statistical models to study this phenomenon. This is because the acoustic correlates for read speech cannot be constrained to be exactly the same for different instances of the same phoneme. Thus, the acoustic features need to be clustered in some sense in order to limit their variation. Qin and Carreira-Perpiñán [9] defined the mapping to be non-unique if, for a particular acoustic cluster, the corresponding articulatory mapping may be found in more than one cluster. Further studies [10] showed non-uniqueness by finding multiple modes of articulatory distributions mapping onto a single acoustic distribution for almost all phonemes in the databases. In [11] non-uniqueness was estimated ranging from 0.4% of the data for the upper lip to 21.7% of the data for the tongue tip.

In our previous work [12], we defined an acoustic-articulatory mapping for one acoustic frame, x_t, to be non-unique if the conditional distribution $\rho_{Y|X}(y|x_t)$ where y and x are articulatory and acoustic vectors, respectively) had more than one peak. The conditional distribution was modeled as a Gaussian Mixture Model (GMM). We performed multiple peak detection on the conditional probability density function and were thus able to find more than one high probability region in the articulatory parameter space for acoustic frames belonging to almost all phonemes. The study also tried to inquire into whether the non-uniqueness was higher for those articulators or parts of the trajectory that are not important for the production, as suggested in some studies (e.g., [3], [5]). However, we could find no statistical evidence to show that the non-uniqueness was lower for the points in the articulatory trajectory closer in time to what we considered critical for the pronunciation. This observation needed a good explanation, being rather non intuitive.

The above studies based on statistical models of acoustic-to articulatory inversion largely consider discrete modes of articulator positions corresponding to the findings in [7]. The variance of the articulator positions was considered a sufficient description of the predictability for the continuous case. However, a Gaussian distribution has the maximum entropy (upper bound) among all real valued distributions with the same variance [13]. A normal distribution is thus called a maximum entropy distribution. In other words, consider two variables, one a single normal distribution and the other with multiple modes. If both these distributions have the same variance, the normal distribution will have a higher prediction entropy than the distribution with multiple modes. Thus, given that the acoustic-to-articulatory inverse mapping is non-unique with several frames having multiple modes in the conditional distribution, calculating the variance alone does not give an acceptable account of the predictability in the mapping.

This paper follows our previous work while trying to answer the following main questions:

1) Is the probability of finding non-uniqueness in the mapping higher for important articulators corresponding to the place of articulation?
2) How does one distinguish the discrete [7] and continuous [6] cases of non-uniqueness?
3) What is the relationship between the acoustic-articulation predictability and non-uniqueness of the mapping?

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

In Section II of this paper, we show how to estimate the conditional probability distribution of the articulatory positions (i.e., the probability of finding the articulator at different positions), given the acoustic features, for every time frame in the database. The conditional entropy (i.e., the entropy of the conditional distribution) is a good indicator of the predictability of the mapping. The larger the range of positions the articulators can assume, given the acoustics, the lower the predictability of the mapping and higher the entropy. We propose that the estimation of the upper bound on the conditional entropy provides an insight into the extent of non-uniqueness in the continuous [6] case as well as when there is more than one mode in the probability distribution of articulatory positions, i.e., the discrete [7] case. We also describe a method to estimate which articulator is important for production of a certain phoneme. Using acoustic-articulatory data described in Section III a study relating the non-uniqueness in the mapping of each articulator to the importance of the particular articulator in the production of the phoneme is made in Section IV. Finally, the conclusions of the study as well as future directions for research are presented in Section V.

II. Theory and methods

This section largely deals with describing the mathematical theory behind how we estimate the non-uniqueness of the acoustic-to-articulatory mapping in both the discrete (multi-modal) and continuous senses using statistical models. The central idea is the method to estimate the probability distribution of the articulatory vector, given single acoustic vector in the data.This can be estimated while keeping the acoustic vector constant, without considering a window of variation as had been done in previous empirical studies in the domain. As described in the previous section, finding more than one discrete high probability region in the conditional distribution is an indication of non-uniqueness in the mapping in the multi-modal sense.

There are several ways to model an unknown probability distribution; we use GMM because it is easy to generalize the GMM to an arbitrary multi-variate real-valued density function, and estimate its parameters. Besides, GMMs have been used successfully in several previous studies to model the acoustic-to-articulatory mapping (e.g., [12], [14]). While the previous studies model the joint acoustic-articulatory distribution, we model the conditional distribution as a GMM, as described in Section II-A.

There are many ways to detect peaks in a multivariate density function (e.g., including non-parametric bump search, parametric optimization and other methods). In this article we describe one such method to detect the total number of maxima in a distribution in Section II-B. While there may be other methods of coming to similar results, the method used in this article is based on calculating the upper bound of the entropy of the distribution.

Knowing this upper bound also gives us a fair idea about the predictability of the conditional distribution, and hence the acoustic-to-articulatory mapping, i.e., an idea about how difficult or easy it is to predict the articulation, given an acoustic vector. This provides us with the perfect means of measuring non-uniqueness in the continuous sense.

A. Non-Uniqueness as a Function of the Conditional Distribution

One can model the conditional probability density function $\rho_{X|Y}(y|x_t)$ of the articulator space $y \in \mathbb{R}^d$, for a given acoustic Vector $x_t \in \mathbb{R}^D$, at time frame t: $1 \le t \le T$ as a GMM (λ_M) with M Gaussian components as

$$\rho_{Y|X}(y|x_t; M) = \sum_{m=1}^{M} \rho(x_t, \lambda_M)\rho(y|x_t, m, \lambda_M) \quad (1)$$

where

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

$$\rho\left(m|x_t,\lambda_M\right) = c_{m,t}^{Y|X} = \frac{c_m^X N(x_t;\mu_m^X,\Sigma_m^X)}{\sum_{n=1}^{M} c_n^X N(x_t;\mu_n^X,\Sigma_n^X)} \quad (2)$$

and c_m are the weights for the individual Gaussian components and

$$\rho\left(y|x_t,m,\lambda_M\right) = N(y;\mu_{m,t}^{Y|X}\Sigma_m^{Y|X} \quad (3)$$

$\mu_{m,t}^{Y|X}$ and $\Sigma_m^{Y|X}$ are the mean vector and the covariance matrix of the conditional probability distribution.

B. Estimating the Entropy of the Conditional Distribution Modeled as a GMM

A Gaussian distribution has the maximum entropy among all real valued distributions with the same variance [13]. Huber *et al.* extended this to postulate that a distribution modeled with the number of Gaussians that correspond to the exact number of modes in the distribution would have the lowest upper bound of the entropy [15]. If the distribution has two modes, then modeling the data with only one Gaussian gives a higher upper bound on the entropy estimate than modeling it with two Gaussians. On the other hand, if the distribution actually has only one mode, then modeling it with two Gaussians gives a higher upper bound on the entropy estimate. This property is illustrated in Figs. 1 and 2.

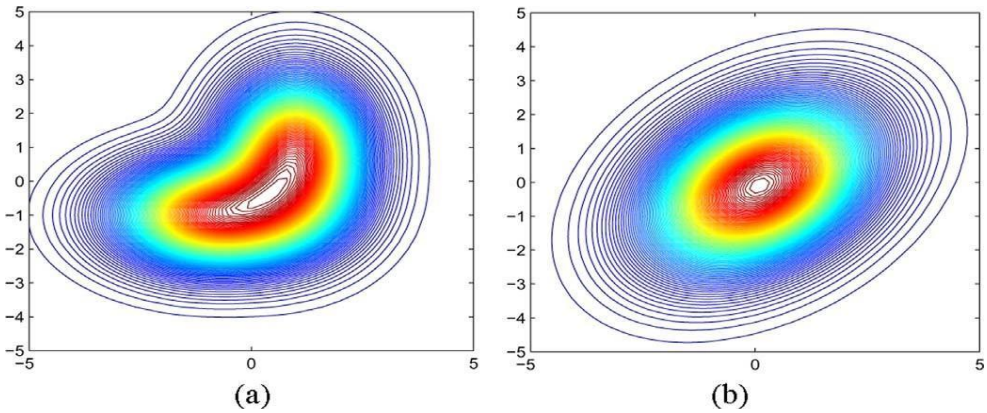

Fig. 1. A non-Gaussian distribution (with an arbitrary scale) with a single mode. The contour lines represent equal probability regions. The estimated entropy upper bound when modeling it with a single Gaussian component *(M=1)* is 3.60 (figure to the right) while when modeling it with 2 Gaussian components the upper bound on entropy is 3.66 (figure to the left). Thus even though the fit is better with *M=2*, the upper bound of the entropy is higher. The variance is the same for both models, the log of which is proportional to the entropy estimated with *M=1*.

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

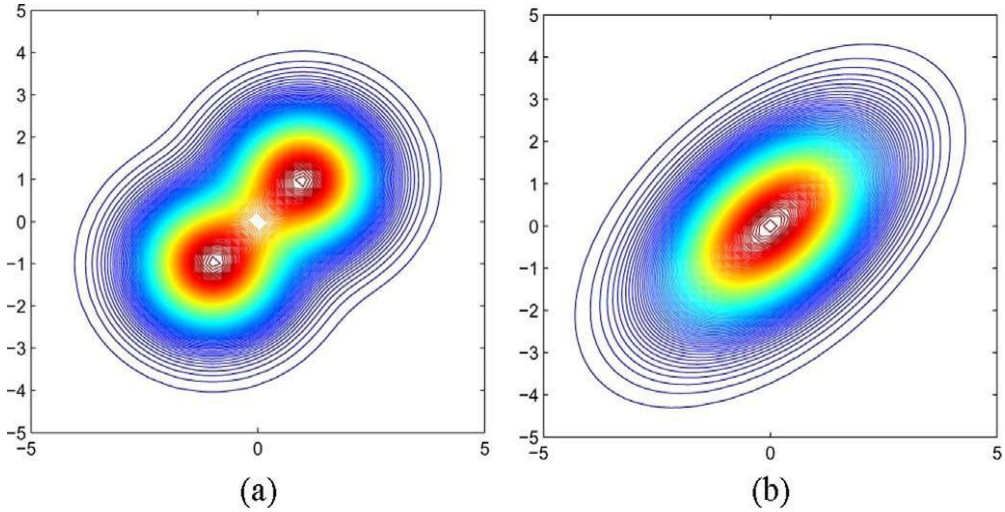

Fig. 2. A non-Gaussian distribution with two modes. When modeling it with 2 Gaussian components *(M=2)* the estimated entropy upper bound is 3.31 (figure to the left) and lower than when modeling it with one Gaussian component, where the estimated entropy upper bound is 3.38 (figure to the right). Since the estimated variance is exactly the same, irrespective of the number of Gaussians it is modeled with, the predictability is higher than what the variance suggests.

It should be noted that a reduced model that has the lowest upper bound on the entropy is not an optimal model in terms of the log-likelihood of the goodness of fitting the distribution function, as can be seen from Fig. 1. Using a single mode for modeling this distribution provides a better (tighter upper bound) entropy estimate, but fails to model the distribution accurately. So the number of Gaussians used for minimizing the upper bound of the entropy is not the optimum number of Gaussian components for fitting the distribution. On the other hand, this parameter is ideal for selecting the number of modes as is required to detect discrete modes of articulation. If the true distribution of the data indeed had only one mode, the upper bound of the entropy would be equal to the true entropy and is proportional to the log of the variance.

We use the method suggested in [15] to find the upper bound of the entropy of a random variable, in this case the conditional distribution of the acoustic-articulatory mapping, modeled by a GMM (cf. Appendix II-A). This method assumes an arbitrarily high number of Gaussian components, *M*, used to model the conditional distribution, $\rho_{Y|X}\left(y^a|x_t\right)$ for articulator *a*, at time frame *t*. By merging the Gaussian components (cf. Appendix A) in such a way that the merged distribution differs the least from the whole distribution in terms of the Kullback-Leibler Divergence [16], we can make an approximation of the same probability function with a smaller number of components, *M-1*. The upper bound of the entropy, calculated with the decreased number of components may be either lower or higher than the upper bound of the entropy calculated with *M* components. If the upper bound is reduced, then we attempt decreasing the number of components further. On the other hand, if the upper bound of the entropy increases, then it means that we need a minimum of *M* Gaussians to model the distribution. If the lowest entropy upper bound, $\Xi_u(y^a|x_t)$, is for modeling the distribution with a single Gaussian component, then we can say the inversion mapping is uni-modal. If modeling the distribution with a number of Gaussian components $\aleph_{a,t} > 1$ provides the lowest entropy upper bound, then we can infer that the inversion mapping is non-unique in the discrete (multi-modal) sense for the said frame and articulator. Details about the mathematics of this procedure are presented in Appendices A and B.

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

C. Important Articulators

If an articulator position is crucial for distinguishing it from other phonemes in the language, one can say that the articulator is important for the production of the phoneme. We use this idea to find out which articulators are important for the production of certain phonemes. Here a crucial position is determined by whether the said articulatory position can improve a classifier to distinguish between one phoneme and the rest. This is similar to a parameter selection or parameter weighting approach where one estimates how important a parameter is for the classification task. Here the classification task is separating one phoneme from the rest of the phonemes in the database. For the current work we used the Support Vector Machine Projection Recurrent Feature Elimination (SVM-Projection RFE) algorithm [17], in order to rank the articulatory parameters according to their importance. This can be achieved by sorting the angles made by the SVM hyperplane with respect to the articulatory dimensions.

One problem is that the number of instances (frames) for a single phoneme is much less than the frames from all the other phonemes and the SVM is always biased towards the class with the higher number of patterns. Another problem is the presence of outliers, which could cause a change in the orientation of the hyperplane. To avoid these two problems, the data in both classes (one phoneme versus the rest) are clustered using *K*-means clustering. The SVM is applied on the *K* cluster-centroids from each class. Thus, outliers are filtered out and the number of patterns from each class remains constant.The importance of each articulator can be established from $w_p \in R^{\tilde{A}}$ for each phoneme (Ã=2A, i.e., twice the number of articulators).

III. Data and experiments

In order to study the presence of non-uniqueness in an empirical sense, we used the MOCHA-TIMIT database [18] with simultaneous acoustic and articulatory measurements for two speakers, ‘msak’ (male speaker) and ‘fsew’ (female speaker). The database for each speaker consists of 460 phonetically balanced short sentences. The sentences were orthographically transcribed and then aligned with the respective phonemes (46 phonemes in total) using a Hidden Markov Model (HMM) based forced alignment. This study was restricted to 24 consonants for which the selection of the most important articulator is easily verified and intuitive.

The acoustic parameterization was a difficult choice. The acoustic parameters should be able to represent the perception of articulation in a reasonable manner. Our study pertains to consonants, for which formant frequencies do not provide adequate representation of their acoustics. While low dimensionality of the acoustic space could create null spaces in the acoustic-articulatory transformation, a larger number of acoustic parameters would make the parameter space very sparse and the GMM models would be ill-conditioned. Last but not the least, while studying transients like unvoiced stop consonants, the silence region before the burst will definitely be non-unique. Therefore we need an acoustic parametrization longer than a single frame.

In order to solve the above problems the following acoustic parameterization scheme was selected. We first chose the first 18 Mel Frequency Cepstral Coefficients (MFCCs) from acoustic windows of 25 ms shifted by 10 ms. In order to avoid non unique cases related to the silence region before the burst for plosives or during the central noisy phase of the fricatives, the 18 MFCCs were then concatenated for 11 consecutive acoustic frames. Thus, each acoustic instance parameterized 125 ms of the acoustic data. While the choice of 125 ms of acoustics is arbitrary, it is a reasonable compromise between not including information outside that of the phoneme and making it longer than silence regions in the plosives. A small discussion about the effect of choosing a smaller window is discussed in Section IV-D.

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

Further, Principal Component Analysis (PCA) was performed on the 198-dimensional acoustic features and 62 (in Section II-A, *D=62*) principal components were taken, which represented 98% of the variance in the concatenated features. This method of dimensionality reduction ensured that, even though one value of the acoustic parameters may have variations in the missing dimensions, the variation would be lower than that of the considered dimensions. Thus, every acoustic parameter vector may in fact represent a small variation in the actual acoustic signal, but we assume that this unparameterized variation is not significant as compared to the variations caused by changes in articulation. Finally, it is relevant to mention that many state-of-the-art algorithms for statistical based inversion (e.g., [19]) employ similar types of acoustic parametrization.

The articulatory data consisted of 14 channels of EMA measurements, which included the *X*-axis and *Y*-axis trajectories of coils placed on 7 articulators (in *B, A=7*): the Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), TongueBack (TB),TongueDorsum (TD) and Velum (VE) along the midsagittal plane of the vocal tract. The measurements were made with a resolution of 0.01 millimeters, but the effective resolution was estimated to be around 0.43 mm on average [20]. The drift in the EMA data was corrected by the algorithm suggested in [21]. 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. There were 10,358 data frames for the male speaker and 12,372 data frames for the female speaker after removing the frames corresponding to silence.

The articulatory space for each GMM was the two-dimensional space (in Section II-A, *d=2*) (along the *X* and *Y* directions of the midsagittal plane) for each of the 7 articulators. Thus, for each speaker we had 7 conditional articulatory-acoustic models for $\rho_{Y|X}(y|x)$, one for each articulator a. The GMM was modeled with an arbitrarily high number of mixture components, in this case *M=128* components. We were forced to restrict ourselves to 128 Gaussian components due to memory restrictions in our system. However, since this was much larger than the typical number of modes expected (i.e., between 1 and 5), the restriction was reasonable. The GMM with full covariance matrices for the Gaussian component was trained with 50 different initializations using the Expectation Maximization (EM) algorithm [22] with 100 iterations each or until convergence. Of the different initializations, we picked the model with the lowest log-likelihood error to avoid local minima in the EM training. In order to avoid ill-conditioned covariance matrices, we trained the GMMs for one articulator at a time and set a variance flooring threshold to be 0.001% of the standard deviation of the data. We calculated the upper bound of the conditional entropy of the prediction mapping, $\Xi_u(y_a|x_t)$, as well as the number of articulator modes $\aleph_{a,t}$, for each acoustic vector in the data.

Based on the SVM-Projection RFE method (cf. Section II-C), consonants were assigned the most important articulator from among the Jaw and Lips (LJ, UL and LL), Tongue Tip (TT), Tongue Back (TB) and Tongue Dorsum (TD), depending on which of the 14 articulatory parameters had the highest weight. A Radial Basis Function (RBF) kernel was used with *K=200*, the number of articulatory clusters before we ran the algorithm.

IV. Results

Table 1 shows the percentage of frames that are non-unique in the discrete sense. These results closely follow the number of peaks in the conditional distribution calculated in [12]. We find a difference between the two methods only for 0.001% of the frames in the database. Though a peak in the probability distribution is not strictly equivalent to a mode in the distribution, peaks and modes can be considered equivalent for the present problem since the representation of the distribution is made by a reduced number of Gaussian components. We therefore use these terms interchangeably in this article.

As can be seen from Table 2, the upper bound of the entropy is more or less constant for the different articulators in contrast to the percentage of frames with more than one mode (cf. Table 1, where there is a large difference in the number of modes for different articulators). The maximum range of variation and therefore minimum predictability

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

is for the upper lip coil (UL), while the tongue tip coil (TT) has the minimum average upper bound of the entropy. The female subject in general shows a higher average entropy (upper bound), again the opposite of the multi-modal case in Table 1. The average upper bounds of the conditional entropy for the uni-modal case are not substantially different as compared to the case with multi-modal articulatory positions. The notable exceptions are the LJ and TT coils where the average entropy is substantially lower for the multi-modal case (especially for the female speaker). This shows that even if there may be only one articulatory mode of production, the range of positions they assume may not necessarily be lower than when multiple modes of articulation exist.

Table 3 shows the variation of Ξ_u over different phoneme classes. The articulator positions of stop consonants and sonorants consonants (like nasals, approximants and liquids) are more difficult to predict from the acoustics (given their higher mean Ξ_u) than fricatives.

Table 1. The percentage of frames with more than one mode for different articulators

Speaker	LJ	UL	LL	TT	TB	TD	V	Total
msak	4.01%	7.5%	6.27%	17.98%	10.76%	12.4%	2.7%	8.82%
fsew	7.16%	5.75%	6.77%	13.5%	6.8%	10.2%	4.1%	7.7%

Table 2. The mean upper bounds of the conditional entropy for different articulator coils

Speaker	# Modes	LJ	UL	LL	TT	TB	TD	V	Total
fsew	=1	1.73	2.04	1.76	1.29	1.35	1.51	1.26	1.57
fsew	>1	1.01	1.70	1.51	0.91	1.5	1.54	1.30	1.31
fsew	All frames	1.68	2.02	1.74	1.24	1.36	1.51	1.26	1.55
msak	=1	1.49	1.73	1.33	1.06	1.26	1.36	1.5	1.40
msak	>1	1.33	1.71	1.39	1.00	1.32	1.53	1.62	1.34
msak	All frames	1.48	1.72	1.33	1.05	1.27	1.38	1.50	1.39

Table 3. Comparison of non-uniqueness measures for different phoneme classes

Speaker	Vowels	Diphthongs	Stop Consonants	Fricatives	Other Sonorants
% of frames with multiple modes from phoneme classes					
fsew	5.3%	5.7%	11.9%	12.4%	11.5%
msak	6.2%	5.8%	12.3%	13.2%	10.4%
Average upper bound of entropy for different phoneme classes					
fsew	1.57	1.55	1.62	1.46	1.58
msak	1.42	1.46	1.47	1.25	1.47

A. Comparative Analysis of Non-Uniqueness

When comparing individual frames, frames with multiple modes did not exhibit significantly higher Ξ_u as compared to uni-modal frames in general, as indicated by the results in Table 2. If we want to study the relationship between $\aleph$ and Ξ_u for every articulator and phoneme, we have some difficulties because some articulators and phonemes tend to have a higher probability of having multiple modes or have a higher Ξ_u with respect to other articulators or phonemes. For example, more frames show multi-modality for the TT coil when compared to the other articulator

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

coils. Similarly, stop consonants more frequently have non-unique articulator positions than fricatives. Hence, we need to be able to compare the frequency of non-uniqueness for an articulator a_1 for phoneme p_1 with that of articulator a_2 for phoneme p_2 (e.g., we may need to compare the non-uniqueness of the UL coil for phoneme */p/* with that of the TT coil for phoneme */t/*). Thus, in order to have parity in comparison, we need to perform a two-way normalization across articulators and phonemes. The two-way normalized frequency of non-unique occurrences ($nf^{\aleph}_{a,p}$) of an articulator *a*, occurring within a phoneme *p* is by first normalizing with respect to the phoneme and then with respect to the articulator as follows.

$$nf^{\aleph}_{a,p} = \frac{\left(\frac{\#\{\forall\, t:(\aleph_{a,t}>1 \cap L(t)=p)\}}{\#\{\forall\, t:(\aleph_{a,t}>1\}}\right)}{\left(\frac{\sum_{a=1}^{A}\#\{\forall\, t:(\aleph_{a,t}>1 \cap L(t)=p)\}}{\sum_{a=1}^{A}\#\{\forall\, t:\aleph_{a,t}>1\}}\right)} \qquad (4)$$

Where *L(t)* is the phonemic label of frame *t* and *#{.}* is the number of elements in the set. Thus $nf^{\aleph}_{a,p}$ indicates how frequent it is for an articulator *a* to occur in discrete non-unique articulatory modes when the said frame is labeled as phoneme *p*, in relation to other phonemes, and to the other measured articulators. Similarly, the normalized entropy (upper bound) is calculated as follows

$$n\Xi_{a,p} = \left(\frac{\mu\left\{\Xi\left(y^a|x_t\ \forall t{:}L(t)=p\right)\right\}-\mu\left\{\Xi\left(y^a|x_t\ \forall t\right)\right\}}{\sigma\left\{\Xi\left(y^a|x_t\ \forall t\right)\right\}-\mu\{\mu\{\Xi\left(y^a|x_t\ \forall t\right)\}\forall\alpha\}}\right) \qquad (5)$$

where $\mu\{.\}$ is the mean of the set and $\sigma\{.\}$ is the standard deviation of the set.

B. Relationship Between the Two Types of Non-Uniqueness

Figs. 3 and 4 show the comparison between the normalized non-uniqueness ($nf^{\aleph}_{a,p}$) and the normalized average conditional entropy $n\Xi_{a,p}$ (upper bound) for each articulator *a* and phoneme *p*. The different phonemes are located in the $nf^{\aleph} - n\Xi$ plane. The different phonemes are clustered in this plane, using *K*-means clustering into 2 to 4 clusters in order to elucidate some aspects of the observations. The ASCII symbols used to denote the phonemes in the figures are explained in terms of IPA symbols in Table 4.

Consider Fig. 3(a) and (b), showing the comparison between the normalized non-uniqueness frequency and relative entropy for the LJ coil. For both the speakers, phonemes such as 's' (*/s/*), 'z' (*/z/*), 'th' (*/θ/*) have a low $n\Xi$, but high $nf^{\aleph}$ for both speakers. Studies on production strategies [23] indicate that the jaw position is important for these fricatives, which is corroborated by the observation of low relative entropy.

Fig. 3(c)–(f) show similar plots for the UL and LL coils.The three labial phonemes *'p'*, *'b'* and *'m'* (*/p, b, m/*) form a distinct cluster with high $nf^{\aleph}$ and low $n\Xi$. This is a strong example of high $nf^{\aleph}$ for articulators that are important for producing the phoneme, as decided by the SVM weights. Phoneme 'ω' (/w/) for which the lips are critical for production has a low $n\Xi$ as well as a low $nf^{\aleph}$, especially for the LL coil. On the other hand, phonemes for which the lips are not important for production such as '*zh*', '*sh*', '*s*', '*z*', '*jh'*, '*ch*', '*t*', '*d*', '*y*' (/ʒ, ʃ, s, z, dʒ, tʃ, t, d, j/) have a high $n\Xi$, but a low $nf^{\aleph}$.

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

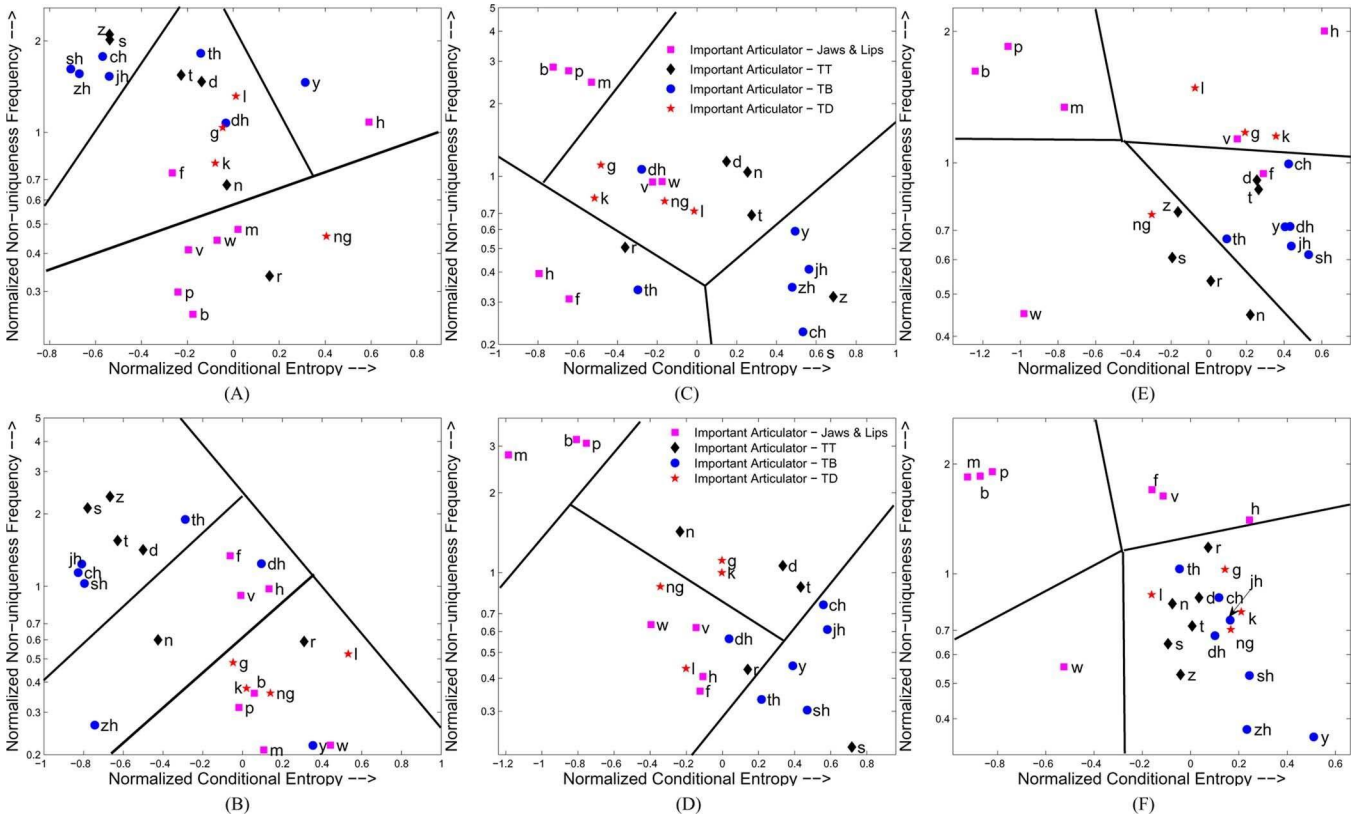

Fig. 3. Comparison of the $nf_{a,p}^{\aleph}$ and normalized $n\Xi_{a,p}$ for articulators Lower Jaw (LJ), Upper Lip (UL) and Lower Lip (LL), depending on the phonemes. The phonemes are clustered using K-means clustering, the straight lines representing boundaries of these clusters. The symbols used depend on the articulator that is considered important (based on the SVM hyperplane orientation) when uttering the particular phoneme. (A) Lower Jaw (LJ), male speaker (B) Lower Jaw (LJ), female speaker (C) Upper Lip (UL), male speaker (D) Upper Lip (UL), female speaker (E) Lower Lip (LL), male speaker (F) Lower Lip (LL), female speaker.

Fig. 4(a)–(f) show the comparisons for the coils on the tongue, TT, TB and TD. For both speakers, the alveolar and post-alveolar consonants, '*zh*', '*sh*', '*s*', '*z*', '*jh*', '*ch*', '*t*' and '*n*' (/ʒ, ʃ, s, z, dʒ, tʃ, t, n/) are seen to have average or above average $nf^{\aleph}$ but low $n\Xi$ for the TT and TB coils. The low $n\Xi$ is expected, since the tongue tip and tongue back are important for the production of these phonemes. On the other hand, labial sounds like */p, b/* and */m/* have high $n\Xi$ and low $nf^{\aleph}$, the opposite of the situation for the UL and LL coils. The velar phonemes '*k*', '*g*' and '*ng*' (/k, g, ŋ/) have low $n\Xi$ and high $nf^{\aleph}$ for the TD coil. Here the view that the tongue dorsum is important while pronouncing these phonemes is vindicated with the low $n\Xi$ estimate. However, the articulators important to produce these phonemes are seen to have a higher $nf^{\aleph}$. However, phonemes for which the tongue dorsum may not be important such as '*f*', '*v*' and '*w*'(/f, v, w/) have a high $n\Xi$ and also a high $nf^{\aleph}$. In almost all the plots, although there are differences in exact location of the different phonemes in the $nf^{\aleph} - n\Xi$ plane, the two speakers are seen to show similar strategies for their tongue movements, with similar patterns emerging for almost all the consonants.

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

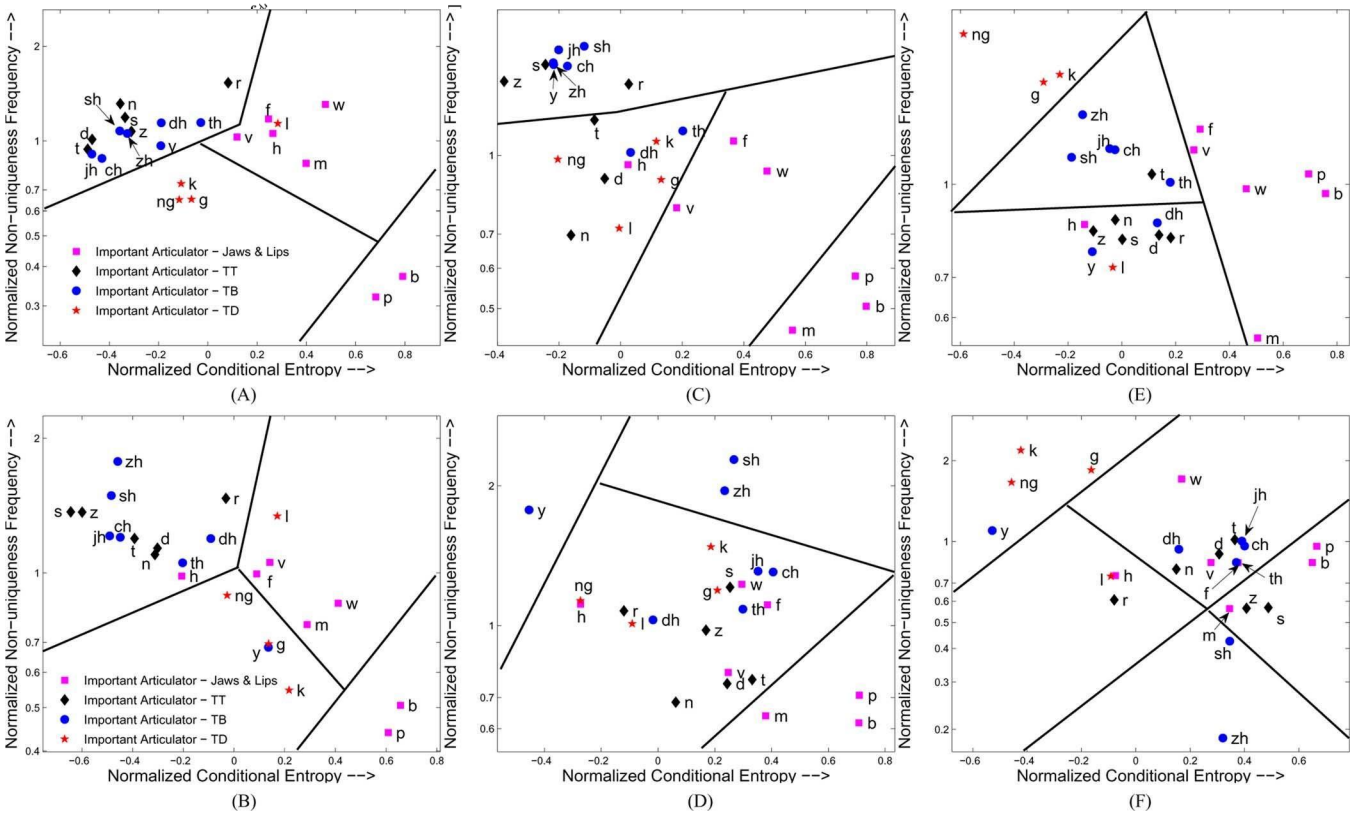

Fig. 4. Comparison of the and normalized for articulators Tongue Tip (TT), Tongue Back (TB) and Tongue Dorsum (TD), depending on the phonemes. The phonemes are clustered using K-means clustering, the straight lines representing boundaries of these clusters. The symbols used depend on the articulator that is considered important (based on the SVM hyperplane orientation) when uttering the particular phoneme. (A) Tongue Tip (TT), male speaker (B) Tongue Tip (TT), female speaker (C) Tongue Back (TB), male speaker (D) Tongue Back (TB), female speaker (E) Tongue Dorsum (TD), male speaker (F) Tongue Dorsum (TD), female speaker.

Table 4. The list of phonemes used in this study

ASCII Symbol	IPA Repr.	ASCII Symbol	IPA Repr.	ASCII Symbol	IPA Repr.
Stop consonants					
p	p	t	t	k	k
b	b	d	d	g	g
Nasals, approximants and other sonorants					
m	m	n	n	ng	ŋ
l	l	r	ɹ	w	w
y	j				
Fricatives					
f	f	s	s	sh	ʃ
v	v	z	z	zh	ʒ
ch	tʃ	j	dʒ	th	θ
dh	ð	h	h		

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

C. Interpretation of the Results

It is clear that stop consonants are at the extreme ends of both the relative entropy (upper bound) and the relative non-uniqueness (in the multi-modal sense) frequency scales. For most consonants it becomes clear that it is relatively easier to predict the articulator position (low relative entropy upper bound) that is important for producing the phonemes, the frequency of multiple articulatory models is relatively higher. Our other experiments show that this trend is consistent for all vowels and diphthongs as well, although they tend to cluster around the average upper bound of entropy for all the articulators. The best claim that we can make based on these observations is that it is not possible to directly relate the presence of multiple modes in the articulatory sub-space to difficulty in prediction. What is relatively clearer is that, for consonants, the difficulty of predicting the articulatory positions from the acoustics is relatively lower when the articulator is important for its production.

While there are several examples of phonemes with multiple articulatory modes for the important articulators, there are almost no examples of multiple modes for articulators that are not important for the production of that phoneme. However, these unimportant articulators have a larger range of variations in the articulator positions, corresponding to non-uniqueness in the continuous sense. This means that co-articulatory effects on unimportant articulators cause an increase in the range of positions, but not in the number of frames with multiple non-unique articulator modes. On the other hand, there are few instances of multiple articulatory modes due to co-articulation of unimportant articulators. The exceptions are the coils on the tongue, for frames falling under phonemes '*f,v* and *w*' having a high $nf^{\aleph}$ for the TD coil. These may be the cases where co-articulation effects may be non-unique in the multi-modal sense.

The multi-modal sense of non-uniqueness, while being very specific, is all the more interesting in hind-sight, because it involves cases where the speaker uses the important articulators in a non-unique way to produce normal speech without altering the acoustics (cf. Figs. 3 and 4).

D. Validity of the Results

Based on statistical methods, such as in this paper, it is not easy to assess the validity of the results. Are the non-unique predictions an artifact of the data? How are they affected by the modeling choices? Is the parameterization affecting the results? What do multiple modes in the conditional probability mean? This section tries to address these factors one by one.

The first most crucial aspect is the attribution of evidence for non-uniqueness to drift, rotation or detachment of the coils. It is possible that the data that is being considered in fact has a drift and therefore the different peaks or clusters in the articulation may, in fact, correspond to the positions before and after the displacement of the coil in question.

In order to verify if this is the case, for every frame that was found to be non-unique, we calculated the 100 closest neighbors from among the data-set in the acoustic-articulatory space. From among the neighbors, we classified the vectors according to their proximity to one of the modes that were detected. For each set of 100 neighbors, we divided the time-series data (i.e., the data ordered according to the sequence in which it was collected) until we had at least one sample from each mode in each time division. For each such time-division in the data, we calculated the entropy of the probability of the neighbors being closer to one peak rather than to another. If the hypothesis that the estimated non-uniqueness was a result of local displacements in the coil were true, it would mean that within one such time-division, one would find more samples closer to one of the peaks rather than to the other, while for other time-divisions, the distribution would be different. Such one-sided, biased distributions would have a low entropy (we call this the time-spread entropy), close to zero. On the other hand, if for every time-division, the distribution between the different neighbors were even (entropy is close to one), then it means that the non-uniqueness was not due to local displacements but was in fact a global phenomenon in the database. We found that the average

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

time-spread entropy for all the articulator coils was between 0.77 and 0.98. The hypothesis that the non-uniqueness was due to local displacements could be rejected strongly for all the articulators (p-value>0.95).

The second question we tried to validate was whether the GMM model we adopted was good enough to parameterize the acoustic-articulatory space. Since the largest model we could adopt was 128 Gaussian components due to memory restrictions, this may not be a sufficiently accurate description of the data. Secondly, the EM algorithm may not actually have reached a global minimum in terms of the log-likelihood error. In order to verify how good the model was, we performed an inversion experiment. The model could predict the articulator positions from the acoustics with an RMS (Root Mean Square) error of 1.34 mm for the male speaker (msak) and 1.47 mm for the female speaker (fsew). This was without applying any smoothing or continuity constraints. However, when the error was estimated on only the frames that were predicted as unique, the RMS errors were 0.68 mm and 0.83 mm, values quite close to the average resolution of the acquired data (0.43 mm) and much lower than the theoretical limit [10], found to be around 0.91 mm for the same data-set. Of course, the training and testing data were exactly the same, in this experiment, which says how well the model could fit the data, rather than generalize it. It is obvious that the data would be described better if modeled with a greater number of Gaussian components. Given a better model, even if the error estimated on the data was definitely reduced, the number of non-unique frames would increase. The trend observed when we used GMM models of different numbers of components showed an increase in non-uniqueness from 1.1% ofthe frames to 8.8% ofthe frames as the number of GMM components was increased from 8 to 128. There was no significant difference between M=64 and M=128.

The GMM model, while being a good tool to model data of unknown probability distributions, has some short-comings. One of the most important of them is that they are real valued, unbounded and multi-dimensional. Thus, the conditional probability distribution, $\rho_{Y|X}(y|x)$, may exist even beyond the permissible limits of the vocal tract anatomy. For this reason, looking at the whole distribution or making calculations on the moments (for example, the variance estimate) of the distribution can be misleading. For this reason, the definition of non-uniqueness in the multi-modal sense, where one can be sure that none of the detected peaks fall outside the permissible range of articulations, is more accurate. On the other hand, the upper bound of the entropy gives an idea about the non-uniqueness in the continuous sense, but may be more inaccurate.

Another important question is whether the said articulatory modes in fact correspond to different phonemes, because of insufficient resolution of the acoustic parameterization. The answer to the question is that the acoustic parameterization is in fact insufficient for the purpose of categorization into respective phonemes. Based on the acoustic representations alone, using a K-nearest neighbor classifier with K=100 neighbors the phoneme classification task gives an average of 53.4% with a leave-one-out cross-validation. This is in fact quite close to human phoneme recognition on short words [24]. While the confusion over phonemic identity based on the acoustics is a known artifact, it should not be the case that the two modes consistently correspond to different phonemes. Thus we needed to test the hypothesis that the nearest neighbors of the two peaks in the distribution consistently belong to different phonemes. In order to test this hypothesis we assessed the phonemic labels of the top 10 closest articulatory points (neighbors) in the data to the peaks in the conditional distribution. We observed that for 77 to 94% of the non-unique frames, at least one neighbor was assigned to the same class label as the current frame for each of the multiple peaks. This of course was subject to the errors in the HMM based phonemic alignment. This showed that, while confusions did occur in terms of the phonemic identity, the detected multi-modality was not entirely due to insufficient resolution of the acoustic parametrization. In fact, for 21 to 32% of the frames, the first three closest neighbors to each of the modes in the conditional articulatory distribution correspond to the same phoneme as that of the current frame. Thus even though the acoustic resolution is insufficient for classification, the non-uniqueness results themselves are not due to this artifact.

The final question we tried to answer was to observe the effect of the number of acoustic frames we used for parameterization. In the current tests, we used an acoustic parameterization of 11 consecutive acoustic frames for

Manuscript accepted for publication in published in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

125 ms of speech. This is expected to reduce the estimated non-uniqueness in the data, because it is easier to obtain some context from the previous and following acoustic data. However, one must note that although the context is obtained from the acoustics, no context is obtained from the previous prediction of the articulations, and thus the augmentation of the acoustic frame towards a larger context does not correspond to applying continuity constraints to the prediction. However, the empirical results presented will definitely depend on the context taken. When we calculated the percentage of non-unique frames using different acoustic durations we found the maximum effect was for stop consonants, followed by fricatives. The percentage of non-unique frames did not vary as much for other sonorant consonants. For example, in the TT coil, when the acoustic vector duration was reduced from 125 ms to 45 ms, the percentage of non-unique frames from among those corresponding to stop consonants increased from 24.37% to 35.95% and from 26.4% to 34.4% for fricatives. The corresponding change was from 20.8% to 26.1% of the frames corresponding to other consonant sonorants in the database for the TT coil.

V. Conclusions and future work

This article presents statistical tools to study non-uniqueness in the acoustic-to-articulatory mapping, using simultaneously recorded acoustic-articulatory data. The main contribution is the ability to analyze the non-uniqueness in two directions while constraining the acoustic distribution to be constant. The first direction is based on finding the number of modes in the articulatory probability density function conditioned on the acoustics. A multi-modal conditional distribution points to non-uniqueness that is less dependent on measurement or modeling errors. The second method is to estimate an upper bound to the conditional entropy, which gives a general sense of the range of variation in the articulation, given the acoustics. While this also gives a good account of non-uniqueness, it is difficult to distinguish it from the errors in modeling and measurement.

The article also describes the relationship between nonuniqueness in the multi-modal sense and predictability of the articulation from the acoustics based on the MOCHA-TIMIT database. It is generally not true that higher non-uniqueness means higher upper bounds for the entropy and thus lower predictability. It was also found that non-uniqueness in the multi-modal sense is not necessarily higher for the unimportant articulators. In fact, for most phonemes, even though the entropy (upper bound) of prediction was relatively higher for the unimportant articulators, the occurrence of non-uniqueness in the multi-modal sense was often relatively infrequent. On the contrary, the non-uniqueness in the multi-modal sense was often relatively more frequent in the important articulators for the phonemes, even though they were more easily predicted (with low entropy upper bounds). The results of this article can not only be used to improve acoustic-to-articulatory inversion by explicit modeling of nonuniqueness, it is also useful in understanding and applying to fields like infant language acquisition, second language learning by adults, intra-oral articulation and speaker adaptation.

The reason why a speaker tends to utilize more than one configuration of the important articulators is not very straightforwardly explained. This may be due to co-articulatory or prosodic constraints in the speech material. An important future work is to validate whether the non-uniqueness in the discrete sense is compensatory in nature, maintaining area-functions, or has different area functions for all sounds that can be produced non-uniquely. This can be done by building an articulatory model and using the predicted non-unique positions to synthesize the acoustics. One main problem that impedes such a study is to find a unique speaker specific transform from articulator flesh points to vocal tract area functions. It has been shown that predicting the entire vocal tract based on 3 or 4 discrete flesh points is possible [25]. The next step is to verify whether the acoustics produced by the synthesizers using the two predicted non-unique configurations are perceptually distinguishable or not. If they are not distinguishable, one can then verify the exact means of effecting an instance of non-uniqueness.

This study is also important when trying to interpret the Motor Theory of speech perception [26] and the Direct realist theory of speech perception [27]. These theories assert that all speech is perceived by mapping the acoustics to an articulatory configuration corresponding to each distinct phoneme. We have shown that speakers are able to produce non-unique configurations to produce similar acoustic sounds. Although the final vocal tract configuration

Manuscript accepted for publication in
IEEE Transactions on Audio, Speech, and Language Processing, 20(10), 2672 –2682
DOI: 10.1109/TASL.2012.2210876

may or may not be constant while producing sounds using these non-unique configurations, the individual articulators (which are controlled by different muscles) seem to be able to produce almost the same acoustics from different positions. Another theory, the Quantal Theory of speech perception [28] asserts that small changes in the articulation cause sudden switching in the acoustic features and vice-versa. This study while illustrating such behavior also shows that small acoustic changes may also be produced by discrete multi-modal articulations. Thus, trying to interpret these theories of speech perception in the light of this study would be interesting future work.

Appendix

Merging Gaussian Probability Distributions: The merging of two Gaussian distributions in a GMM is based on the method suggested in [16]. Given a GMM with *M* Gaussians with parameters, $\{c_m, \mu_m, \Sigma_m\}$, merge the Gaussian components *i*: $1 \le i \le M$ and *j*: $1 \le j \le M$, if $i \neq j$ and if *i* and *j* minimize

$$B(i,j) = \frac{1}{2}[(c_i + c_j)log|\widehat{\Sigma_{ij}}| - c_i log|\Sigma_i| - c_j log|\Sigma_j| \quad (6)$$

with

$$\widehat{c_{ij}} = c_i + c_j \quad (7)$$

$$\widehat{\mu_{ij}} = \frac{c_i\mu_i + c_j\mu_j}{\widehat{c_{ij}}} \quad (8)$$

$$\Sigma_{ij} = \frac{c_i\Sigma_i + c_j\Sigma_j}{\widehat{c_{ij}}} + \frac{c_i c_j(\mu_i - \mu_j)(\mu_i - \mu_j)^T}{\widehat{c_{ij}}^2} \quad (9)$$

being the parameters of the merged Gaussian.

Algorithm for Finding the Upper Bound to the Entropy: To locate the modes of the conditional distribution for one acoustic frame *t*, for each articulator *a*, $\forall t: 1 \le t \le T$ (*T* is total number of frames in all utterances) and $\forall a: 1 \le a \le A$ (*A* is the number of articulators):

1) Estimate the conditional probability distribution of the acoustic to articulatory mapping, $\rho_{Y|X}(y^a|x_t: M)$, using an arbitrarily high number of components *(M)*.
2) Calculate the upper bound of the entropy [15] of $\rho_{Y|X}(y^a|x_t: M)$ modeled by $\lambda_M(c_{m,a,t}^{Y|X}, \mu_{m,a,t}^{Y|X}, \Sigma_{m,a}^{Y|X}: 1 \le m \le M)$, using the equation

$$\Xi_u(y^a|x_t;\ M) = \sum_{m=1}^{M} c_m^{Y|X}\left(\frac{1}{2}log((2\pi e)^d|\Sigma_m^{Y|X}|) - logM\ log c_m^{Y|X}\right) \quad (10)$$

3) Keep decreasing *M* by merging the Gaussian components as described in Section A until *M=1* or $\Xi_u(y^a|x_t; M+1) < \Xi_u(y^a|x_t; M)$. For each M, find new parameters for λ_M, namely, $[\widehat{c}_{m,a,t}^{Y|X}, \widehat{\mu}_{m,a,t}^{Y|X}, \widehat{\Sigma}_{m,a}^{Y|X}: 1 \le m \le M]$.
4) Calculate the upper bound for the entropy, $\Xi_u(y^a|x_t; M)$, from (10) at each step.
5) If $\Xi_u(y^a|x_t; M+1) < \Xi_u(y^a|x_t; M)$ for any *M*, the final number of modes is $\aleph_{a,t} = M + 1$ and the final entropy, $\Xi_u(y^a|x_t) \le \Xi_u(y^a|x_t; M+1)$.
6) If entropy keeps decreasing with *M* until *M=1*, then the final number of modes, $\aleph_{a,t} = 1$ and $\Xi_u(y^a|x_t; 1)$ is the final upper bound for the entropy.

By this method, we not only find the number of modes in the distribution but also estimate a tighter upper bound to the conditional entropy (cf. Equation (10), which is shown to be quite close to the true entropy of the distribution [15]. The estimate of non-uniqueness in the discrete sense is provided by the number of modes, $\aleph_{a,t}$. For the continuous case, we base our measure of non-uniqueness on the entropy upper bound estimate. An upper bound to the conditional entropy also gives us a lower bound to the predictability of the articulators given the acoustics. The estimated entropy (upper bound) is directly proportional to the log of the variance for the uni-modal case. For the multi-modal case ($\aleph_{a,t}>1$), we are able to calculate a tighter upper bound to the entropy, thereby obtaining a better description of predictability (and non-uniqueness) than by approximating it by a single normal distribution (which would be the case if we only considered the variance of the articulatory distribution). Even though the method is not optimum in terms of merging the different Gaussian distributions (since it is a greedy search), this is sufficient for estimating the reduction in the entropy bound. A more optimal scheme of merging would affect the final entropy estimate, but not the final number of modes.

References

[1] P. Mermelstein, “Determination of the Vocal-Tract shape from measured formant frequencies,” *J. Acoust. Soc. Amer.*, vol. 41, no. 5, pp. 1283–1294, 1967.

[2] M. R. Schroeder, “Determination of the geometry of the human vocal tract by acoustic measurements,” *J. Acoust. Soc. Amer.*, vol. 41, no. 4B, pp. 1002–1010, 1967.

[3] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” *J. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1535–1555, 1978.

[4] B. Lindblom, J. Lubker, and T. Gay, “Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation,” *J. Phonetics*, vol. 7, pp. 147–161, 1979.

[5] M. Stone and E. Vatikiotis-Bateson, “Trade-offs in tongue, jaw, and palate contributions to speech production,” *J. Phonetics*, vol. 23, no. 1–2, pp. 81–100, 1995.

[6] C. Kroos, A. Geumann, and P. Hoole, “Tongue-jaw trade-offs and naturally occurring perturbation,” *J. Acoust. Soc. Amer.*, vol. 105, pp. 1355–1355, 1999.

[7] F. H. Guenther, C. Y. Espy-Wilson, S. E. Boyce, M. L. Matthies, M. Zandipour, and J. S. Perkell, “Articulatory trade-offs reduce acoustic variability in American English productions,” *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2854–2865, 1999.

[8] A. Nieto-Castanon, F. H. Guenther, J. S. Perkell, and H. D. Curtin, “A modeling investigation of articulatory variability and acoustic stability during American English production,” *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3196–3212, 2005.

[9] C. Qin and M. Á. Carreira-Perpiñán, “An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping,” in *Proc. Interspeech*, 2007, pp. 74–77.

[10] D. Neiberg, G. Ananthakrishnan, and O. Engwall, “The acoustic to articulation mapping: Non-linear or non-unique?,” in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1485–1488.

[11] C. Qin and M. Carreira-Perpiñán, “The geometry of the articulatory region that produces a speech sound,” in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, 2009, pp. 1742–1746.

[12] G. Ananthakrishnan, D. Neiberg, and O. Engwall, “In search of nonuniqueness in the acoustic-to-articulatory mapping,” in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 2799–2802.

[13] S. Park and A. Bera, “Maximum entropy autoregressive conditional heteroskedasticity model,” *J. Econometrics*, vol. 150, no. 2, pp. 219–230, 2009.

[14] T. Toda, A. W. Black, and K. Tokuda, “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model,” *Speech Commun.*, vol. 50, pp. 215–227, 2008.

[15] M. F. Huber, T. Bailey, H. Durrant-Whytem, and U. D. Hanebeck, “On entropy approximation for Gaussian mixture random vectors,” in *Proc. Multisensor Fusion and Integration for Intell. Syst.*, Seoul, South Korea, 2008, pp. 181–188.

[16] A. Runnalls, "Kullback-Leibler approach to Gaussian mixture reduction," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 3, pp. 989–999, 2007.

[17] E. S. Youn, "Feature selection in support vector machines," M.S. thesis, The Graduate School of the Univ. of Florida, Gainesville, 2002.

[18] A. Wrench, "A Multi-channel/multi-speaker articulatory database for continuous speech recognition research," Queen Margaret Univ. College, Tech. Rep..

[19] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, 2008.

[20] P. Hoole, "Issues in the acquisition, processing, reduction and parameterization of articulographic data," in *Proc. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 1996, vol. 34, pp. 158–173.

[21] K. Richmond, "Estimating articulatory parameters from the speech signal," Ph.D. dissertation, The Center for Speech Technol. Research, Edinburgh, U.K., 2002.

[22] J. A. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Int. Comput. Sci. Inst., 1998, Tech. Rep..

[23] C. Mooshammer, A. Geumann, P. Hoole, P. Alfonso, P. van Lieshout, and S. Fucks, "Coordination of lingual and mandibular gestures for different manners of articulation," in *Proc. ICPhS*, 2003, pp. 81–84.

[24] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier, "Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1273–1276.

[25] C. Qin, M. Á. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 2306–2309.

[26] A. Liberman, F. Cooper, D. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychol. Rev.*, vol. 74, no. 6, pp. 431–461, 1967.

[27] R. Diehl, A. Lotto, and L. Holt, "Speech perception," *Ann. Rev. Psychol.*, vol. 55, pp. 149–179, 2004.

[28] K. N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds. New York: McGraw-Hill, 1972, pp. 51–66.