

Cross-modal Clustering in the Acoustic-Articulatory Space

(As a requirement for the GSLT course on Statistical Methods)

- G. Ananthkrishnan

Abstract

This paper explores cross-modal clustering in the acoustic-articulatory space. A method to improve clustering using information from more than one modality is presented. Formants and the Electromagnetic Articulography measurements are used to study corresponding clusters formed in the two modalities. A measure for estimating the uncertainty in correspondences between one cluster in the acoustic space and several clusters in the articulatory space is suggested.

Introduction

Trying to estimate the articulatory measurements from acoustic data has been of special interest in the speech community for long time and is known as acoustic-to-articulatory inversion. Though this mapping between the two modalities is expected to be a one-to-one mapping, early research presented some interesting evidence showing non-uniqueness, in this mapping. 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 (Gay *et al.*, 1981). Mermelstein (1967) and Schroeder (1967) have shown, through analytical articulatory models, that the inversion is unique to a class of area functions rather than a unique configuration of the vocal tract.

With the advent of measuring techniques like Electromagnetic Articulography (EMA) and X-Ray Microbeam, it was possible to collect simultaneous measurements of acoustics and articulation during continuous speech. Several attempts have been made by researchers to perform acoustic-to-articulatory inversion by applying machine learning techniques to the acoustic-articulatory data (Yehia *et al.*, 1998 and Kjellström and Engwall, 2009). The statistical methods applied to the problem of mapping brought a new dimension to the concept of non-uniqueness in the mapping. 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. It is almost impossible to show this using real recorded data, unless more than one articulatory configuration produces exactly the same acoustic parameters. However, not finding such instances does not imply that non-uniqueness does not exist.

Qin and Carreira-Perpinán (2007) 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. Evidence of non-uniqueness in certain acoustic clusters for phonemes like /j/, /l/ and /w/ was presented. The study by Qin quantized the acoustic space using the perceptual Itakura distance on LPC features. The articulatory space was clustered using a nonparametric Gaussian density kernel with a fixed variance. The problem with such a definition of non-uniqueness is that one does not know what is the optimal method and level of quantization for clustering the acoustic and articulatory spaces. A later study by Neiberg *et al.* (2008) argued 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. Using an approach based on finding the Bhattacharya distance between the distributions of the inverse mapping, they found that phonemes like /p/, /t/, /k/, /s/ and /z/ are highly non-unique.

In this study, we wish to observe how clusters in the acoustic space map onto the articulatory space. For every cluster in the acoustic space, we intend to find the uncertainty in finding a corresponding articulatory cluster. It must be noted that this uncertainty is not necessarily the non-uniqueness in the acoustic-to-articulatory mapping. However, finding this uncertainty would give an intuitive understanding about the difficulties in the mapping for different phonemes.

Clustering the acoustic and articulatory spaces separately, as was done in previous studies by Qin and Carreira-Perpinán (2007) as well as Neiberg *et al.* (2008), leads to hard boundaries in the clusters. The cluster labels for the instances near these boundaries may be estimated incorrectly, which may cause an over estimation of the uncertainty. This situation is explained by Fig. 1 using synthetic data where we can see both the distributions of the synthetic data as well as the Maximum A-posteriori Probability (MAP) Estimates for the clusters. We can see that, because of the incorrect clustering, it seems as if data belonging to one cluster in mode A belongs to more than one cluster in mode B.

In order to mitigate this problem, we have suggested a method of cross-modal clustering where both the available modalities are made use of by allowing soft boundaries for the clusters in each modality. Cross-modal clustering has been dealt with in detail under several contexts of combining multi-modal data. Coen (2005) proposed a self supervised method where he used acoustic and visual features to learn perceptual structures based on temporal correlations between the two modalities. He used the concept of slices, which are topological manifolds encoding dynamic states. Similarly, Belolli *et al.*(2007) proposed a clustering algorithm using Support Vector Machines (SVMs) for clustering inter-related text datasets.

The method proposed in this paper does not make use of correlations, but mainly uses co-clustering properties between the two modalities in order to perform the cross-modal clustering. Thus, even non-linear dependencies (uncorrelated) may also be modeled using this simple method. However the method proposed, does not take into account dependencies which are time related, which the other methods specified above, take into account.

Theory

We assume that the data is a Gaussian Mixture Model (GMM). The acoustic space $Y = \{y_1, y_2 \dots y_N\}$ with ' N ' data points is modelled using ' I ' Gaussians, $\{\lambda_1, \lambda_2, \dots \lambda_I\}$ and the articulatory space $X = \{x_1, x_2 \dots x_N\}$ is modelled using ' K ' Gaussians, $\{\gamma_1, \gamma_2, \dots \gamma_K\}$. ' I ' and ' K ' are obtained by minimizing the Bayesian Information Criterion (BIC). If we know which articulatory Gaussian a particular data point belongs to, say, γ_k The correct acoustic Gaussian ' λ_n ' for the the ' n^{th} ' data point having acoustic features ' y_n ' and articulatory features ' x_n ' is given by the maximum cross-modal a-posteriori probability

$$\begin{aligned} \lambda_n &= \arg \max_{1 \leq i \leq I} P(\lambda_i | x_n, y_n, \gamma_k) \\ &= \arg \max_{1 \leq i \leq I} p(x_n, y_n | \lambda_i, \gamma_k) * P(\lambda_i | \gamma_k) * P(\gamma_k) \end{aligned} \quad (1)$$

The knowledge about the articulatory cluster can then be used to improve the estimate of the correct acoustic cluster and vice versa as shown below

$$\gamma_n = \arg \max_{1 \leq k \leq K} p(x_n, y_n | \lambda_i, \gamma_k) * P(\gamma_k | \lambda_i) * P(\lambda_i) \quad (2)$$

Where $P(\lambda|\gamma)$ is the cross-modal prior and the $p(x,y|\lambda,\gamma)$ is the joint cross-modal distribution. If the first estimates of the correct clusters are MAP, then the estimates of the correct clusters of the speech segments are improved recursively. Finally a soft clustering is obtained which maximizes the cross-modal a-posteriori probability in both the modes. We call this method Maximum Cross-Modal A-Posteriori Probability (MCMAP). Proving the convergence of the algorithm is beyond the scope of this paper. However, the algorithm converged for all the experiments within 50 iterations. In Fig. 2, we can see that the estimate of the correct cluster is slightly better than a simple a posteriori probability in Fig. 1.

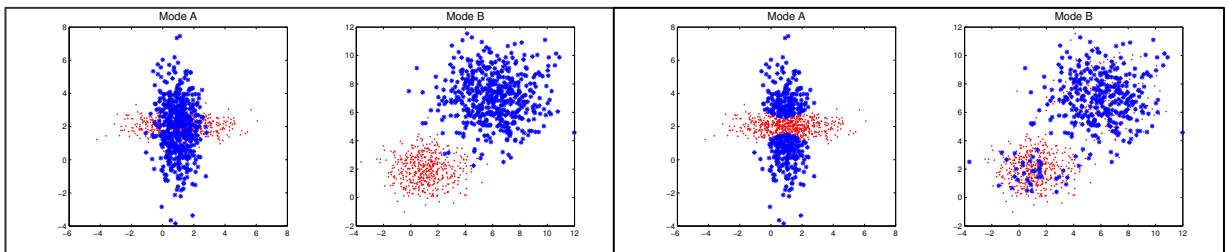


Figure 1. The figures above show a synthesized example of data in two modalities. The figures on the left show how MAP hard clustering may bring about an effect of uncertainty in the correspondence between clusters in the two modalities.

The proposed method is interesting because it is unsupervised and makes use of knowledge from both the modalities efficiently. Clustering the joint space of the two modalities is another approach although it assumes that the number of clusters in each of the modalities is exactly the same. So the proposed method

gives an added advantage since each of the modalities could have a different number of clusters. However, comparing the two methods is beyond the scope of this paper presently.

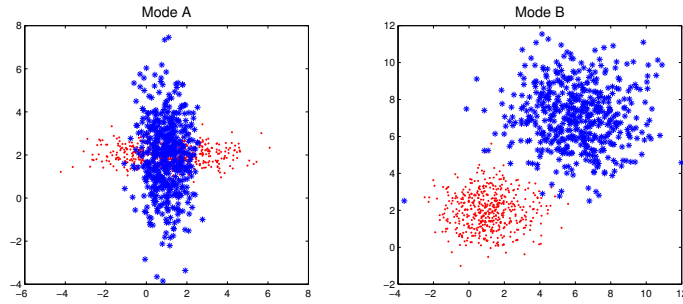


Figure 2. The figures shows an improved performance and soft boundaries for the synthetic data using cross-modal clustering, here the effect of uncertainty in correspondences is less.

The uncertainty of the clustering for acoustic to articulatory inversion in a particular phoneme, can be estimated by the cross-modal prior, i.e. $P(\gamma|\lambda)$. We propose that the measure of uncertainty in the cross-modal cluster correspondence, 'U', is given by the Entropy of $P(\gamma|\lambda)$. This seems to be an intuitive method of ascertaining uncertainty, because a lower entropy means that given a cluster in one of the modalities, there is close to one-to-one correspondence between another the cluster in the second modality.

$$U_{\lambda_i} = -\sum_{k=1}^K P(\gamma_k | \lambda_i) \log_K(P(\gamma_k | \lambda_i)) \quad (3)$$

$$U = \sum_{i=1}^I U_{\lambda_i} * P(\lambda_i)$$

Log to the base 'K' is taken in order to normalize the effect of different number of articulatory clusters. The entropy is a good measure of the uncertainty in prediction, and thus forms an intuitive measure for our purpose. It is always between 0 and 1 and so comparisons between different cross-modal clusterings is easy. 1 indicates very high uncertainty while 0 indicates one-to-one mapping between corresponding clusters in the two modalities.

Experiments and Results

The MOCHA-TIMIT database (Wrench, 2001) was used to perform the experiments. The data consists of simultaneous measurements of acoustic and articulatory data for a female speaker. The articulatory data consisted of 14 channels, which included the X and Y-axis positions of EMA coils on 7 articulators, the Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (V). Only vowels were considered for this study and the acoustic space was represented by the first 5 formants, obtained from 25 ms acoustic windows shifted by 10 ms. The articulatory data was low-pass filtered and down-sampled in order to correspond with acoustic data rate. The uncertainty (U) in clustering was estimated using Equation 3 for the British vowels, namely /ʊ, æ, e, ɒ, ɑ:, u:, ɜ:r, ɔ:, ʌ, ɪ, ə, ə/. The articulatory data was first clustered for all the articulatory channels and then was clustered individually for each of the 7 articulators.

Fig. 3 shows the clusters in both the acoustic and articulatory space for the vowel /e/. We can see that data points corresponding to one cluster in the acoustic space (F1-F2 formant space) correspond to more than one cluster in the articulatory space. The ellipses, which correspond to initial clusters are replaced by different clustering labels estimated by the MCMAP algorithm. So though the acoustic features had more than one cluster in the first estimate, after cross-modal clustering, all the instances are assigned to a single cluster.

Fig. 4 shows the correspondences between acoustic clusters and the LJ for the vowel /ə/. We can see that the uncertainty is less for some of the clusters, while it is higher for some others. Fig. 5 shows the comparative measures of overall the uncertainty (over all the articulators), of the articulatory clusters corresponding to each one of the acoustic clusters for the different vowels tested. Fig 6. shows the correspondence uncertainty of individual articulators.

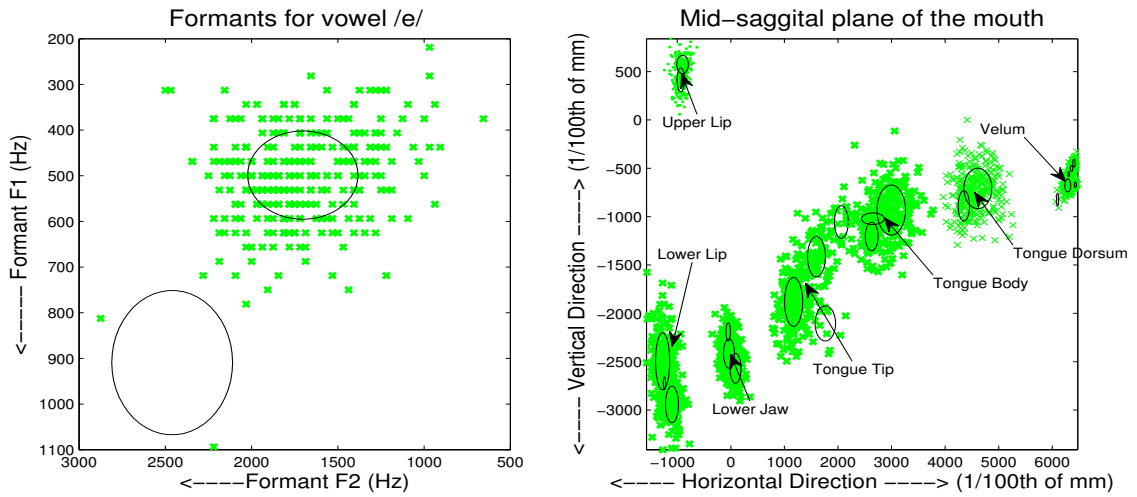


Figure 3. The figure on the left side shows the formant space for the vowel /e/ and the corresponding articulatory positions on the right hand. The ellipses show the locations of the estimated gaussians. One can observe that a single cluster in the acoustic space may be spread over several clusters in the articulatory space.

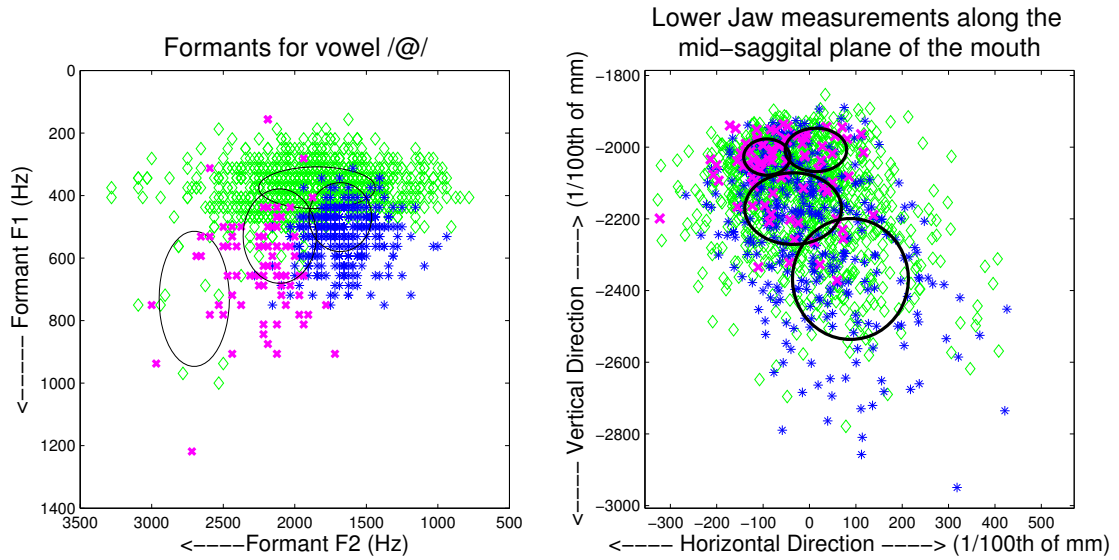


Figure 4. The figure on the left side shows the formants for the phoneme /ə/. The figure on the right indicate the measurements of the lower jaw for the corresponding instances. The ellipses indicate the locations of the estimated gaussians. We can see that some clusters is located within a small region in the articulatory space, while a few other clusters are spread all over.

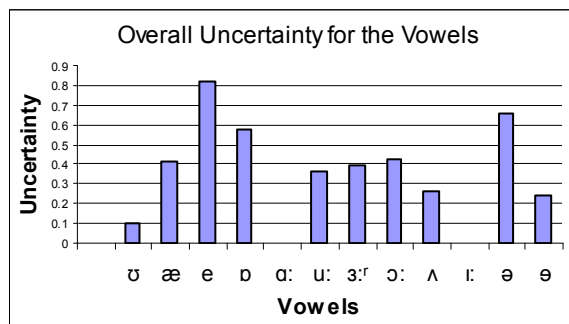


Figure 5. The figure shows the overall uncertainty (for the whole articulatory configuration) for the British vowels.

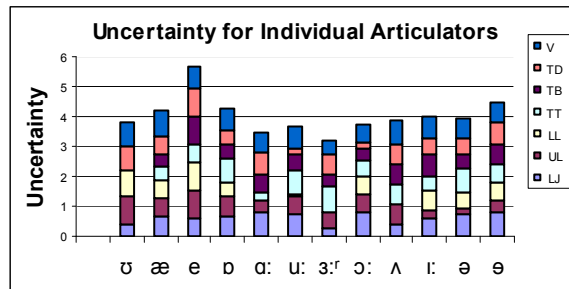


Figure 6. The figure shows the uncertainty for individual articulators for the British vowels.

Discussion

From Fig. 5 it is clear that the shorter vowels seem to have more uncertainty than longer vowels which is intuitive. The higher uncertainty is seen for the short vowels /e/ and /ə/, while there is almost no uncertainty for the long vowels /ɑ:/ and /ɪ:/. The overall uncertainty for the entire configuration is usually around the lowest uncertainty for a single articulator. This is intuitive, and shows that even though certain articulator correspondences are uncertain, the correspondences are more certain for the overall configuration. When the uncertainty for individual articulators is observed, then it is apparent that the velum has a high uncertainty of more than 0.6 for all the vowels. This is due to the fact that nasalization is not observable in the formants very easily. So even though different clusters are formed in the articulatory space, they are seen in the same cluster in the acoustic space. The uncertainty is much less in the lower lip correspondence for the long vowels /ɑ:/, /u:/ and /ɜ:/ while it is high for /ʊ/ and /e/. The TD shows lower uncertainty for the back vowels /u:/ and /ɜ:/. The uncertainty for TD is higher for the front vowels like /e/ and /ə/. The uncertainty for the tongue tip is lower for the vowels like /ʊ/ and /ɑ:/ while it is higher for /ɜ:/ and /ʌ/. These results are intuitive, and show that it is easier to find correspondences between acoustic and articulatory clusters for some vowels, while it is more difficult for others.

Conclusion and Future Work

The algorithm proposed, helps in improving the clustering ability using information from multiple modalities. A measure for finding out uncertainty in correspondences between acoustic and articulatory clusters has been suggested and empirical results on certain British vowels have been presented. The results presented are intuitive and show difficulties in making predictions about the articulation from acoustics for certain sounds. It follows that certain changes in the articulatory configurations cause variation in the formants, while certain articulatory changes do not change the formants.

It is apparent that the empirical results presented depend on the type of clustering and initialization of the algorithm. This relation must be explored further and must be expressed for other methods of clustering such as vector quantization hierarchical clustering and k-means clustering. Future work must also be done on extending this paradigm to include other classes of phonemes as well as different languages and subjects. It would be interesting to see if these empirical results can be generalized or are special to certain subjects and languages and accents.

The proposed soft clustering approach is a general method and can be applied in several contexts. It would be specifically interesting to see whether it is better than modeling the joint spaces of two modalities rather than the two separately.

References

- Bolelli L., Ertekin S., Zhou D. and Giles C. L. (2007) K-SVMMeans: A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets. International Conference on Web Intelligence, 198–204.
- Coen M. H. (2005) Cross-Modal Clustering. Proceedings of the Twentieth National Conference on Artificial Intelligence, 932-937.
- Gay, T., Lindblom B. and Lubker, J. (1981) Production of bite-block vowels: acoustic equivalence by selective compensation. J. Acoust. Soc. Am. 69, 802-810, 1981.
- Kjellström, H. and Engwall, O. (2009) Audiovisual-to-articulatory inversion. Speech Communication 51(3), 195-209.

Mermelstein, P., (1967) Determination of the Vocal-Tract Shape from Measured Formant Frequencies, *J. Acoust. Soc. Am.* 41, 1283-1294.

Neiberg, D., Ananthakrishnan, G. and Engwall, O. (2008) The Acoustic to Articulation Mapping: Non-linear or Non-unique? *Proceedings of Interspeech*, 1485-1488.

Qin, C. and Carreira-Perpinán, M. Á. (2007) An Empirical Investigation of the Nonuniqueness in the Acoustic-to-Articulatory Mapping. *Proceedings of Interspeech*, 74–77.

Schroeder, M. R. (1967) Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am* 41(2), 1002–1010.

Wrench, A. (1999) The MOCHA-TIMIT articulatory database. Queen Margaret University College, Tech. Rep, 1999. Online: <http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html>.

Yehia, H., Rubin, P. and Vatikiotis-Bateson. (1998) Quantitative association of vocal-tract and facial behavior. *Speech Communication* 26(1-2), 23-43.