Investigating Non-Uniqueness in the Acoustic-Articulatory Inversion Mapping

THE UNIVERSITY *of* EDINBURGH

B005324

Master of Science

Speech & Language Processing

University of Edinburgh

2011

Table of Contents

B005324

Declaration

I have read and understood The University of Edinburgh guidelines on Plagiarism and declare that this written dissertation is all my own work except when I indicate otherwise by proper use of quotes and references.

Terence Simms

B005324

Copyright

Acknowledgements

I would like to thank Dr. Korin Richmond & Prof. Steve Renals for all their excellent help and guidance throughout the completion of this dissertation, for all their valuable discussions and for their swift and courteous responses to my many emails.

I would also like to thank Chao Qin & Miguel Á. Carreira-Perpiñán for taking the time to help me in whatever way they could.

Finally, I would like to thank my parents for their unconditional support, and I'd especially like to thank Deborah, without whom I would not have survived this past year.

Abstract

The task of inferring articulatory configurations from a given acoustic signal is a problem for which a reliable and accurate solution has been lacking for a number of decades. The changing shape of the vocal-tract is responsible for altering the parameters of sound. Each different configuration of articulators will regularly lead to a single distinct sound being produced (a *unique* mapping from the articulator to the acoustics). Therefore, it should be possible to take an acoustic signal and invert the process, giving the exact vocal-tract shape for a given sound. This would have wide-reaching applications in the field of speech and language technology, such as in improving facial animation and speech recognition systems. Using vocal-tract information inferred from the acoustic signal can facilitate a richer understanding of the actual constraints in articulator movement.

However, research concerned with the inversion mapping has revealed that there is often a multi-valued mapping from the acoustic domain to the articulatory domain. Work in identifying and resolving this *non-uniqueness* thus far has been somewhat successful, with Mixture-Density Networks (MDN) and articulator trajectory systems presenting probabilistic methods of finding the most likely articulatory configuration for a given signal. Using an subset of an EMA corpus, along with a combination of an instantaneous inversion mapping and a non-parametric clustering algorithm, I aim to quantify the extent to which acoustically similar vectors to a given phone can exhibit qualitatively different vocal-tract shapes. Categorical identification of acoustically similar sounds that can have shown a multi-valued mapping in the articulatory domain, as well as identifying which articulators this occurs for, could be key to resolving issues in the reliability and quality of the inversion mapping.

Chapter 1. Introduction

Acoustic-Articulatory Inversion Mapping

The task of inferring articulatory configurations from a given acoustic signal is a problem for which a reliable and accurate solution has been lacking for a number of decades. The changing shape of the vocal-tract is responsible for altering the parameters of sound. Each different configuration of articulators will regularly lead to a single distinct sound being produced (a *unique* mapping from the articulator to the acoustics). Therefore, it should be possible to take an acoustic signal and invert the process, giving the exact vocal-tract shape for a given sound. This would have wide-reaching applications in the field of speech and language technology.

One such application of this technology implements the inversion mapping to provide more realistic lip and mouth movements to computer-generated characters in video games and in computer generated animations. As this uses just the acoustic signal, this is more cost-effective and accurate method of providing facial animation than is in use currently, wherein an actor has reflective pellets attached only to the lip articulators, resulting in less realistic mouth animation. A successful inversion method could also aid the advance of many speech synthesis and recognition systems. Using vocal-tract information inferred from the acoustic signal can facilitate a richer understanding of the actual constraints in human articulator movement.

However, research concerned with the inversion mapping has revealed that there is often a multi-valued mapping from the acoustic domain to the articulatory domain. Work in identifying and resolving this *non-uniqueness* thus far has been somewhat successful, with Mixture-Density Networks (MDN) and articulator trajectory systems presenting probabilistic methods of finding the most likely articulatory configuration for a given signal. Though these are currently the most reliable and accurate inversion mapping systems, as discussed in chapter 2, evidence for non-uniqueness in the acoustic-articulatory mapping is still a

rigorously researched field in speech technology.

Overview

Categorical identification of acoustically similar sounds that can have shown a multi-valued mapping in the articulatory domain, as well as identifying which articulators this occurs for, could be key to resolving issues in the reliability and quality of the inversion mapping. The experiments undertaken in this dissertation project share similar methodological steps to previous studies that have investigated non-uniqueness in the acoustic-articulatory inversion mapping (Qin & Carreira-Perpiñán (2007), Neiberg, Ananthakrishnan & Engwall (2008)).

Using an subset of an EMA corpus, consisting of electromagnetic articulography coordinate data for 6 articulatory positions (3 sets of x & y coordinates on the midsagittal plane for the tongue, 2 for the lips and one for the lower-incisor) and concurrent labelled audio data from 1263 utterances (799,159 pairs of acoustic-articulatory vectors), I will be quantifying the extent to which acoustically similar vectors to a given phone can exhibit corresponding qualitatively different vocal-tract shapes. Furthermore, I will be investigating for which phones and for which articulators this non-unique inversion mapping occurs, and in what quantity.

To achieve this, I will be selecting a specified reference frame, taking the 1500 closest vectors in the acoustic data-set to that reference frame, and finding the corresponding articulatory coordinates for each and every articulator. As a distance measure, I will be using a *k*D-tree, which is an efficient technique of partitioning multidimensional data such as parametrised speech, and can subsequently be used to perform a *nearest neighbour* search using Euclidean distance. Finally, I will be using a non-parametric Gaussian-blurring mean-shift algorithm to efficiently cluster the data that exists in distinctly dense regions in articulatory space. This will be discussed further in my detailed methodology (chapter 3).

A broad empirical investigation such as this has not been performed on the new *mngu0* EMA corpus, and as such this study will hopefully provide both a further insight into the quality of

the *mngu0* corpus (after Richmond (2009) & (2011)), whilst also facilitating a comparison of the extent of acoustic-articulatory non-uniqueness in the corpus with previous & future studies (e.g. Qin & Carreira-Perpiñán (2007)).

Layout

This dissertation consists of three sections; *Literature Review & Background, Methodology* and *Conclusion & Further Work.*

Chapter 2. Literature Review

The Ill-Posed Problem

The acoustic-articulatory inversion mapping has long been sought after, but the nature of the task has been found to be *ill-posed (*Richmond 2002). An ill-posed problem refers to a problem for which multiple possible solutions can occur, and this has been found to be applicable to the inversion task as multiple articulatory configurations can map to a single acoustic vector. Evidence for this has come from a number of fields, and this acoustic-articulatory non-uniqueness is a problem for which multiple solutions have been proposed.

Bite-blocks

Experiments using bite-blocks in the 1970s provided some of the first qualitative evidence suggesting that non-uniqueness existed. Using bite-blocks of either 2.5mm or 22.5mm thickness, Lindblom, Lubker & Gay (1979) fixed the position of 6 speaker's jaws to see if they could produce four Swedish vowels (/i,u,o,a/). Then, in comparing the formants of the resulting acoustic signal with the formants of the speaker's vowels without any impediment, they discovered that the speakers utilised a degree of *articulatory compensation* in their impeded speech. This meant that the speakers varied their articulation of a vowel to match the formants of their regular unimpeded vowels. Although the jaw was fixed in an unnatural position, the sounds produced were perceptually well within the variation of their normal vowels.

Inversion

Tube-Models

A small proportion of research into the inversion mapping attempted to infer the size and shape of the vocal tract for a given sound. These analyses utilise the concept of the source-filter model of speech production.

The source of a sound during speech production can be categorised into two types; voiced speech, where regular vibration of the vocal folds occurs in response to pulmonic airflow from the lungs, and voiceless speech (eg. vowels), where irregular vibrations are caused by the constriction of the vocal-tract itself (eg. fricatives). The filter in speech production is the vocal-tract, where a change in its shape can alter the spectral envelope of the source sound. A model of speech production system treats the vocal-tract as a tube that is closed at one end. The tube resonates at certain frequencies, and these resonances, or harmonics, can be seen as peaks in the spectral envelope at integer multiples of the source frequency, called the *fundamental frequency*. The frequency levels of the first and second peaks, or *formants,* can often be used to differentiate different vowels. (Ladefoged 1996)

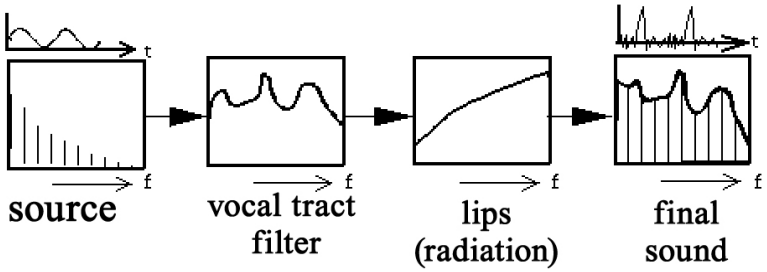


Figure 1. The Source-Filter model of Speech Production

Changing the length or area of the tube means that a different wavelength now fits the tube, resulting in a different sound. In speech production, different articulatory configurations can modify this speech signal. Wakita (1979) in particular used mathematical analysis of the acoustic signal from vowels to infer the area function of the vocal-tract. However, combining the complexity of multiple articulatory positions and the complex shape of the vocal tract with the lack of a measure for nasal sounds and unvoiced consonants meant that this particular analytical method was ill-suited to the inversion task.

Synthesis Models

Using synthesised speech as a means of studying the acoustic-articulatory inversion introduced trained models as a viable inversion method. The first articulatory synthesis models used methods that took samples of articulatory parameters from a model of the articulatory feature space and synthesised the corresponding acoustic signal based on these parameters. Databases, or *codebooks*, of these acoustic-articulatory pairs could then be made available for studying the range of articulatory configurations possible. Also, it allowed a lookup of articulatory parameters to be performed for a given sound.

A study by Atal et al (1978) investigated the inversion mapping using 4 parameters that would best describe the articulatory process; vocal-tract length, the size of the mouth opening, distance from the glottis and the area function at maximum constriction. From these parameters, over 30,000 different vocal-tract configurations were generated. These parameters were then used to generate acoustic counterparts using calculated formant and frequency. The acoustic-articulatory pairs were computer-sorted (due to the large size of the resulting database). A later study by Rahim, Kleijn, Schroeter & Goodyear (1991) used an articulatory model by Mermelstein (1973) to sample randomly from a set number of "reasonable" articulatory shapes and generate an acoustic-articulatory parameter pair database. This synthesised articulatory model was used as the basis for training a Multi-layer Perceptron(MLP).

An MLP is a feed-forward neural network that can be used to map input data onto output data. Each node in the network can be considered a separate processing unit that is activated by a specified function, such as sigmoidal or hyperbolic tangent functions, which respond to parametrised acoustic data. Multiple layers of nodes are connected together, with a *hidden* layer consisting of a given number of nodes. The number of hidden nodes to choose depends on the dimensionality and type of data. Too many nodes, and you risk over-fitting the data to random noise. Each connected node has a weight assigned to it. Using a supervised learning method called *backpropagation*, the error of the output units is calculated and compared with the target output using an error function. Connection weights are then iteratively altered after propagating through the network to see which connection weight contributes most to the

calculated error. (Bishop 1995, Richmond 2002).

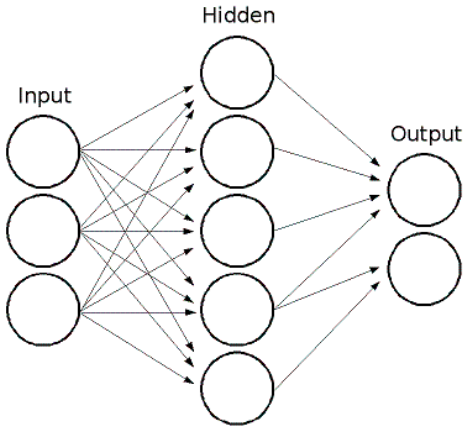


Figure 2. An MLP network consisting of 3 input nodes, 5 hidden nodes and 2 output nodes.

In Rahim et al. (1991), preprocessing the synthesised model of generated acoustic-articulatory parameter pairs required clustering in two stages. Initially, the acoustic data was clustered into 32 distinct regions in acoustic space, and then each cluster was segmented into a further 4 corresponding regions in articulatory space in an attempt to resolve non-uniqueness. For the acoustic and articulatory data, a *cepstral* distance measure (derived by a Fourier Transform of the log spectrum(Jurafsky & Martin 2009)) and a log-area distance measure were used, respectively. Each cluster was used in training a separate MLP, and optimum articulatory trajectories (the path the parameters take) were recovered using dynamic-programming.

The quality of such a system was ultimately dependent on the accuracy of the synthesised speech data in approximating real human speech. In this respect, this inversion model was an insufficient representation of real speech. The acoustic-articulatory pairs were generated by randomly sampling articulatory configuration parameters. This becomes a problem when assessing the validity of whether or not these articulatory configurations actually occur in real human speech. Each articulator has the ability to move to a certain extreme, such as the lips protruding far forward whilst the tongue tip is retroflexed as an extreme example, but they would hardly move to these extremes in real human speech. Furthermore, the importance placed on the uncommon sampled regions may affect the accuracy of the weights updates during backpropagation training. The number of articulatory configurations physiologically available is therefore not a good approximation of the frequency of these configurations in real speech.

Mimic algorithms

Later, combinations of human speech data and synthesised vector pairs that made the synthesised output iteratively match the real acoustic data. This *mimic* method is too computationally expensive to be used with a large dataset, but offers a fast and accurate method of estimating acoustic-articulatory parameters of synthesised data when a smaller dataset is used to train an MLP, with a study by Kobayashi, Yagya & Shirai (1991) showing a 10 fold increase in estimation speed over basic mimic model systems.

Human Articulatory Data Systems

While synthesis models of acoustic-articulatory speech parameters laid the groundwork for how the inversion mapping could be performed, the introduction of large databases of human articulatory data and concurrent acoustic data have enabled the possibility of learning trained acoustic-articulatory models of speech. These databases facilitate a more accurate inversion mapping system, as the true range of articulatory configurations can e analysed. Moreover, these trained inversion systems have the measured articulatory data itself for use in evaluating system performance and accuracy.

X-Ray microbeam cinematography

X-Rays have been used previously as a method of visualising the movements of articulators. However, recent advances in the technology have allowed more precise measurements of articulatory movement with a reduced exposure to X-Rays, which are widely recognised as being harmful under prolonged or repeated exposure. X-Ray microbeam cinematography is one such technique for providing human articulatory data and is widely used, with a number of studies in particular using the Wisconsin X-Ray microbeam cinematography Database (XRDB).

This particular technique uses a number of gold pellets on each of the speaker's articulators, and a narrow beam of X-rays and an X-Ray detector on either side of the speaker's head. Using the known positions, velocity and accelerations of the pellets, scans are taken of the estimated articulator postions using the microbeam and the pellet locations cast a shadow on the X-ray detector. (Westbury 1994)

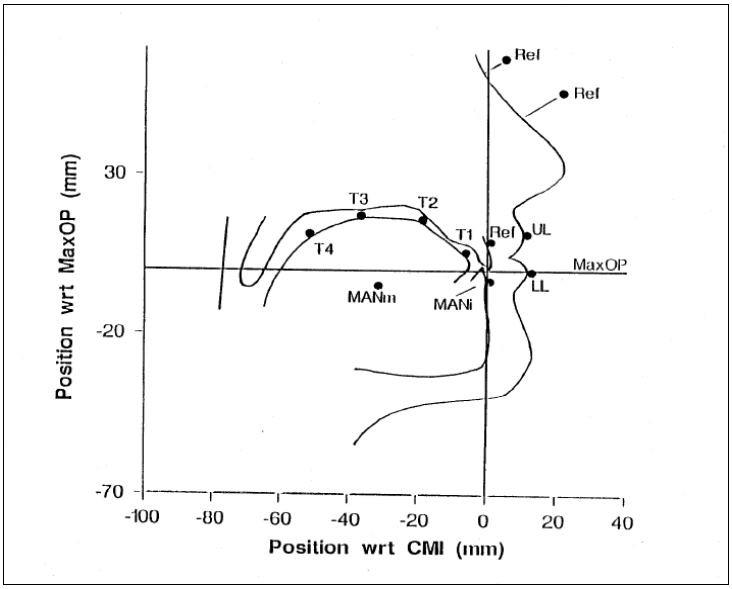


Figure 3 – Pellet locations used in X-Ray microbeam cinematography. (Image from the Westbury (1994))

X-Ray microbeam cinematography, whilst safe, is an expensive procedure, with only a handful of facilities available. Mis-tracking of the pellets can also occur, resulting in disappearing pellets. Gathering a large amount of human speech data using X-Ray microbeam cinematography is therefore both costly and somewhat impractical on this scale.

Electromagnetic Articulography

Electromagnetic Articulography (EMA) is a more recent method for obtaining human articulatory data that is both less invasive and less costly than X-Ray microbeam cinematography. In Electromagnetic Articulography, the position of receiver coils attached to articulators can be made known by calculating voltage differences between the coils and a

transmitter electromagnet. A current is induced in a coil when the electromagnet that is alternating at a given frequency comes within a certain distance from the coil. Each coil responds to a unique frequency, meaning that using a number of transmitter electromagnets allows multiple articulatory configurations to be measured relative to the transmitter. (Perkell et al 1992)

Older EMA systems, such as the Carstens AG100 and AG200, took only 2D measurements (x & y coordinates). Using only 2 dimensions, this method introduced errors whenever a coils position moved from being parallel with the transmitter. The accuracy of these systems depended on the orientation of the coil because the transmitter was situated directly above the middle of the coil. Any movement away from this optimum relative position resulted in the measured voltage difference to decrease and a subsequent increase in error to occur when calculating the coils position relative to the transmitter. New 3D EMA systems such as the AG500 avoid this problem by taking 5 measurements from each sensor coil in 3D space, allowing the coils to rotate and move naturally along with the articulator. (Richmond 2009)

There are a number of advantages to using this new 3D EMA system to obtain human articulatory data. Previously, as with X-Ray microbeam cinematography and 2D EMA systems, the speaker's head wasn't allowed to move around, as it would lead to the mis-tracking of pellets or errors the induced voltage measurements. This also meant that the number of utterances that could be recorded in a given session was severely limited. A 3D EMA system allows much longer recording sessions and therefore a much larger corpora of human articulatory data. (Richmond 2009). For this dissertation, I will be using the *day one* subset of a new EMA corpus (*mngu0*), which will be discussed in detail in Chapter 4.

Inversion Studies using Human Articulatory Data

A study by Papcun et al (1992) used X-ray microbeam human-articulatory data to perform the inversion task, specifically analysing the movement of the tongue during production of 6 American-English stop consonants. The acoustic waveform was parametrised into frames using an overlapping Welch window of length 15.98ms. A Fast Fourier Transform (FFT) was

performed to convert the frames to the frequency domain, the resulting coefficients were normalised, and the articulatory data was interpolated with the acoustic data-frame shift and normalised. A separate MLP was then used to train each of the three articulatory positions of the tongue (tip, blade and dorsum). Backpropagation gradient descent reduced root mean square (RMS) error between input and output until an optimum was achieved. RMS error is a frequently used measure in inversion systems, as it can tell you the distance between the estimated and actual articulatory trajectories.

The aim of this particular study was also to see which articulators were *critical* to the production of a particular phone, and Papcun et al (1992) found that for the critical articulator trajectories the RMS error was higher than for non-critical articulators trajectories. They concluded that their neural network was better able to estimate larger movements, such as those by critical articulators, than the smaller movements of non-critical articulators, which had a smaller range of movement.

A further study by Zach & Thomas (1994) into MLP use in the inversion task utilised a different error function that improved the accuracy of the trained neural network. Using an MLP trained using a *correlational & scaling error* (COSE) function resulting in a vowel classification accuracy of 87% compared to the 73% accuracy received by MLPs trained using a standard squared error function. However, this work, along with most other work using human articulatory data for the inversion task, focuses on a finite number of sounds, rather than whole continuous utterances of speech.

Furthermore, an MLP inversion system succumbs to the problem of ill-posed/multi-valued mappings. An MLP output is the estimate of the conditional average of the target vectors given the input vectors. The network is trained using a sum-of-squares error function backpropagating through the network. However, this error-function essentially performs unimodal Gaussian regression on the target data. Therefore, an MLP assumes a single Gaussian distribution for the target data points, and so non-Gaussian distributions and multi-valued mappings are not modelled. The accuracy of the MLP output is therefore limited to normally distributed data.(Richmond (2002), Richmond, King & Taylor (2003)).

Mixture Density Networks

Richmond (2001) and Richmond, King & Taylor (2003) proposed using an augmented MLP with a Gaussian Mixture Model (GMM) in a Mixture Density Network (MDN) to perform acoustic-articulatory inversion. The MDN provides a full conditional probability density function of an articulatory position in the entire articulatory domain given an acoustic vector. The GMM typically has three components, defined by control parameters of the mixing coefficients (or priors), the means and the variances. Adding more components means that potentially any distribution could be modelled, though too many could lead to over-fitting.

Here, an MDN was trained using Scaled Conjugate Gradient optimisation, which has been shown to converge to the optimal solution 20 times faster than standard gradient descent, used previously in MLP training. Training the MDN also minimizes the negative log likelihood of the target data given the component parameters. The MDN takes an input vector and, after backpropagation training, maps it to the control parameters of the mixture components.

In Richmond *et al* (2003), EMA articulatory data and corresponding acoustic data from the Multichannel Articulatory (MOCHA) database for a female Southern English speaker (*fsew0*) was used to train both an MLP and an MDN. Richmond *et al* propose an evaluative method of comparing the two systems by interpreting the MLP output (mean and variance) and the global variance of each channel per frame as a Gaussian PDF, meaning likelihood can be calculated. As the hypothesis in this study was to see if an MDN could provide a better model of possible articulatory points than an MLP, then an MDN would achieve a higher mean likelihood of the target data given the frame-wise probability density functions for each articulatory channel.

In comparing the two methods, it was found that an MDN did indeed show a higher mean likelihood for the target data than the MLP, with an average improvement of between 2.6% to 21.8% for all articulators, except the velum. They also showed a higher likelihood of the y coordinate over the x coordinate for the speaker, which could not be explained without more speakers being tested.

A more recent study undertook by Richmond (2009) used a new dataset (the same as used in this dissertation project) that was found to be more consistent than that of the MOCHA EMA data. Inconsistency in the *fsew0* articulatory data occurred when coils became detached during the recording, which is a common occurrence in EMA techniques, especially with 2D articulographs when movement of the head is restricted and can cause measurement errors, as discussed previously. The new dataset, as discussed in chapter 4, is demonstrably far more consistent than *fsew0*. Using this *mngu0* dataset, by including velocity and acceleration features of the articulatory data to train an MDN. An additional maximum likelihood parameter generation algorithm (MLPG) was applied to the resulting PDFs, obtaining the most probable trajectory for each frame and each articulator. This system is called a Trajectory MDN (TMDN), and had been developed in a previous study by Richmond (2007), using the *fsew0* dataset.

Comparing these two studies, the average RMS error for the *fsew0* dataset was 1.54mm, while the new *mngu0* dataset resulted in an improved 0.99mm RMS error, which demonstrates the importance in obtaining a consistent dataset.

MDNs have also been shown to be an excellent method of examining articulatory non-uniqueness in the inversion mapping, as Figure 4 shows. As the MDN outputs a frame-wise probability density, this can be used to visualise probability density for an articulators range of movement (x or y coordinate) in what is called a *probabiligram*. The denser a frame, or sequence of frames, the darker it will appear on the graph. Conversely, frames with a high PDF variance will appear lighter and wider. For frames that exhibit non-uniqueness or multi-modal distributions, multiple dense regions will appear, as evident in Figure 4.

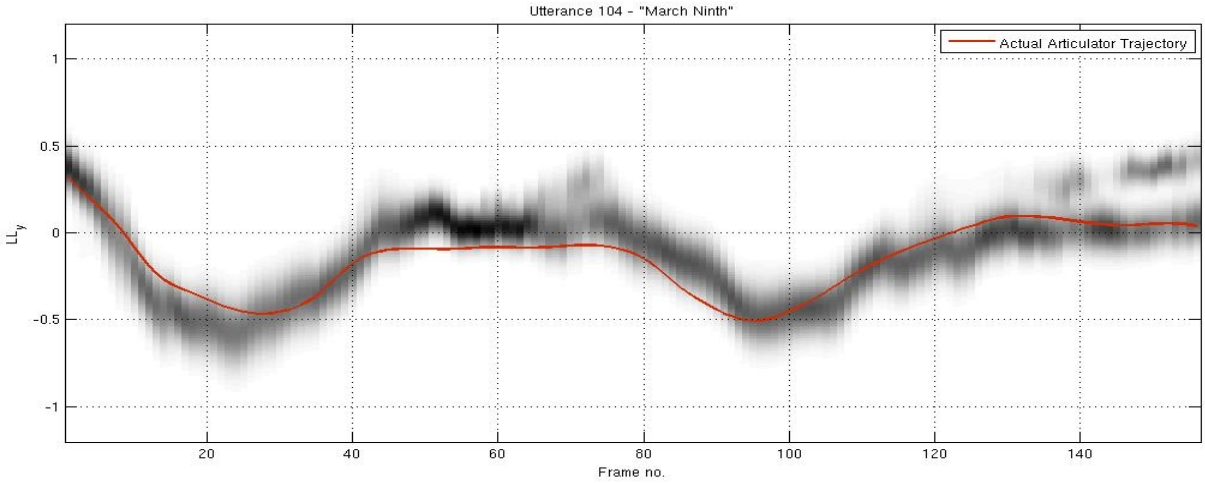


Figure 4 – Frame-wise MDN output of the y coordinate of the lower-lip (LL) for the Words "March Ninth". At frame 140 onwards, clear multi-modality can be seen.

Previous studies in Non-Uniqueness

Previous work concerning the study of non-uniqueness have used widely different methodologies in an attempt to both resolve and quantify the extent of non-uniqueness in the inversion mapping.

Roweis (1999)

Roweis (1999) used a large database of paired acoustic Line spectral pairs (LSP) and X-Ray microbeam articulatory data to empirically investigate the inversion mapping. Roweis selected a key frame, such as the centre frame of a given phone, and, using a Mahalanobis distance measure, found the nearest 1000 data points to it from the entire database in LSP space. The corresponding articulatory vectors for each articulator (x,y) were then scatter-plotted to reveal the distribution and spread of the data points. Roweis discovered that a large proportion of the resulting plots showed a wide spread of data points for a single frame. Multi-modal distributions were also apparent in the scatter plots, indicating varying possible configurations of articulatory positions that can produce a single sound. Going further, Roweis examined the spectral envelopes of these 1000 LSP data points and found that these

distributions did indeed have similar spectral shapes.

Qin & Carreira-Perpiñán (2007)

The first systematic large-scale empirical study into non-uniqueness in the inversion mapping was carried out by Qin & Carreira-Perpiñán (2007), and aimed to quantify exactly how much of human speech could be considered to exhibit a non-unique inversion mapping. Using the Winconsin X-Ray microbeam database (XRDB), which provided acoustic and articulatory pairs of vectors (x,y) for a single speaker, Qin & Carreira-Perpiñán used statistical machine learning techniques to investigate the number of modes returned from the inversion mapping. Their method consisted of two stages; *inversion* and *clustering*.

First, using Linear predictive coefficients (LPC) for accurate spectral representations of the vocal tract, a single representative acoustic frame was chosen and fixed. From this frame, the Itakura distance measure, approximate to perceptual distance, was used to search for acoustically similar frames within a given threshold. Here, a threshold of 0.04 was chosen as the database frames were 0.06-0.1 apart. Too large a threshold would result in too many clusters, and too small a threshold would result in too few. Given the returned set of acoustic vectors, a set of corresponding articulatory vectors could be obtained.

To cluster the set of articulatory vectors, Qin & Carreira-Perpiñán (2007) used a non-parametric mode-finding algorithm called *Mean-Shift*. This is a hill-climbing algorithm that treats data points as probability density functions and can be used to find the modes in a given cluster. This will be discussed in further detail in chapter 3.. As the number of clusters were not known, a density was defined using a non-parametric kernel density estimate, and this bandwidth was used to determine which data point belongs to which cluster.

The inversion and clustering steps were performed on every acoustic-articulatory pair in the 45000+ database, and resulted in approximately 5% of all acoustic vectors showing multi-modal distributions of corresponding articulatory vectors. Inspection of the resulting point clouds for a selection of vowels such as /ae/ /u:/ and /y/ showed uni-modality, whilst

approximants such as /w/, /l/ and /ɹ/ showed non-uniqueness and multiple modes in the resulting point clouds. For those phonemes that showed multi-modality, Qin & Carreira-Perpiñán then examined the spectral envelopes of the returned acoustic frames and found that they were of similar shape. Overall, they concluded that whilst non-uniqueness did occur in human speech, it was infrequent.

Qin & Carreira-Perpiñán (2010)

Following their work on articulatory non-uniqueness in the inversion mapping, Qin & Carreira-Perpiñán (2010) investigated the use of conditional density modes in performing the inversion task. This time they focussed on the multiple articulations of American English alveolar approximant /ɹ/. This particular phone can be produced in two different ways; 'bunched', where the tongue dorsum is raised and the tongue tip is lowered, and 'retroflex' where the tongue tip is raised and the tongue dorsum is lowered. Using the XRDB and pairs of acoustic-articulatory vectors, a Gaussian Mixture Density model was learnt using the EM algorithm. Conditional densities (p(x|y)) were estimated from this model for every acoustic frame of /r/, and a variation of the mean-shift algorithm was used to initialise at each centroid of the conditional mixture data.

Carreira-Perpiñán (2000, 2003) had previously researched mode-finding in Gaussian mixtures. According to scale-space theory, for a 1 dimensional mixture of Gaussians, the number of modes is equal to the number of components when the scale is small, and this number decreases as the scale gets larger. Conversely, Carreira-Perpiñán found that when the dimensionality of the Gaussian mixture is larger than 1 and the components aren't isotropic, the number of modes can be higher than the number of components.

Ultimately, the 'smoothest' trajectory of modes is then recovered by using dynamic programming to minimize the objective function over all the modes. This trajectory is the sequence of modes that move the slowest (minimum-energy of motion).

The algorithm chosen used the density model to predict and recover multiple articulatory

configurations, and was successful in the task, finding clear instances of non-uniqueness for alveolar approximant /ɹ/ in initial, intervocalic and final positions. However, the mode-finding algorithm used was computationally costly, especially when increasing the number of components in the Gaussian mixture, but this complexity could be reduced by using an accelerated mean-shift algorithm, or by thresholding out mixture components far from the acoustic reference frame.

Neiberg, Ananthakrishnan & Engwall (2008)

Neiberg, Ananthakrishnan & Engwall (2008) used Gaussian Mixture Models (GMMs) to try to quantify the non-linearity and non-uniqueness of the inversion mapping seen in previous studies. Using normalised EMA data and MFCCs (at a sampling rate of 125Hz) for the acoustic data, Neiberg *et al* chose to ivnestigate clusters in the resulting distribution. The inversion data was fitted to a GMM and the Expectation Maximization (EM) algorithm was used to find Maximum Likelihood (ML) estimate vectors of the data, with Bayesian Information Criterion being minimized to result in optimum clustering.

To study non-linearity, Neiberg *et al* (2008) took the articulatory data points corresponding to the acoustic Gaussian and modelled a corresponding number of Gaussians. If there were multiple articulatory Gaussians for the corresponding acoustic Gaussian, then that mapping could be said to be non-linear. Keen to stress that multi-modality in a given inversion mapping does not necessarily mean that non-uniqueness has occurred, Neiberg *et al* (2008)presents a further measure to differentiate non-linearity and non-uniqueness. If the acoustic data points have exactly the same distribution in acoustic space, but varying distributions in articulatory space, then that mapping can be said to be non-unique. Using Non-Gaussianity of the data (examining kurtosis) & Bahattacharaya distance as a measure, Neiberg *et al* define non-uniqueness as the inverse of Bahattacharaya distance weighed by its Gaussianity.

Investigating non-linearity and non-uniqueness for a number of consonant phonemes in the database, Neiberg *et al* found that stop consonants and fricatives showed high levels of non-

uniqueness, whilst liquids (eg. /l/ & /r/) were shown to be highly non-linear but unique in the inversion. Neiberg *et al,* postulated that voiceless alveolar fricative /s/ and voiced /z/ showed this articulatory non-uniqueness due to an inability of the EMA data to fully capture the exact location of the tongue tip for a given acoustic frame.

Anathakrishnan, Neiberg & Engwall (2011)

In a study by Anathakrishnan, Neiberg & Engwall (2011), continuity constraints were proposed as a way of resolving all instances of non-uniqueness found in the inversion mapping. Estimating the trajectory of an articulator, or how it actually moves, can be achieved using the acceleration and velocity coefficients of the sampled articulatory data and had previously been shown to occasionally reduce instances of non-uniqueness in phones such as /ɹ/ by Qin & Carreira-Perpiñán (2010).

Acoustic/articulatory vector pairs from the MOCHA-TIMIT database were used. The acoustic data was parametrised as Mel Frequency Cepstral Coefficients (MFCCs), so as to provide a good representation of the vocal tract and to eliminate the chance that instances of silence in the data would be deemed non-unique. Principal Component Analysis (PCA) was used to reduce the dimensionality of the acoustic features to be closer to the articulatory data, and the x and y coordinates of the articulatory data was low-pass filtered and down-sampled to match the acoustic frame-shift. Conditional distribution of the inversion mapping was constructed by modelling the data to a GMM, and the peaks in this conditional probability function were obtained using Box Induction, which is a bump-hunting algorithm.

Plotting the resulting conditional probability function for the inversion mapping over a sequence of frames in an utterance produced a ‘path’ of the most likely movement of the articulator. Anathakrishnan *et al* classified the instances of non-uniqueness into two distinct paths types: ‘Along the same path’ (ASP), where the path has a minimized continuity constraint before and after the non-unique instance, and ‘With a change in path’ (WCP), where different paths minimize the continuity constraint before and after the non-unique instance. ASP was found to be more frequent during a lengthier sequence of non-uniqueness,

and overall the continuity constraints reduced prediction error by as much as 53%. However, up to 22% of non-unique mappings could not be accurately predicted using continuity constraints or by using the mean value of the two paths, which was unexplained by the authors. Ultimately, non-uniqueness is left unresolved by Anathakrishnan *et al*, but the paper offered further evidence that applying continuity constraints can help resolve a substantial proportion of non-unique instances found in the inversion mapping.

Discussion

The outline for this dissertation is to empirically quantify the amount of non-uniqueness in a given dataset. Depending on the dataset, and how you try to define it, non-uniqueness can occur in 5% of the dataset (Qin & Carreira-Perpiñan 2007) or it can be shown to exist to some level in nearly all phonemes (Neiberg *et al* 2008). A conditional probability density estimate of the data, as used by Qin & Carreira-Perpiñan (2010) or by the MDN models of Richmond (2001, 2009) and Richmond et al (2003), can provide an increasingly accurate estimation of articulatory configurations conditioned on the corresponding acoustic frames. However, in this project I not concerned with estimating the most likely acoustic-articulatory inversion mapping. Rather, I am trying to quantify from the acoustic-articulatory pairs, how much of the dataset when inverted yields a multi-modal cloud, and concurrently, which articulators and phones contribute most to non-uniqueness in the dataset.

The studies by Qin & Carreira-Perpiñan (2007, 2010) focused on using a mode-finding algorithm to determine the number of modes in a point cloud, with the latter paper using these modes to derive the single most likely trajectory from the resulting modes. For this study, I intend to use a variation of this mean-shift algorithm not to find the modes in the dataset, but to determine the articulatory distribution of the data through the number of clusters found (discussed in detail in chapter 3). This non-parametric segmentation algorithm is suited to my aims as, though it does not find the individual modes of a dataset, it is a very fast clustering algorithm that, depending on the threshold you use, will segment an elongated cloud or distinct dense areas in the cloud into a clear number of clusters.

The size of the data-set, nearly (800 000) vectors, means that if I was to truly determine the percentage of non-unique frames, using the definition of mean-shift mode-finding, then the order of complexity would be very large ($O(In^2)$, given I iterations and n data-points).. Furthermore, I believe that many instances of non-uniqueness have as much to do with the parametrisation of the acoustic data itself. As will be discussed, acoustic data that has been normalised and parametrised may, if grouped by parametric similarity, be shown as more similar as they may actually be, perceptually (this sentence doesn't make sense). In this respect, a non-unique articulatory cloud may not show multiple articulatory configurations. Rather the parametric similarity of a number of acoustic frames may be misleading.

Following Neiberg et al (2008), I will be quantifying the level of articulatory non-uniqueness that occurs for selected phones in the dataset, in order to determine which phones are quantifiably the most unique in the data-set. Extending this premise, I intend to discover which articulators contribute most to non-uniqueness. The range of movement for a particular articular, such as the lower incisor, may be lesser than, say, the tongue tip, which must be taken into account. Discovering for which phone and articulator non-uniqueness occurs in the mngu0 acoustic-articulatory dataset would be pertinent and this particular data-set is demonstrably more consistent and larger than the datasets used by Qin & Carreira-Perpiñan (2007, 2010), Neiberg et al (2008) and Anathakrishnan & Engwall (2011).

Chapter 3. Methodology

Overview

The content of this methodology is concerned with, and segmented into two main steps. First, there is the inversion step, which here consists of using a data-partitioning technique to allow a distance measure to be quickly and efficiently performed on a set of acoustic vectors. This will allow acoustically similar sounds to be compared parametrically, extracted, and paired with corresponding articulatory vectors. As the acoustic and articulatory frames are frame-aligned, this is a straightforward task. Secondly, in order to both explore and quantify articulatory non-uniqueness in the inversion mapping, a method of clustering the data into distinct articulatory parts will be employed. This will allow a thorough exploration of the inversion mapping.

Inversion

KD-Trees

Binary Space-Partitioning (BSP) is a commonly used method of handling complex spatial data, and is typically used in computer graphics. A data-set is recursively divided into two, partitioning the data based on a set of requirements. The subdivision of the space into convex sets by hyperplanes means the resulting data-structure is a hierarchical tree structure with a set of linked nodes. *k*D-trees are a special case of BSP and multidimensional data-structures. Here, the data is used to determine the splitting value and the splitting dimension for each node. The multidimensional space is partitioned into hyperrectangles, which are defined at each node, and the left and right children of the node are made known to the node. Each node is split according to the defined splitting value and dimension. This split can be represented by a hyperplane perpendicular to the splitting dimension's specified direction (a hyperplane

has (k -1) dimensions in k dimensional space) (Moore (1991)).

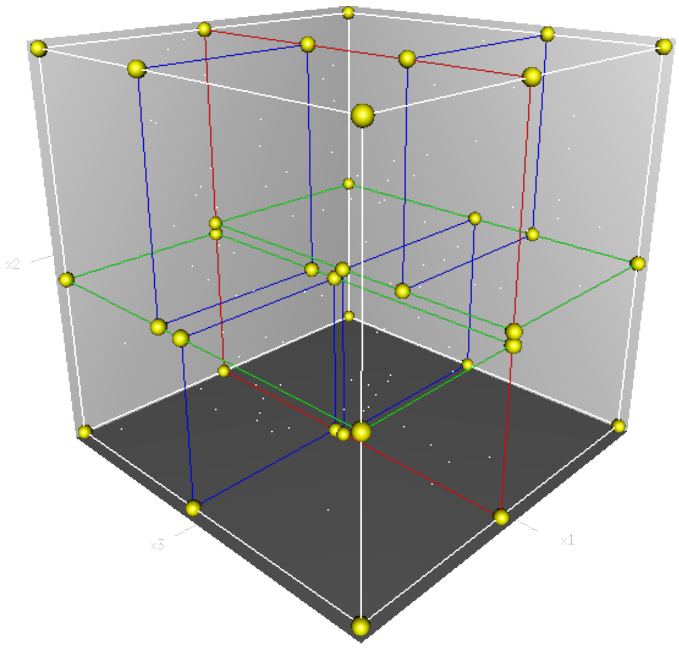


Figure 5. A 3D representation of a kD-tree. Beginning with the root cell (white), each section is partitioned into smaller and smaller subsections. (image from Wikipedia commons)

Construction

The construction of the kD-tree begins at the root, taking the training set (here, the acoustic frames), and working recursively to split each child of a node until termination. The dimension with the greatest variance is chosen as the splitting dimension, and is calculated by computing the sample variance S_{ii} from the data-set and choosing to split in the dimension where S_{ii} is the largest. The location of the split is chosen by the median value of the data along the splitting dimension, and this in turn results in a more balanced tree. If the data is unevenly distributed, with the result being skinnier hyperrectangles, then the location of the split is chosen by the mean value of the data along the splitting dimension.

In order to construct a kD-tree of the LSF acoustic data for use with a nearest neighbour search, I used code[1] that took in 799,159 LSF acoustic frames in 40 dimensions and built a partitioned tree of the data. From here, a nearest-neighbour search could be performed from a given query point.

1 This code was kindly provided by Korin Richmond.

Nearest-Neighbour Search

The task of finding the closest points in a space is a thoroughly researched problem, and primarily Euclidean distance is the measure used for determining the distance between any two points. Linear search methods, considered to be naive approaches to the problem, compute the distances from a given point to every other point in the data-set, and keep track of the closest points. This method has a complexity of *n(O(nd))*, given the number of training examples. The data structure of a *k*D-tree allows a nearest neighbour search that does not require comparison of a query point with every other point in the data-set (Moore (1991)). Given the large number of acoustic frames in the data-set, the *k*-nearest neighbour search will be an accurate and fast search method.

The nearest neighbour search can be performed once the *k*D-tree is constructed, and comprises of a depth-first search and the intersection of a hyperrectangle and a hypersphere. The average case complexity of a nearest neighbour search is O(lg *n*), with the worst case complexity being O(*n*). The advantage of this *instance-based learning* method is that it just involves storing the data, and that it does not perform an exhaustive search.

The nearest neighbour algorithm is comprised of the following steps

Working recursively from the root node, the algorithm moves down the tree to the leaf node that contains the query point in the leaf node's hyperrectangle.

This leaf node is saved as the current best estimate of a nearest neighbour.

A hypersphere is constructed around the query point and the radius of this hypersphere is the distance to the current leaf node.

At the root node, nothing can be ruled out, and the nearest neighbour must fall within the hypersphere. The algorithm checks the left child to see if there is a closer nearest neighbour than the current best candidate.

If the left node is closer to the query point, then that is the new current best candidate.

Going back to the parent of the candidate node, the algorithm checks the sibling of the left

node to see if any of the right hand children have hyperrectangles that intersect with the current best candidate's hypersphere.

- If so, then a depth-first search can be performed to see if it is a new best candidate.
- If not, then the regions and nodes that don't intersect with the hypersphere can be disregarded.

The number of nearest neighbours that were to be returned for a given frame is largely dependent on the size of the data-set and the parametrisation of the acoustic and articulatory data. Too many nearest neighbours returned could result in acoustic data that is too far away from the query point being included in the search. Too few, and you could miss out on a range of possible articulatory configurations for a given frame. Following the study by Roweis (1999), the number of frames to be returned from a single query point was decided to be 1500. This returned a reasonable distribution of frames that, when plotted, would show regions of articulatory and acoustic feature space with clear distribution of dense regions.

In performing the instantaneous inversion mapping, the nearest neighbour *k*D-tree search method returns the indices of the desired number of acoustic frames, sorted by distance to the original query point. As the acoustic and articulatory data-sets are aligned by frame, the inversion simply required extracting the corresponding articulatory frames. Scatter plots of this inversion can be seen in Figure 6 for the voiceless dental fricative /θ/. The x and y axes correspond to the x and y coordinates of each EMA articulatory coil. As is evident from the distributions in the scatter plots, especially for the lower lip and tongue tip for this particular phone /θ/, there is visual evidence to suggest a non-unique inversion by the multiple densely plotted regions. In a number of plots, there is quite a wide spread of data-points that suggests that there may be more than one cluster and therefore perhaps more than one articulatory configuration.

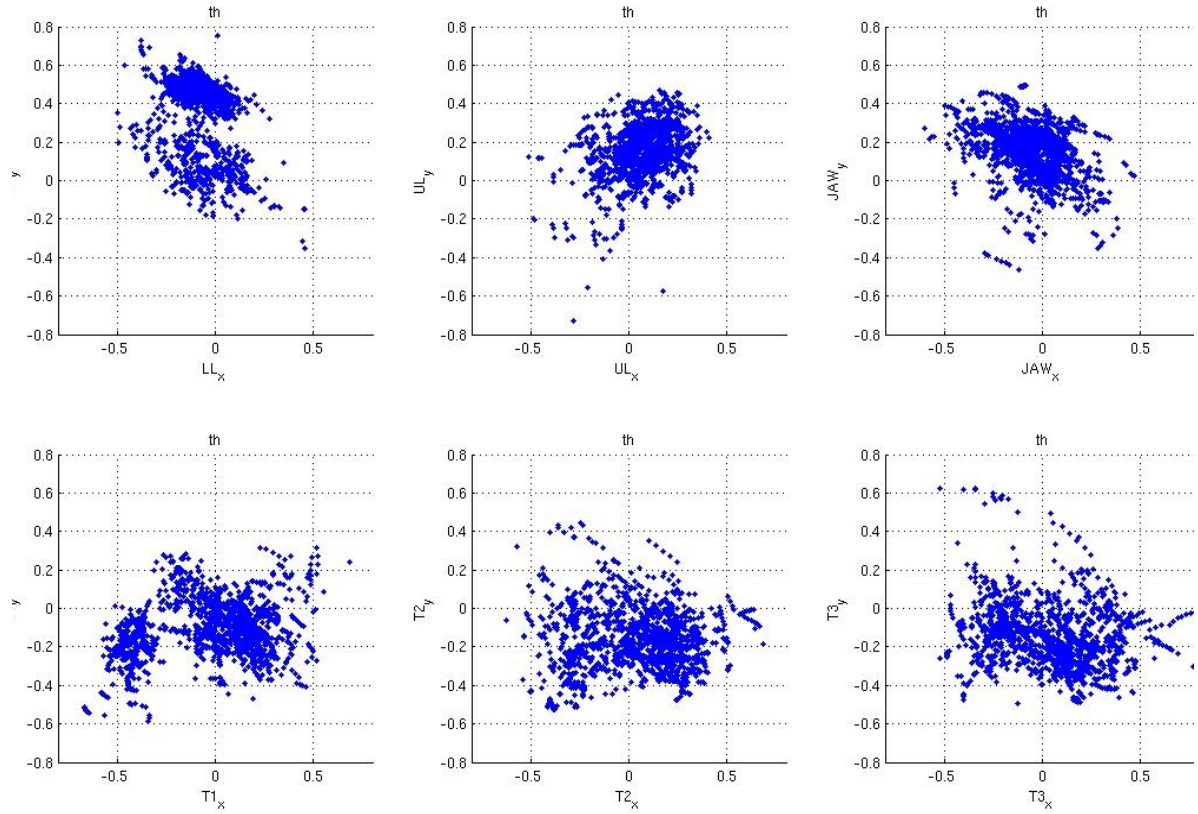


Figure 6: Inversion results for 1500 frames of /θ/ in *day one* of the *mngu0* corpus for 6 articulators: *from top left*; lower lip (LL), upper lip (UL), lower incisor (JAW), tongue tip (T1), tongue blade (T2) & tongue dorsum (T3).

Query Point Selection

The acoustic-articulatory data that was provided also contained phone labelling that was produced by automatic forced-alignment, using *Multisyn* and the *Combilex* lexicon. The lexicon is an orthographic-phonemic representation of speech and is provided frame-aligned with the acoustic-articulatory data. This meant that for a given frame, the phone label was available which could help disambiguate certain frames if their distribution raised suspicions as to their identity. However, the labelling is not a fine-grained phonetic representation of the

speech data, and so care must be taken in investigating how similar two acoustically similar frames are. The context of a phone must also be taken into account, as *coarticulation* is prevalent in continuous and conversational speech. Here, the end of one sound will influence the sound of another, as is anticipatory assimilation, such as with nasal consonants shifting their place of articulation in anticipation of a following stop consonant (eg. 'Hand bag' sometimes becomes /hæmbæg/ in continuous speech.). Along these lines, I decided to omit diphthongs, triphthongs and affricates, as any frame found could possibly be unrepresentative of the entire sound due to the complexity of the sound to begin with.

Using the label set, I returned all the unique phone labels that occurred across the 1263 utterances. This was to be the basis for selecting a representative phone for the data-set. The returned acoustic nearest neighbours were in sequences of frames corresponding to the utterances in the database. In order to find a representative frame for a given phone as a query point, I used the sequence of frame indices to find the corresponding utterance and chose an appropriate context for each phone. This was a manual task, as determining what exactly was an appropriate and representative context was both ambiguous and relied on trial and error. Listening to the corresponding audio file was one such method of figuring out if the labelling was correct, and generally the centre frame from the returned vectors was chosen. Producing a scatter plot of a few samples per phone and visualising the distribution to see if it was widely different per sample was another such method.

Clustering

The scatter plots of the inversion for each phone can obscure the true density of the data, with seemingly elongated point clouds perhaps actually concealing multiple dense regions that could indicate a non-unique articulatory distribution for that phone. In Qin & Carreira-Perpiñán (2007), they indicate that multi-modal distributions can be concealed by elongated articulatory clouds. An unsupervised learning method for automatic clustering of subsets in the data should disambiguate the point clouds. Though finding the modes of each point cloud is not the aim of this project, the number of clusters that emerge from a given point cloud can reveal the distribution of phone articulations. A spread distribution in the point cloud could

hold many clusters, whilst a denser and more specific point of articulation should present a single cluster. Initially, both K-means clustering and Gaussian mixture model clustering were explored.

K-Means Clustering

K-means clustering is a fast parametric clustering technique that accepts two inputs; The data itself and the number of clusters you want to find. If the number of clusters in the data is known, then K-means clustering is an efficient method of partitioning the data into pre-specified clusters. The algorithm works by initially choosing a certain number of random points, making them the initial centroids in each cluster, that being the point of the cluster most representative. Every point in the data is assigned to the nearest cluster, and the clusters are updated so that the centroid is the average of all the points in the cluster. The resulting clusters are all spherical, or elliptical. The algorithm itself has a complexity of O(knT), where T is the number of iterations taken and k and n are the number of clusters and the number of data-points, respectively. (Mackay (2003)).

The problem with using K-means clustering in this study stems from the fact that we do not know how many clusters there are in our data. As previously discussed, visualising the scatter plots often showed elongated point clouds that could potentially hold a number of modes or clusters, but we do not know exactly how many, and assigning a set number of clusters would not be suitable to this task. Furthermore, the shapes of clusters in the point clouds may not be spherical or elliptical, rather they may be arbitrarily shaped.

Gaussian Mixture & Expectation Maximisation

One possible solution to the issue of an unknown number of clusters was to fit a Gaussian mixture to the returned nearest neighbours using an Expectation Maximisation (EM) algorithm. Using the returned Gaussian mixture and the returned nearest neighbours, a clustering algorithm partitions the data based on the number of components in the Gaussian mixture (Bishop (1995)). The number of components required varies depending on the data

itself, and therefore requires visual inspection to ensure how many clusters there actually are.

Following this method, I attempted fitting a Gaussian mixture to the data using either 1,2, 3 or 4 components, running the EM algorithm and clustering the data. For a single component, there was only a single cluster for both visually possible uni-modal and multi-modal distributions across all the articulators. For 2 or more components, the point clouds were clustered into as many distinct clusters as there were components, which meant that the results were often ambiguous as to how accurately the clusters represented the actual data. Though not to the same extent as with k-means clustering, the number of resulting clusters appeared to be dictated by the number of components. In some examples, visual inspection of the resulting clustering indicated that some clusters were perhaps unnecessarily created in dense regions of the cloud where one would have been more appropriate. It appeared to be a case of the clustering algorithm trying to represent all the components of the mixture model. In other examples, correctly distinct clusters were identified, with outliers being represented by a component in the mixture.

Mean-Shift

In order to cluster the data in a fast & non-parametric way, I chose to use a Gaussian-Blurring Mean-Shift. This algorithm was first proposed by Fukanaga & Hostetler (1975), and was adapted and renamed as the 'blurring mean-shift' by Cheng (1995), and proposes a way of using a Gaussian Mean-Shift algorithm to initialise a converging clustering algorithm that results in a similar number of clusters as provided by mean-shift, but at a much faster rate. This algorithm is therefore well suited to my task, as the linear order convergence rate of a Gaussian mean-shift (p=1) would be computationally costly for a dataset as large as *day one* of *mngu0*, which I will be using for this project. A Gaussian blurring mean-shift converges in a cubic order (p=3). I will provide a detailed description of both algorithms here.

Gaussian Mean-shift

Gaussian mean-shift is an algorithm that is widely used in image segmentation tasks, such as separating out and identifying each distinct colour from the pixels in a picture. Proposed by Fukanaga & Hostetler (1975), the mean-shift is a powerful non-parametric algorithm that can be utilised to perform tasks in computer vision, such as image-segmentation, image tracking and smoothing but can also be used for mode-finding and clustering (Qin & Carreira-Perpiñán (2007, 2010) . As it does not require any prior knowledge of the number of clusters, and does not assume a spherical shape, its use has become increasingly popular in a number of fields of study. The Gaussian Mean-shift is performed in three main stages, and is essentially a local-maxima finding algorithm similar to the Expectation Maximisation algorithm. These stages are that of gradient-ascent, iteration and convergence.

Gradient-Ascent & Convergence

The mean-shift algorithm makes the assumption that each point in a data-set is sampled from a probability density function. Each dense region of the point cloud is therefore considered to correspond to a mode or area of local-maxima. Stationary points in a dense region are chosen as candidates for the modes of the cluster, and the gradient-ascent is used to find these points.

At each point of the data-set, hill-climbing is performed based on the density contour. The density of the data-point is calculated using a *kernel density estimate* which returns the weighted mean of the region. In gradient ascent, the data-point is considered the centre of a window. The mean of this window is calculated, and the window then moves so that the newly calculated mean is the new centre of the window. This process is repeated until all points converge to a particular local-maxima, or mode, that are the stationary points of the cluster. The size of the window, or the *kernel density*, is therefore important in determining how the data-points converge through the weight it gives to nearby data-points in re-estimation of the mean, and requires tuning so that the number of clusters returned is neither too few or too many.

Depending on the size of the data-set, the complexity of Gaussian mean-shift can limit its applicability in certain task. Given n data-points, and I iterations, the order of complexity on

standard mean-shift is $O(In^2)$. Acceleration strategies for Gaussian mean-shift have therefore been required that make this algorithm computationally more efficient and its usage more widespread. One such strategy would be to use a less stringent tolerance, or bandwidth, between each mean-shift window. This would mean that the number of iterations would be reduced as larger areas around data-points are considered per mean calculation, and therefore the convergence rate would be more rapid. Care must therefore be taken to ensure that the true modes of each cluster in the data-set are found.

Another method of GMS acceleration is that of the adaptive mean-shift algorithm, wherein the *k* number of nearest neighbours are used as a means of adjusting the bandwidth for each data-point. This way, a closer data-point will require a smaller bandwidth than a further away data-point. This would be useful as not all data is linearly spaced. A final strategy is to iteratively alter the position of the data-points, moving each closer together to speed up convergence. The Gaussian Blurring Mean-Shift (GBMS) takes this idea and expands it for use as a fast non-parametric clustering algorithm.

Gaussian-Blurring Mean-Shift

In blurring mean-shift, each point x_m in a given data-set X containing $\{x_1,...,x_N\}$ moves towards a specific point $f(x_m)$ in a region, and becomes a new data-point called X^y This stage utilises the first step of the mean-shift algorithm. Applying this to each data-point in turn reduces, or *blurs*, the size of the data-set. Iterating this process results in the creation of several incrementally smaller data-sets that can be used vastly reduce the dimensionality of the data-set and thus reduce the complexity of Gaussian mean-shift in clustering (Carreira-Perpiñán (2000)).

Work in developing this technique by Fukanaga & Hostetler (1975) used a kernel density called the Epanechnikov kernel, defined as so;

$$K(u) = \frac{3}{4}(1 - u^2)\,\mathbf{1}_{\{|u|\leq 1\}}$$

Further work by Cheng (1995) demonstrated that convergence was possible with blurring mean-shift using a Gaussian kernel, which better models data distribution by showing the mean μ and variance σ^2 in a bell-shaped curve, the function of which is defined as so;

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

However, though Cheng (1995) demonstrated that GMBS with an Epanechnikov kernel was computationally efficient, Carreira-Perpiñán showed that use of a Gaussian kernel produced a better segmentation of the data as it could allow the data to be modelled more accurately. In this paper (Carreira-Perpiñán (2000)), each data-point was treated as a Gaussian-mixture and was used as the kernel-density estimate needed for the mean-shift procedure. The algorithm proposed by Carreira-Perpiñán demonstrated that by stopping the blurring mean-shift algorithm just as the data has been segmented into distinct regions then rapid convergence occurs. This would result in an accelerated clustering technique that produces excellent segmentation of a point cloud.

Though the values and number of modes in the data-set are not used in cluster convergence, Carreira-Perpiñán (2000) notes that the segmentation quality of the procedure is comparable to standard mean-shift algorithms. For the task of quantifying the level of non-uniqueness apparent in the entire data-set, in distinct phones and across all articulators, the number of resulting clusters found by the GMBS algorithm should give a good indication what articulatory non-uniqueness may be occurring and where, depending on the bandwidth chosen.

The GMBS algorithm moves each data-point, or *centroid*, according to the local average defined by its neighbours in the data-set. At each iteration in the algorithm, each centroid moves closer together, and this changes the shape and size of the data-set. The resulting clusters become compacted after a few iterations, and are distinct from each other. However, once each cluster is compact, further iterations proceed to move each centroid cluster towards

each other until convergence. This increases the possibility of clusters merging, reducing the accuracy of the segmentation in terms of the number of clusters returned. This also increases computation time in the event of clusters not-changing much.

To combat this, Carreira-Perpiñán (2000) proposed a stopping criterion, coming into effect once distinct clusters emerge, using histogram theory. In a histogram, data is partitioned into 10 linearly spaced containers (or bins), storing the frequency of each value in a separate bin. Using this, the stopping criterion compares the non-empty bins of each data-point in each cluster. Considering that each point in a cluster moves equally, comparison of the entropy of each cluster's histograms at two subsequent iterations will determine when the points in a cluster have stopped moving separately. Thus, the algorithm will stop, and distinct but non-converged clusters will be returned.

The algorithm I will be using to cluster my articulatory point clouds is an accelerated variation of this algorithm, developed by Carreira-Perpiñán (2000). Comparable to the connected-components stage of GBMS but applied at every iteration, the accelerated GBMS algorithm replaces the distinct clusters that collapse into a tightly compact point with a single data-point. This point is weighted according to the number of points it replaced in the cluster. This turns out to reduce the complexity of the GBMS dramatically, from $O(kN^2)$ to an accelerated $O(N^2)$, at no reduction in quality.

In this project, the number of clusters obtained is my measure of articulatory non-uniqueness, and so whilst the clusters themselves may contain modes that could be obtained and quantified using a more thorough local-maxima finding algorithm, such as by GMS, I may not always be finding clusters that correspond to the true mode or modes of a dense region of feature space. Interpretation of the results of this study will need to bear in mind this caveat.

Bandwidth tuning

To obtain a description of clusters found in a given articulatory point cloud, the bandwidth setting of the GBMS algorithm, or tolerance level, needed to both accurately reflect the

number of distinct clusters and also not result in too many spurious clusters. An articulator's position and movement during production of a recurrent phone in a recurrent context can naturally vary per occurrence. The human tongue, whilst extraordinarily accurate and consistent in facilitating sounds through its movement and positioning, will not reach the exact same coordinate each and every time. Moreover, the acoustically similar frames that give the corresponding articulatory positions may result in a seemingly larger number of articulatory configurations that is more to do with the parametrisation of the acoustic data than the actual way it was articulated.

Taking these points into consideration, visualising the resulting clusters and fine-tuning the bandwidth is a very good idea. Qin & Carreira-Perpiñán (2007) noted that the bandwidth parameter in clustering can make the difference between spurious modes (and therefore spurious clusters) and one large cluster. Using too small a bandwidth, or too large a bandwidth would result in either an incorrect multiple clusters or a single incorrect large cluster, respectively.

Both the acoustic and articulatory data-set have been z-score normalised, and so the range of the data falls between [-0.9 ; 0.9]. As the GBMS treats each point in the data-set as a probability density function, and because I will be assuming that the data is normally distributed for use with the algorithm, the kernel density estimate of the data is essentially a Gaussian mixture, albeit one where there are as many components as data-points, all with the same covariance and weight. In initial tests, a bandwidth of $\sigma=0.02$ was chosen based on histogram plots of similar and different phones in the data-set. The results showed a huge number of clusters for a number of phones that have previously been shown to have a unique mapping (Qin & Carreira-Perpiñán (2007)), and generally for every frame I tested. Voiceless alveolar fricative /s/, for example, has been shown previously to yield a uni-modal cluster for the tongue tip (T1) in the articulatory domain, but here, with a small bandwidth, spurious clusters arose, as shown in Figure 7.

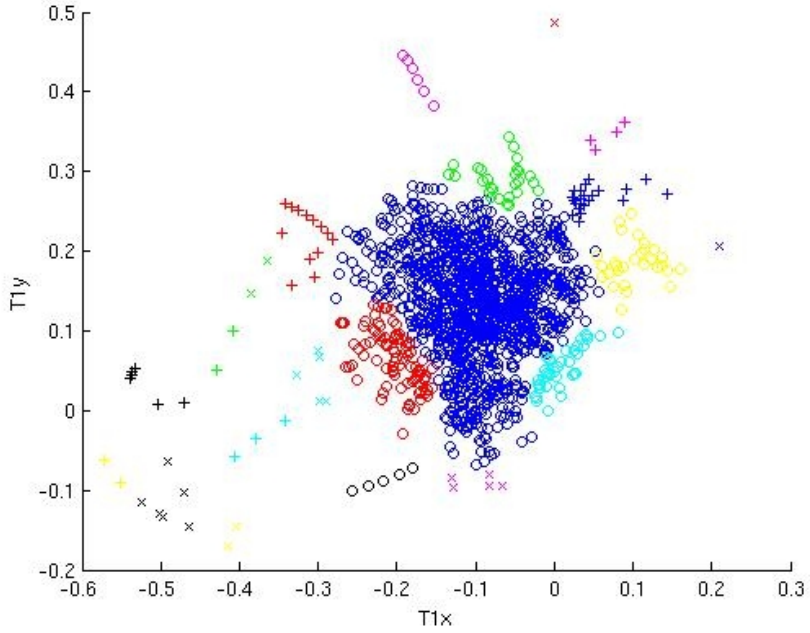


Figure 7. GBMS clustering of tongue-tip (T1) articulatory coordinates associated with 1500 acoustically close frames of /s/. As shown, though predominantly uni-modal, spurious clusters have been identified.

The tongue tip needs to be in a particularly definite position to produce this particular phone, and previous studies have shown the tongue tip to have a lower variance than other articulators. For example, in Richmond, King & Taylor (2003), they ranked the phone dependent averages from the MDN output by average variance σ^2 for the tongue-tip articulator, and showed that fricatives (/s/,/t/,/ʃ/) had a lower variance (σ^2 =0.003) than, for example, /p/ (0.017). The authors commented that the MDN output demonstrated that the variance for a phone was lower when the articulator could be said to exert a strong influence on a phone. Using this as a rule of thumb, visually inspecting the point clouds for a low bandwidth across all articulators and phones indicated that a higher bandwidth would be required. Incrementally altering this figure until the number of clusters returned per phone was in line with the visual distribution of the articulatory points. A bandwidth of σ=0.15 was settled on after rigorous analysis of each phone and articulator, along with guidance from past studies of *criticalness* by Richmond *et al* (2003) and Singampalli & Jackson (2007) that indicated what articulators exert the greatest influence on each phone.

Summary of Method

To summarise, inversion is performed by a *k*D-tree being built around LSF acoustic data from a database of paired, parametrised and normalised acoustic-articulatory data. A nearest neighbour search is then performed to extract 1500 acoustically similar acoustic frames to every frame in the data-set. The returned frame indices correspond to articulatory frames, which can be plotted and clustered in the articulatory domain to reveal whether certain sounds can be produced by multiple articulatory configurations. A frame that is non-unique will be defined in this study as exhibiting more than 1 cluster

Chapter 4. Experiments

In this chapter, first I will describe the data-set used, then I will proceed to describe the experiments and analyse the results obtained.

Articulatory & Acoustic Data

The acoustic-articulatory data set I will be using was provided by the University of Edinburgh, and is called *mngu0*. I will be using a subset of this corpus, day one, which was recorded at Ludwig-Maximilians-Universitat Muchen using a *Cartsens AG500* electromagnetic articulograph. The data set is comprised of 1263 utterances spoken by a male RP speaker, which were derived from prompts generated from newspapers and a text-selection algorithm. This algorithm is used primarily in the speech synthesis system *Festival* and aims to select text that will be phonetically diverse so as to maximise coverage of as many phonetic contexts as possible. (Richmond (2011)[2]).

A total of 6 EMA sensor coils were attached to the speaker's articulators, one for each of the following; tongue dorsum (T3), tongue blade (T2), tongue tip (T1), lower incisor (JAW), upper lip (UL), and lower lip (LL). A total of 12 channels were then obtained, sampled at 200Hz, one for each x and y coordinate per sensor coil on the midsagittal plane. On the second day of recording, a sensor coil was attached to the velum, but is not present in the data set that I will be using. This data has then been z-score normalised, which allows a better comparison of the range of movement of each articulator. This is important, as, for example, a tongue-tip has a greater range of movement than, say, the lower lip. This normalisation is achieved by subtracting the global mean from each channel and dividing by 4 times the channel's global standard deviation. This preprocessing of the data also allows further experiments carried out on this dataset to be accurately compared.

The corresponding acoustic data was recorded using a Sennheiser MK50 hypercardioid microphone, and the acoustic waveforms were saved to wavefiles (RIFF). Parametrisation of

2 This data-set was provided by Korin Richmond & Steve Renals

these waveforms is then required to obtain the spectral envelope of the speech signal.

STRAIGHT analysis (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram) is a method of using the source F0 information from a waveform for extended pitch synchronous spectral analysis to give a smoothed representation of the spectral envelope (Banno *et al* (2007)). The resulting Line Spectral Frequencies (LSF) are excellent estimates of speech parameters. In estimating the formants from a speech signal, you can extract these formants through inverse filtering, calculate the remaining source information and then use that to remove F0 information from the signal leaving a clear representation of the spectral envelope. Line Spectral Frequencies are a popular way of representing spectral envelope information, such as linear prediction coefficients (LPC), as LSFs have a smaller sensitivity to quantization noise. This is particularly useful when transmitting the spectral information over a channel. The spectral envelope of the acoustic data here was calculated using a 5msec frame-shift rate so that it matches the 200Hz sample rate of the articulatory data, silent frames were removed and the resulting LSFs were also z-score normalised. (Richmond (2011)).

Experiment 1

To reiterate, a study by Qin and Carreira-Perpiñán (2007) showed that on an acoustic-articulatory data-set derived by X-Ray microbeam cinematography, approximately 5% of articulatory coordinates, associated with a given number of acoustic vectors closest to any given reference vector, yielded a multi-modal distribution. Their method utilised search strategy of Ikatura-distance for acoustic similarity, a mode-finding mean-shift clustering algorithm, and was initialised for each of the 45,000+ acoustic-articulatory vector pairs. Their acoustic data comprised of LPCs sampled at 147Hz, which matched the sample rate of the articulatory data. Articulatory point clouds that yielded a single mode, or one cluster, were deemed uni-modal, whilst clouds that produce multiple modes or clusters are considered multi-modal, and therefore non-unique.

Similarly, I aim to quantify the proportion of frames that show non-uniqueness, when 1500

acoustically close vectors to the reference frame are plotted in articulatory space. As the *mngu0* data-set consists of 799,159 frames, I have partitioned the amount of frames to be analysed. Though the GBMS algorithm is an accelerated clustering algorithm, the sequence of performing the nearest neighbour search for every multi-dimensional frame (the LSF data is of 40 plus gain dimension per frame) and for each of the 6 articulators is still a computationally expensive process. Partitioning the data into smaller groups for simultaneous processing therefore allows a faster and more convenient solution.

For these experiments, the acoustic and articulatory data is from day one of the *mngu0* EMA corpus, consisting of line spectrum frequencies (LSF) sampled at 200Hz using STRAIGHT analysis to match the corresponding articulatory data. Unlike in Qin and Carreira-Perpiñán's study, *day one* of the *mngu0* corpus does not have articulatory data for the movement of the velum. Furthermore, the number of frames is far greater in this data-set than for their speaker (*jw11,* male, 90 utterances), and the subject speaks with a unspecified American English accent. In American English, depending on the regional accent, an alveolar approximant /ɹ/ can be *retroflexed* or *bunched,* which has been show to contribute to articulatory non-uniqueness (Qin & Carreira-Perpiñán (2010)), particularly for the tongue blade and dorsum over a sequence of frames. This isn't an attribute of RP English, and so the number of clusters would in theory be lower for our speaker for the alveolar approximant consonant /ɹ/. Depending on the number of frames of /ɹ/ in the data-set, this and other phonetic differences between the speakers could determine the level of non-uniqueness in the data-set.

Results

The results for this task showed that the number of frames that yielded clouds with multiple clusters was higher in the *mngu0* data-set than in the XRDB data-set used in a previous study (Qin & Carreira-Perpiñán (2007)). Averaging over the 6 articulators, 19.48% of the 799,159 frames analysed yielded multiple clusters, and could therefore be considered to exhibit non-uniqueness. Predominantly, the data-set was uni-modal in nature, with 80.52% of the frames analysed yielding a single cluster. As is evident in Figure 8, the level of non-uniqueness (NU) differed for each of the different articulators, from 7.5% (T2) to 41% (UL).

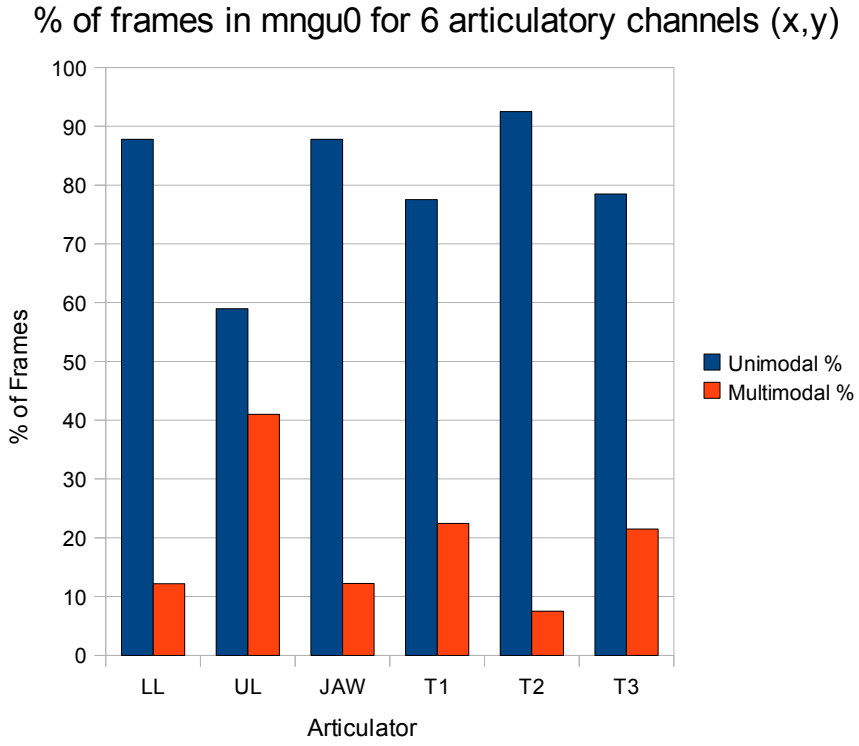


Figure 8 - The percentage of frames in the *mngu0* data-set that showed non-uniqueness (determined by the number of clusters) for 6 articulators.

The lower-lip (LL = 12.19% NU), lower-incisor (JAW = 12.2% NU) and the tongue-blade (T2) exhibit a strongly unique inversion mapping, with the tongue blade in particular showing only 7.51% non-uniqueness. The tongue-tip (T1 = 22.46% NU) and the tongue-dorsum (T3 = 21.49%) exhibited moderately high levels of non-uniqueness, in line with observations made by Neiberg et al (2009) regarding the proportion of non-uniqueness in a given phone for which these two articulators are responsible for. Finally, the articulator that shows the highest number of non-unique frames in the data-set is the upper-lip (UL), with 41.01% of the frames analysed presenting 2 or more clusters.

Articulator	Unique	Non-Unique	Total Frames	Unique %	Non-Unique %
LL	701742	97417	799159	***87.81***	***12.19***
UL	471424	327735	799159	***58.99***	***41.01***
JAW	701662	97497	799159	***87.8***	***12.2***
T1	619668	179491	799159	***77.54***	***22.46***
T2	739142	60017	799159	***92.49***	***7.51***
T3	627420	171739	799159	***78.51***	***21.49***

Table 1 - The number of frames in the data-set analysed for each articulator and the % of unique and non-unique frames.

In this broad empirical test, the phone identification of each frame analysed is not considered. Different phones will present different levels of non-uniqueness, and the context a phone is from can contribute to this. Transitional frames between fully-realised phones are included, so the articulators will consistently be moving towards the next phone. This could contribute to non-uniqueness, but an analysis of these trajectories is not included here. The reason for the large number of frames that have a non-unique mapping for the upper-lip is therefore unclear, and disambiguating phones in the 327,735 non-unique frames would be impractical and misleading. If a transitional frame between two fully realised phones is used to find the nearest 1500 acoustically similar frames, the returned frames may not be from another example of the same phone transition, and a mixing-in of other articulatory configurations from different phone transitions may occur. Moreover, diphthongs (/ɪə/, /aɪ/) and triphthongs (/ɛɪə/), cannot be represented by a single frame, but they are included in this broad empirical analysis, nevertheless.

However, the aim of this particular experiment was to quantify the level of non-unique frames in the data-set regardless of phone type. Overall, it has been found that the methodology used here has highlighted that 19.48% of the *day one* subset of the *mngu0* EMA data-set demonstrates a non-unique articulatory distribution when clustered with 1500 acoustically close vectors

Experiment 2

This experiment was concerned with identifying which sounds can be produced by different vocal-tract shapes, and for which articulators.

To study the non-uniqueness for representative frames of each phone in the data-set, I used the frame-aligned labelling available, whilst omitting diphthongs and triphthongs. Furthermore, I omitted analysis of the mid-central vowel /ə/, as phonetically and phonemically the vowel is seldom precisely identified, with any unstressed central vowel typically being labelled as such.

I ran the same methodology as in Experiment 1, but this time run on a manually constructed matrix of the representative frames, so that each phone could be clearly seen to exhibit either uniqueness or non-uniqueness, as defined by this study. Using this method for each articulator will also allow analysis of the level to which an articulator contributes to non-uniqueness, and for which phones.

Results

Overall, the selected phone data-set, comprising of 34 different phones, presented an average of 26.47% non-uniqueness across all 6 articulators. The upper-lip, in line with the previous experiment, showed the greatest level of non-uniqueness, with 50% the phones tested producing multiple clusters. Conversely, the tongue-dorsum and the lower-lip showed the lowest level of non-uniqueness, with up to 88.24% presenting a uni-modal distribution. However, as will be discussed, the quantity of actual non-uniqueness across all phones is not as clear as first indicated.

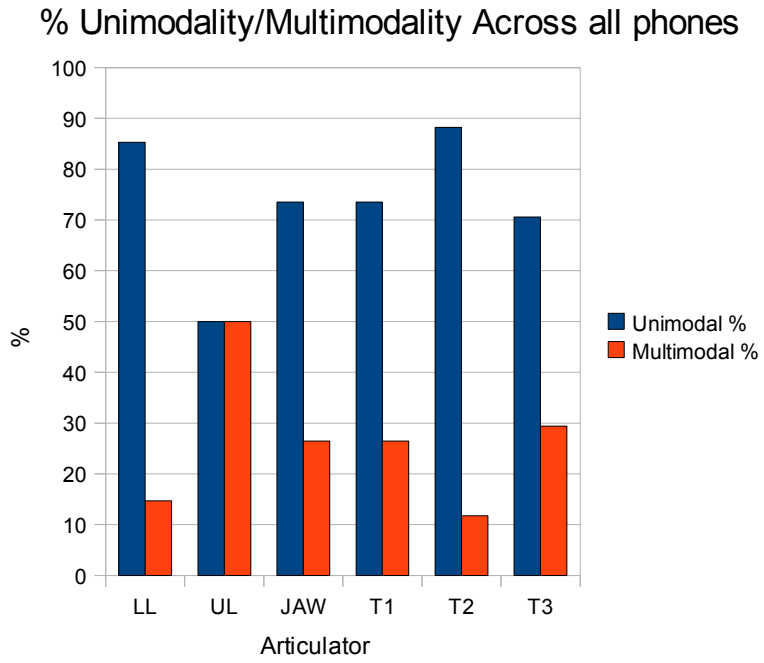


Figure 9 – The percentage of frames in the pre-defined phone test set that showed non-uniqueness (determined by the number of clusters) for 6 articulators.

For the purpose of analysis, the phones will be organised by manner of articulation. I will discussed pertinent discoveries for each phone category.

Fricatives

The results of the inversion and clustering for occurring fricatives in the *mngu0* data-set (/f, v, θ, ð, s, z, ʃ, ʒ/) demonstrate that the lips (UL & LL) contribute to a large proportion of non-uniqueness in the acoustic-articulatory inversion, but that for a few fricatives that is the result of outlying data. For example, while the GBMS algorithm segments voiceless dental fricative /θ/ into 2 clusters, visualising the clustered plot itself shows that all but one of the 1500 data-points is confined to a single moderately dense cluster, as shown in Figure 10. Using the frame-aligned labels, the frame in question is found to belong to /f/. This could either be because of inaccurate automatic labelling of the corpus, or more likely it is due to the acoustic similarity of the query frame of /θ/ and the outlying /f/ frame data. Furthermore,

the very same /θ/ acoustic frames, when inverted, in the tongue-tip articulator channel do yield a demonstrably non-unique distribution (Figure 11). The actual reference frame is located in the blue cluster of Figure 11, and the multi-modal distribution evident is only apparent for the y coordinate of the T1 coil, as shown in the histogram. In Figure 12, the single cluster evident for /f/ corresponds to the unidentified cluster in Figure 11, meaning that the acoustic frames of /f/ were acoustically similar to /θ/ and were treated as such in the inversion.

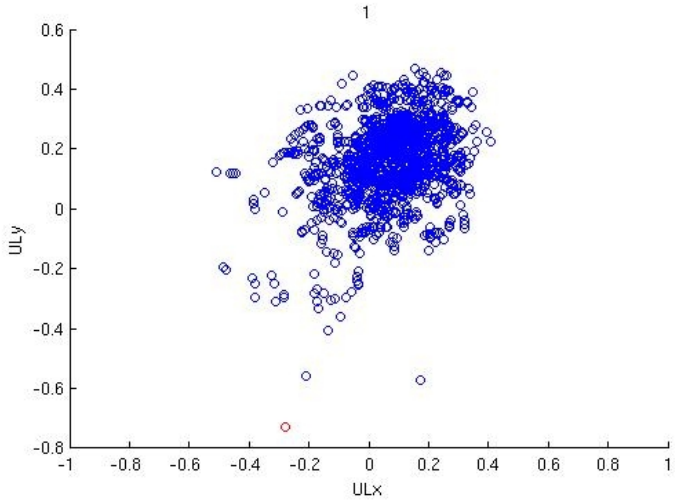
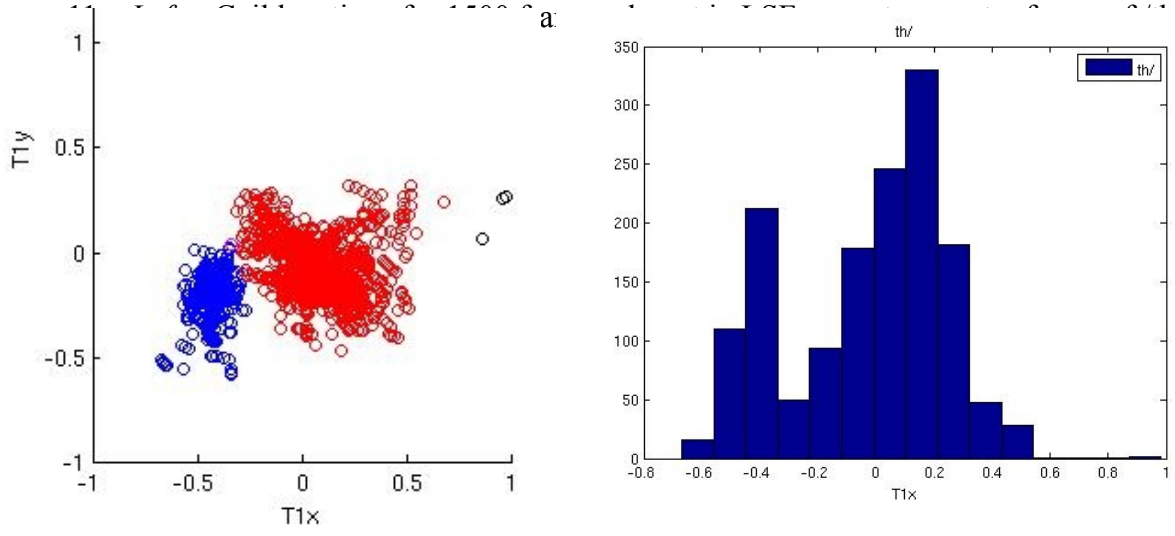


Figure 10 - Coil locations for 1500 frames closest in LSF space to a centre frame of /th/ for the upper-lip (UL). Note that the clustering algorithm has found a single outlying vector marked in red at position x=-0.3/y=-0.75.

a



Right – A histogram of the T1y coordinate, showing a non-unique inversion mapping

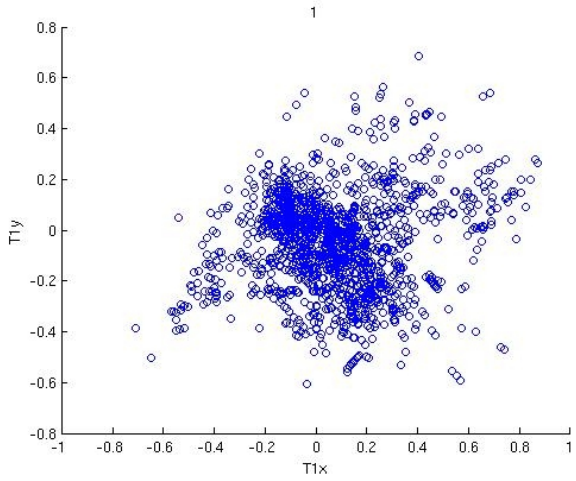


Figure 12 - Coil locations for 1500 frames closest in LSF space to a centre frame of /f/ for the tongue-tip (T1).

In Figure 13, the distribution of the 1500 nearest neighbours for labio-dental fricative /f/ before and after clustering demonstrates that the distribution of the UL articulatory points appears to be non-unique for the y coordinate. However, a histogram of UL_y for all /f/ labelled phones in the data-set (Figure 14) shows no large distribution around -0.4 (UL_y). Further analysis of the phonemically similar phone /v/ through clustering and histograms (Figure 15), though one would not normally perceptually confuse a voiced and voiceless sound, showed an articulatory distribution similar to that of /f/, with mass for UL_y from 0 to 0.2 (Figure 15). A mixing of acoustically similar frames of /f/ and /v/, whilst explaining their similar distributions and subsequent clusters, cannot therefore explain the inclusion of articulatory points around -0.5 for the upper-lip's y coordinate. In this respect, though non-uniqueness is clear in the clustering, other unknown acoustically similar phones must therefore be acoustically close (parametrically at least) to the query frames of /f/ and /v/. Therefore, a conclusion cannot be made as to whether or not these clusters actually indicate true non-uniqueness.

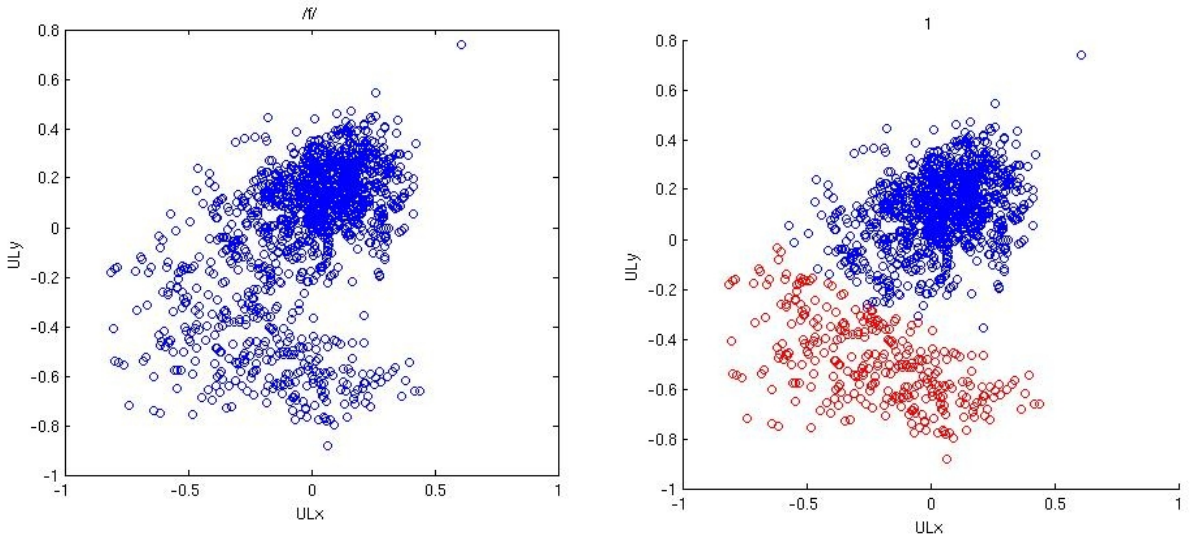


Figure 13 - *Left* - Coil locations for 1500 frames closest in LSF space to a centre frame of /f/ *Right* – GBMS clustering of the data after 1 iteration.

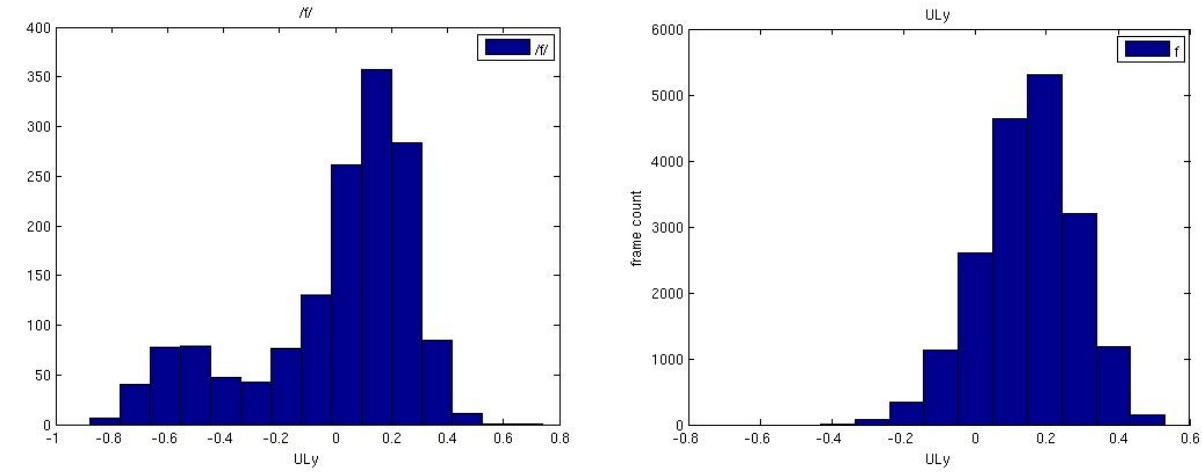


Figure 14 – *Left* - A histogram of the frames for 1500 frames closest in LSF space to a centre frame of /f/ for the UL_y coordinate.

Right - A histogram of all frames in *mngu0* of ‘f’ for the UL_y coordinate.

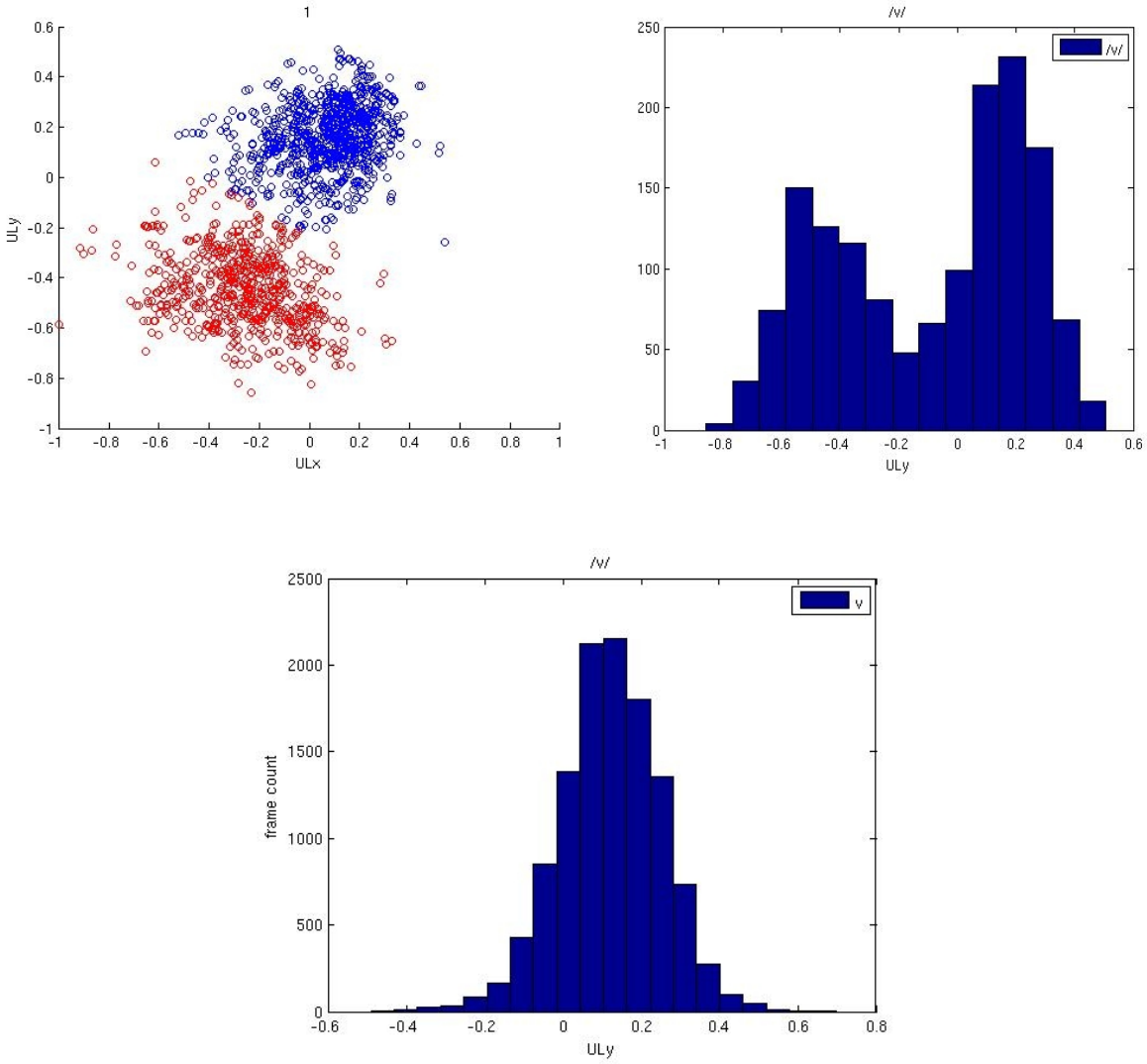


Figure 15. Top left - Coil locations for 1500 frames closest in LSF space to a centre frame of /v/ for the upper-lip (UL) – Top Right - A histogram of the 1500 frames closest in LSF space to a centre frame of /v/ for the UL_y coordinate. - Bottom - A histogram of all frames in *mngu0* of 'v' for the UL_y coordinate.

To demonstrate a uni-modal clustering for fricative, we take the example of /s/ and /z/. Both voiced and voiceless alveolar fricatives /s/ and /z/ are shown to have a unique mapping across all articulators, within the confines of the GBMS clustering, as shown in Figure 2. Previous evidence of phone variance using training MLP models (Richmond, King & Taylor (2003)) indicated that these phones would show low variance, especially for an articulator that is critical to the production of the sound, such as the tongue-tip. Figure 16 shows a single articulatory cluster for the 1500 acoustically similar frames for /s/ & /z/ for the tongue-tip,

converging to a single centroid, and visualising the actual coil locations confirms this.

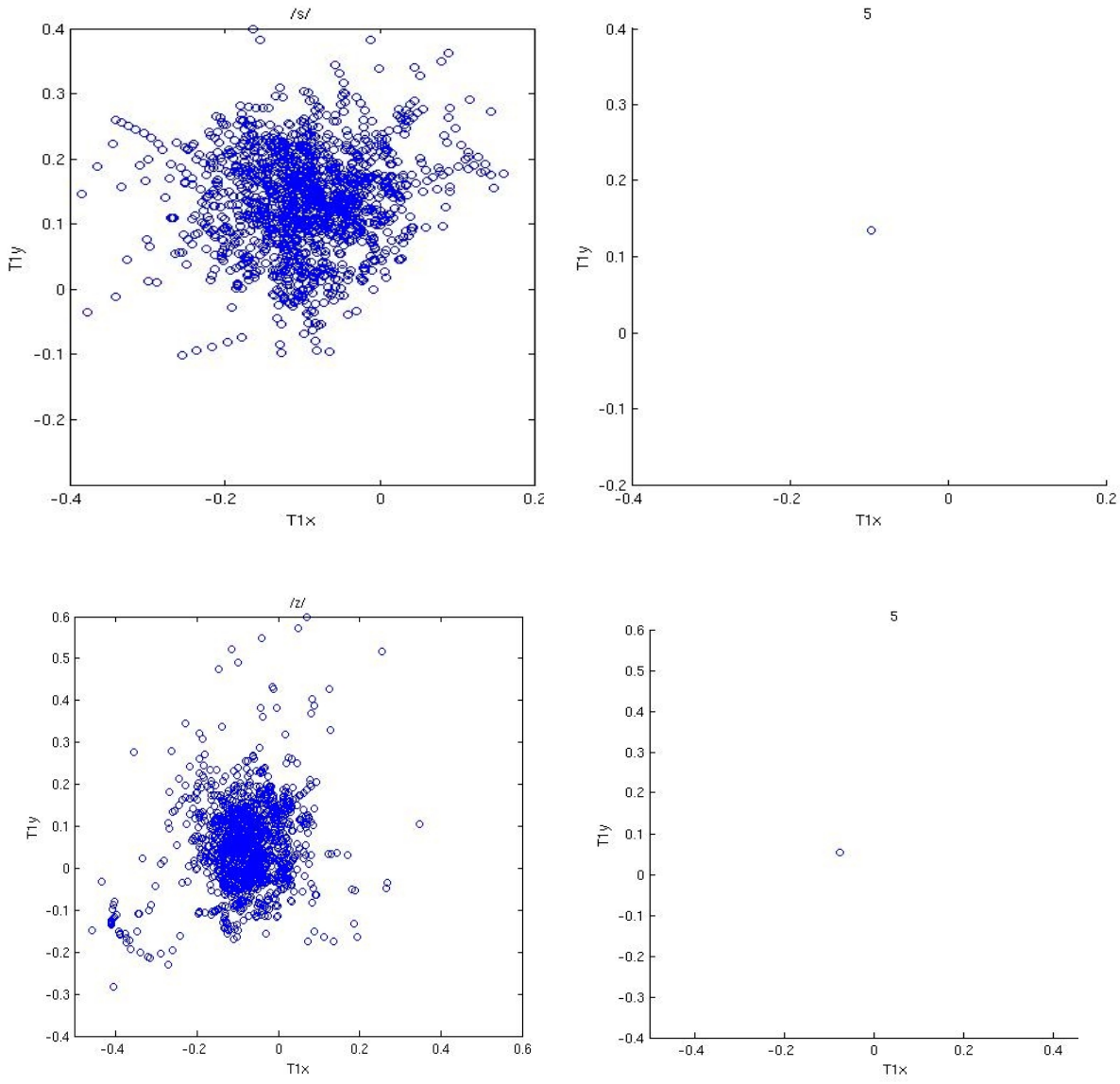


Figure 16 - *Top & Bottom Left* - Coil locations for 1500 frames closest in LSF space to a centre frame of /s/ & /z/, respectively.*Top & Bottom Right* – GBMS clustering of the /s/ & /z/ frames after 5 iterations.

Plosives

As shown in Figure (17), all of the plosives examined exhibited non-uniqueness across at least two articulators. The upper-lip contributed to 23.5% of all instances of plosive non-uniqueness, with the tongue-tip contributing up to 29.4%. Conversely, the tongue-dorsum and blade showed predominantly unique plosive distribution, with only voiceless velar plosive /k/ and voiced velar plosive /g/ showing non-uniqueness on the tongue-dorsum and blade, respectively.

Bilabial plosives /p/ and /b/ yield non-unique clouds in articulatory space for the upper-lip, lower-incisor and the tongue-dorsum, with voiced /b/ also exhibiting a non-unique distribution for the lower-lip. As the lips are essential for the production of the eponymous bilabial stop, this result is curious. Scatter plotting the non-clustered 1500 articulatory frames closest in LSF space to the centre frame of phone /b/ does indeed show a distribution that indicates non-uniqueness, as shown in Figure 17.

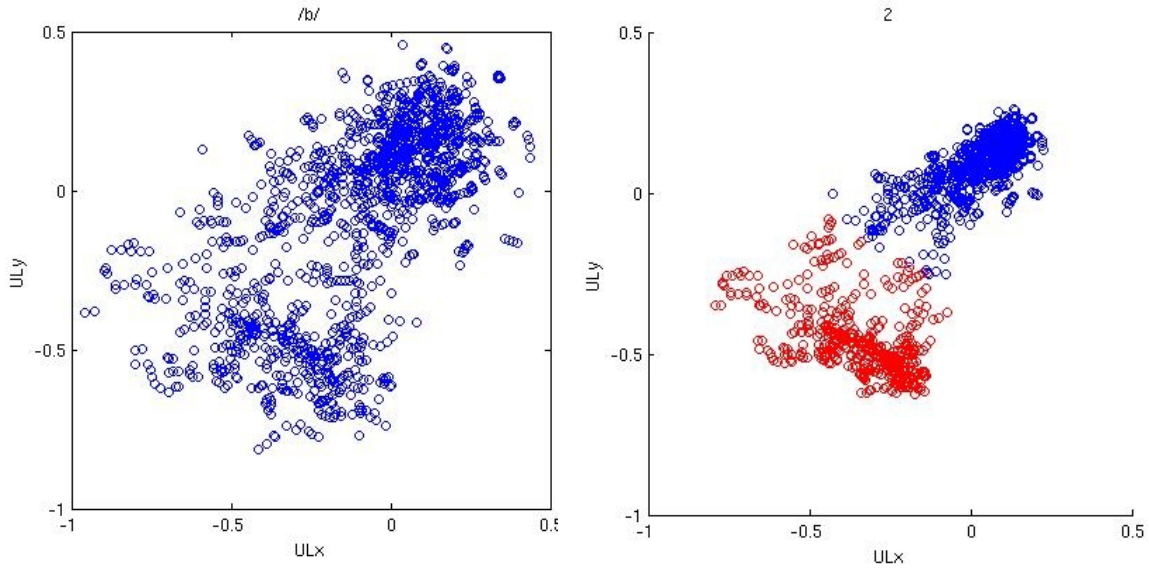


Figure 17 - L*eft* – Coil locations for 1500 frames closest in LSF space to a centre frame of /b/. *Right* – GBMS clustering of the data after 2 iterations. Two clusters have been identified, and points have begun to converge to a central point in each cluster.

To investigate further, I made a histogram showing the distribution of all frames of 'b' for

UL_y (Figure 18i), showing density at 0.2. In Figure 18ii, the 1500 acoustically similar frames of /b/ in UL_y shows a large distribution of frames at both -0.5 and 0.2. Similarly, in Figure 18iii, a histogram of 'p' shows a high UL_y distribution at 0.2, whereas a scatter plot of the 1500 acoustically similar frames in articulatory space (Figure 18iv) shows UL_y distribution at both -0.5 and 0.2. As with the similar distributions of /f/ & /v/, what appears to be occurring is a mixing of acoustically similar frames of /p/ and /b/ when each phone is inverted into articulatory space for this particular articulator through the nearest neighbour search.

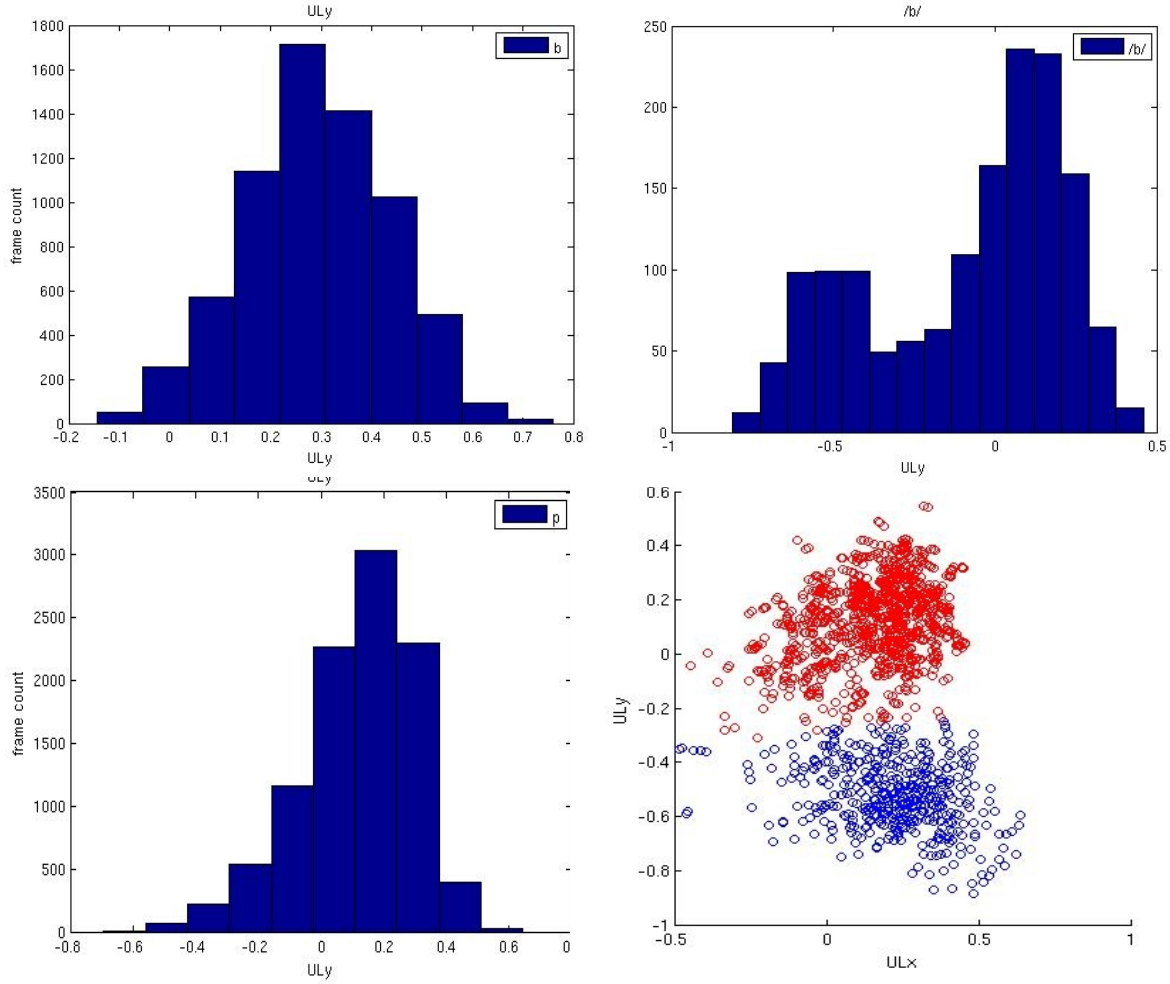


Figure 18 - *Top left i)* Histogram of all frames of 'b' in UL_y – *Top right ii)* A histogram of the 1500 frames closest in LSF space to a centre frame of /b/ in UL_y

Bottom left iii) - Histogram of all frames of 'p' in UL_y – *Botttom right iv)* – Clustered Coil locations for 1500 frames closest in LSF space to a centre frame of /p/

The parametrisation of the acoustics appears to be the cause of this, rather than the /p/ or /b/ actually being produced in multiple articulatory configurations. One possibility could be that the LSF data is not adequately modelling unvoiced frames, such as for /p/, which would explain the mixing in of /p/ and /b/ frames, as well as the previous example of voiceless labio-dental fricative /f/ frames mixing in with voiced /v/. This appears to be limited to the upper-lip (x,y), though non-uniqueness has also been found for non-critical articulators for /p/. The lower-incisor (JAW), tongue-tip (T1) and tongue-dorsum (T3) all show multiple clusters, though not to the same degree of density as for the upper-lip. In Figure 19, for example, the two clusters are evident, but on examining a histogram of all frames of ‘p’ in the corpus for T3_x, there is only a moderate distribution of frames at coordinate -0.4. Likewise, there is a only a moderate to small distribution of frames at coordinate 0.4 for T3_y. In Figure *, the scatter plot with histograms on each axis shows that there is a very small local minima distribution for the acoustically close frames at 0.4 for the T3_y articulator. Furthermore, a histogram for all frames of ‘b’ does indicate a higher distribution at 0.4 for the T3_y articulator. The GBMS algorithm has identified this as a separate cluster, though it is not clear whether or not this is again the result of mixing acoustically close /p/ and /b/ frames due to unvoiced frames being insufficiently represented by LSF coefficients.

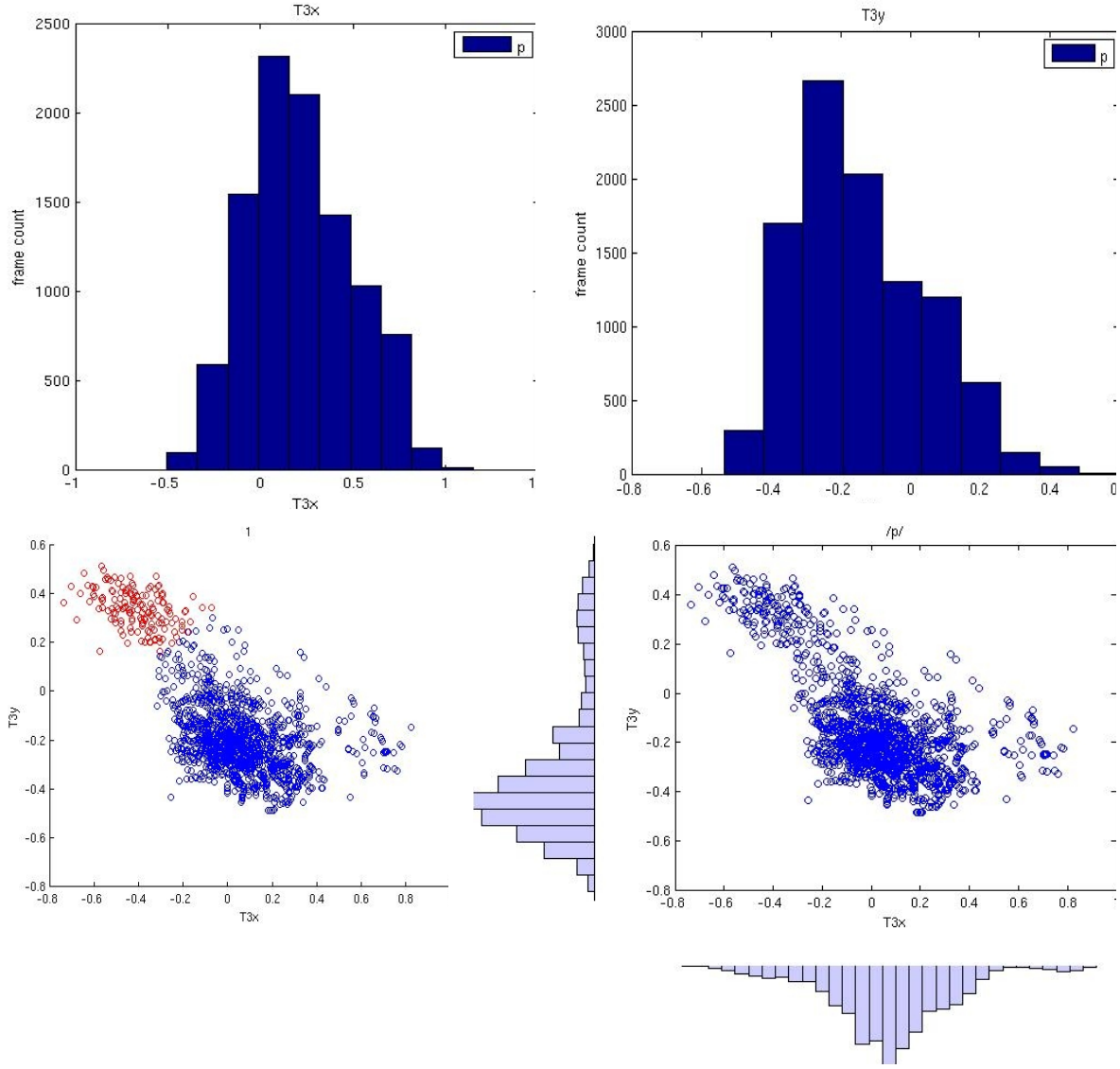


Figure 19 - *Top* - Histograms of the 1500 frames closest in LSF space to a centre frame of /b/ (*left)* and /p/ (*right)* for the UL_y coordinate. - *Bottom* - Coil locations for 1500 frames closest in LSF space to a centre frame of /p/ (*left)* and a combined scatter plot and histogram for the coil locations of 1500 frames closest in LSF space to a centre frame of /p/ in UL_y (*right*).

A final example of curious non-uniqueness for plosives is for the voiced velar plosive /g/ and

the voiceless plosive /k/. Both of these phones exhibit non-uniqueness on the upper-lip (2 clusters identified for each phone), but also are partitioned into two clusters for the tongue-dorsum; their critical articulator. Scatter plots of both /g/ and /k/ (Figure 20 *Top*) both show a clear multi-modal distribution of the 1500 articulatory points. The T3 coordinates that record high densities of frames for both /g/ and /k/ are at approximately positions T3_x=-0.6/T3_y=0.6 and T3_x=0.2-0.5/T3_y=0.2. The actual query frame for both phones in located in the top cluster (x=-0.6/y=0.6), where the visible boundary line corresponds to the soft-palate. What becomes apparent when taking all frames of 'k' and 'g' in the data-set and plotting a histogram for T3(x,y) is that there is no distribution of 'k' or 'g' frames below coordinate 0 for either T3_y channel (Figure 20 *Bottom)*. This is indicates that most or all of the articulatory points below that point should belong to a different phone, and the clustering of these points by the GBMS algorithm indicates the inclusion of other phones.

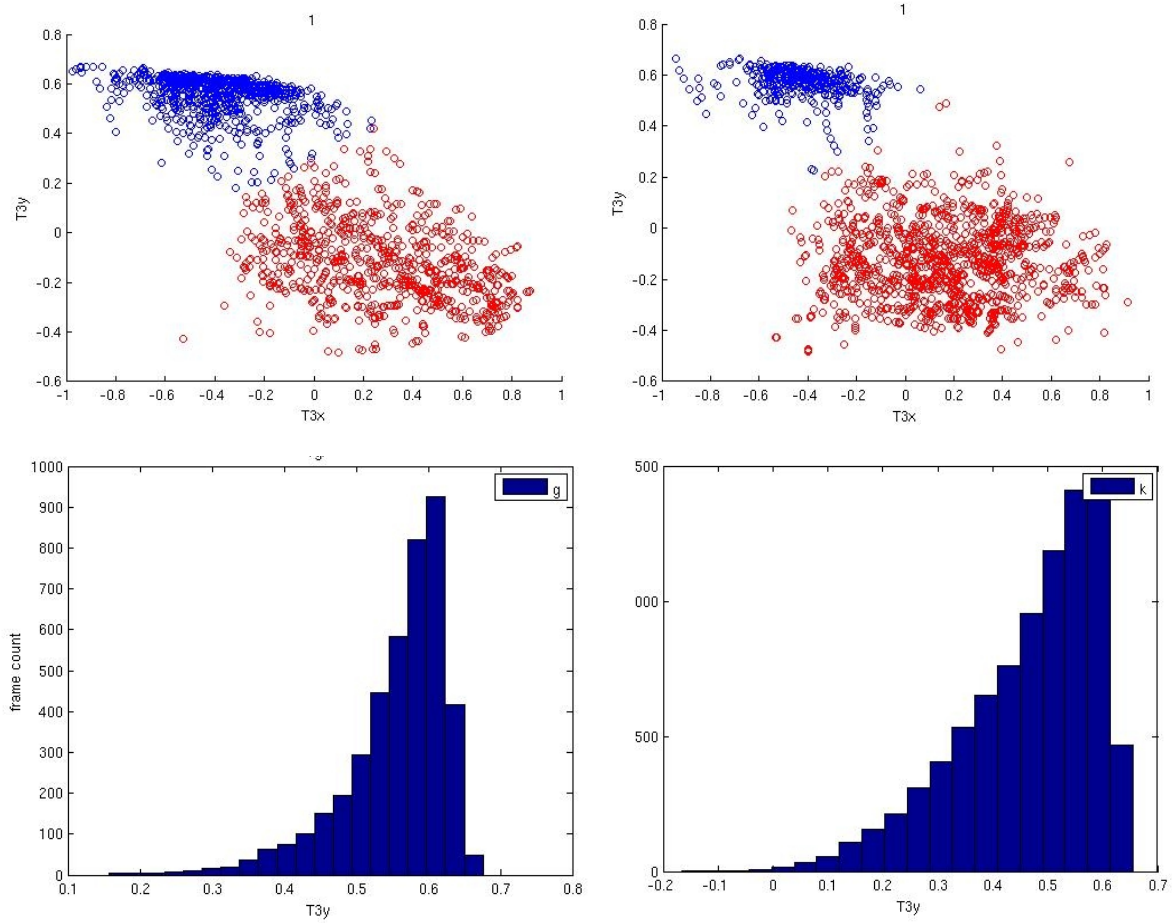


Figure 20 - *Top* - T3 coil locations for 1500 frames closest in acoustic space to a centre

frame. *Left* - /g/. *Right* - /k/. *Bottom* – Histograms for all frames of ‘g’ (*left)* and ‘k’ (*right*) in the data-set.

In investigating the identity of these phones, I found that included in the 1500 acoustically close phones to /g/ were the following; /ʃ, p, b, t, k/. One of the most common frames that was close to /g/, as labelled by the Combilex LTS system, was ‘k_cl’, which is the closure part of the velar plosive /k/. This accounted for 20% of the returned nearest neighbours. However, oddly enough, 24% of the frames 1500 that were acoustically similar to /g/ were comprised of ‘p’, whilst 50% of similar frames to /k/ were comprised of ‘f’. As I doubt that /p/ and /g/, nor /f/ and /k/, are actually acoustically equivalent, this must be another instance where, due to the parametrisation of the acoustic data, the sounds are acoustically similar. Therefore a seemingly non-unique inversion from the acoustic to the articulatory domain such as this one needs to be interpreted carefully, as is evident by one cluster containing entirely different phones.

Nasals

In total, all five types of nasal phones that occurred in the data-set exhibited non-uniqueness over at least one articulator. Of the total non-uniqueness shown by the phone-set, 17.86% came from these nasal phones. The syllabic alveolar nasal stop ‘n!’ showed non-uniqueness for the upper-lip and all three tongue coils (T1, T2, T3). However, inspection of the 2 clusters for T1 shows that all but 2 articulatory frames fall into a single uni-modal distribution. As the tongue tip is necessary for production of the syllabic alveolar nasal stop, this was as to be expected.

Figure 21 (top left) shows 2 clusters in the articulatory point cloud for 1500 frames of ‘n!’ after two iterations of the GBMS clustering algorithm, while Figure * demonstrates clear multi-modality in the distribution of ‘n!’ for the UL_y coil. Of the phone labels that make up this plot, only 6.8% of the frames are for ‘n!’, whilst 37.2% are derived from the alveolar nasal /n/, 26.46% are derived from the bilabial nasal /m/, and 13.8% are derived from the velar nasal /ŋ/. The remaining 15.4% consists of numerous frames of the following phones; /t,

d, b, v, z/. A histogram for all frames of ‘n!’ in the data-set shows only one distribution peak (0.2) for UL_y coil locations, whereas a histogram of all ‘m’ frames for UL_y coil locations shows a distribution peak at -0.5. A combination of 26.46% of 1500 acoustically similar ‘n!’ frames corresponding to ‘m’ and the total distribution of ‘m’ frames in the data-set peaking where the blue cluster of frames is situated is a good indication that non-uniqueness for this particular phone is dependent on bilabial nasal /m/ being mixed in. Once again, though it appears as if there is a non-unique inversion mapping between the acoustic data and the articulatory configurations

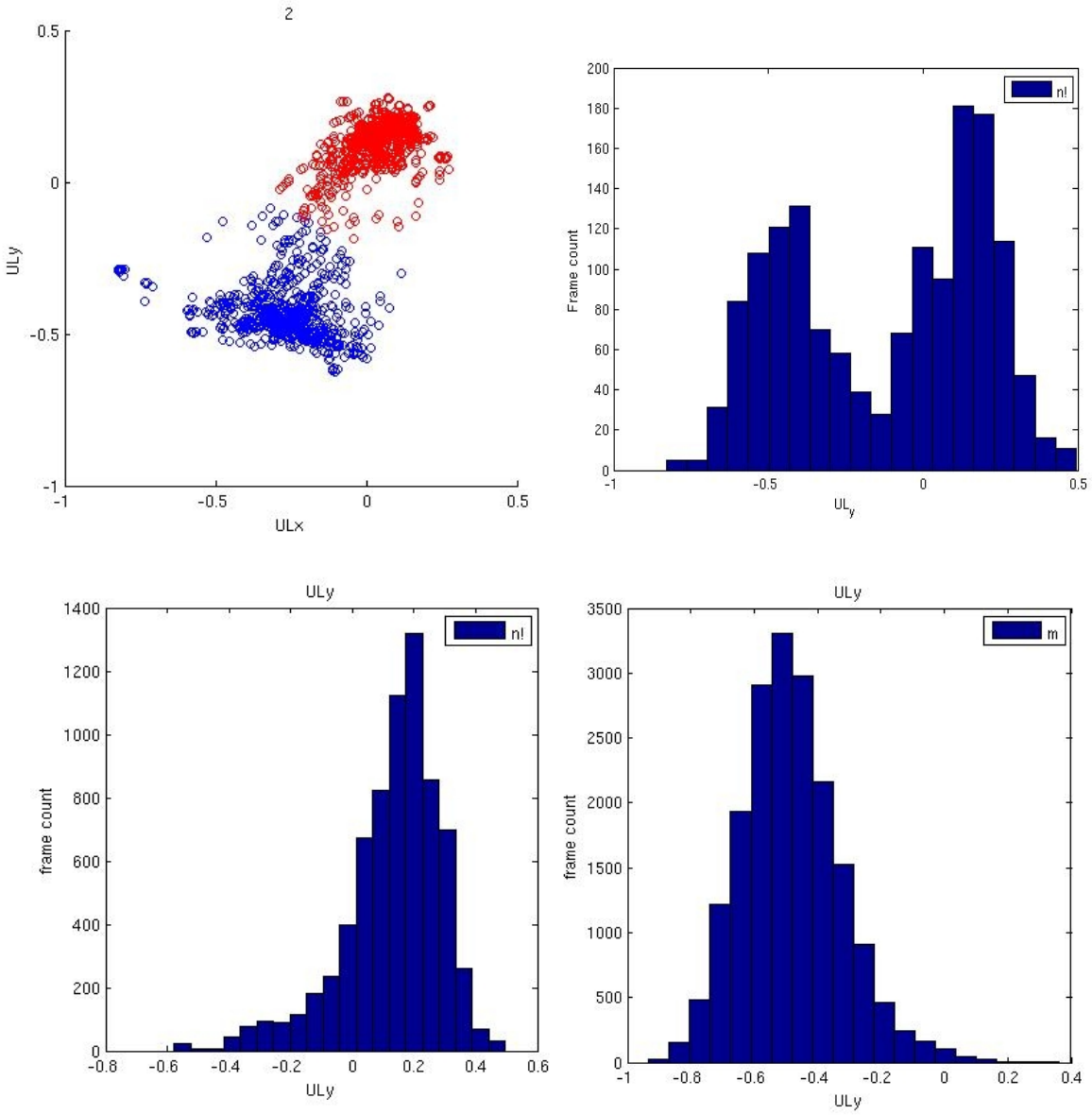


Figure 21 - *Top* - UL coil locations for 1500 frames closest in acoustic space to syllabic alveolar nasal stop ‘n!’ (*left*) and a histogram of the 1500 frames closest in LSF space to a centre frame of ‘n!’ in UL_y. *Bottom* – Histograms for all frames of ‘n’ (*left)* and ‘m’ (*right*) in the data-set.

Approximants

In Qin & Carreira-Perpiñán (2007), the authors found that approximants, specifically American-English alveolar approximant /ɹ/, lateral approximant /l/, and labio-velar approximant /w/, exhibited multi-modality. Here, the alveolar approximant /ɹ/ produced, when inverted, a single cluster for each of the tongue coils (T1, T2, T3). According to previous studies (Qin & Carriera, 2009), a trajectory of each tongue coil for American English /ɹ / showed that multiple tongue-blade shapes occurred (retroflexed and bunched), and but that the modality of the tongue tip and dorsum remained singular. A histogram of this phone’s distribution in *mngu0* for the tongue-blade for the 1500 frames closest acoustically to /ɹ / in articulatory space (Figure 22) does shows a moderately wide spread of articulatory points for the T2_y channel, but the distribution is normal. As the male RP speaker in *day one* of *mngu0* does not have the same accent as the speaker used in Qin & Carriera-Perpinan’s study (2007), in that the RP is a predominantly non-rhotic accent, the variety of different tongue shapes evident for the alveolar approximant /ɹ/ in the articulatory domain will differ. In this respect, the lack of non-uniqueness for /ɹ/ here is not directly comparable to Qin & Carriera-Perpinan’s study.

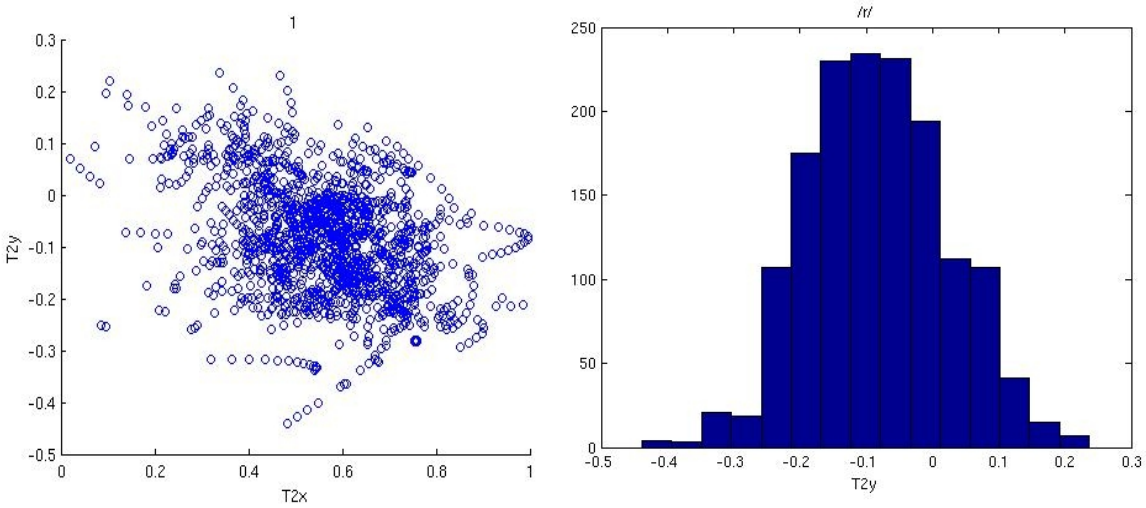


Figure 22 – *Left* - T2 coil locations for 1500 frames closest in acoustic space to alveolar approximant /ɹ/. *Right* - A Histogram of the 1500 frames closest in LSF space to a centre frame of /ɹ/ for the tongue-blade (T2) y coordinate.

The articulatory distribution of lateral alveolar approximant /l/ from the phone-set is uni-modal across the three tongue coils (T1, T2, T3), which is different to what was found in Qin & Carriera-Perpinan's study. Figure 23 (*Top*) shows the single cluster, as identified by GBMS, and the histograms for T2_x and T2_y show a normal distribution for the returned nearest neighbours. This does not, however, necessarily mean that there are not frames from other phones mixed in due to acoustic similarity. A look at the frame identities in the returned data-set shows that, indeed, only 5.6% of the returned frames were drawn from phones labelled 'l' or 'l!' (syllabic lateral stop). Predominantly, vowels such as the mid-central unstressed /ə/, or front close vowels such as /y/ or /i:/, were often deemed acoustically similar to /l/. A histogram of the distribution of all frames of 'l' for T2_x shows that there is a very wide spread of possible T2_x positions in the data-set. The tongue-tip, however, shows an almost exclusively uni-modal distribution (Figure 23 *Bottom*) as it makes contact with the alveolar ridge, which is to be as expected as the variance of a phone is lower for its critical articulator (Richmond, King & Taylor (2003)).

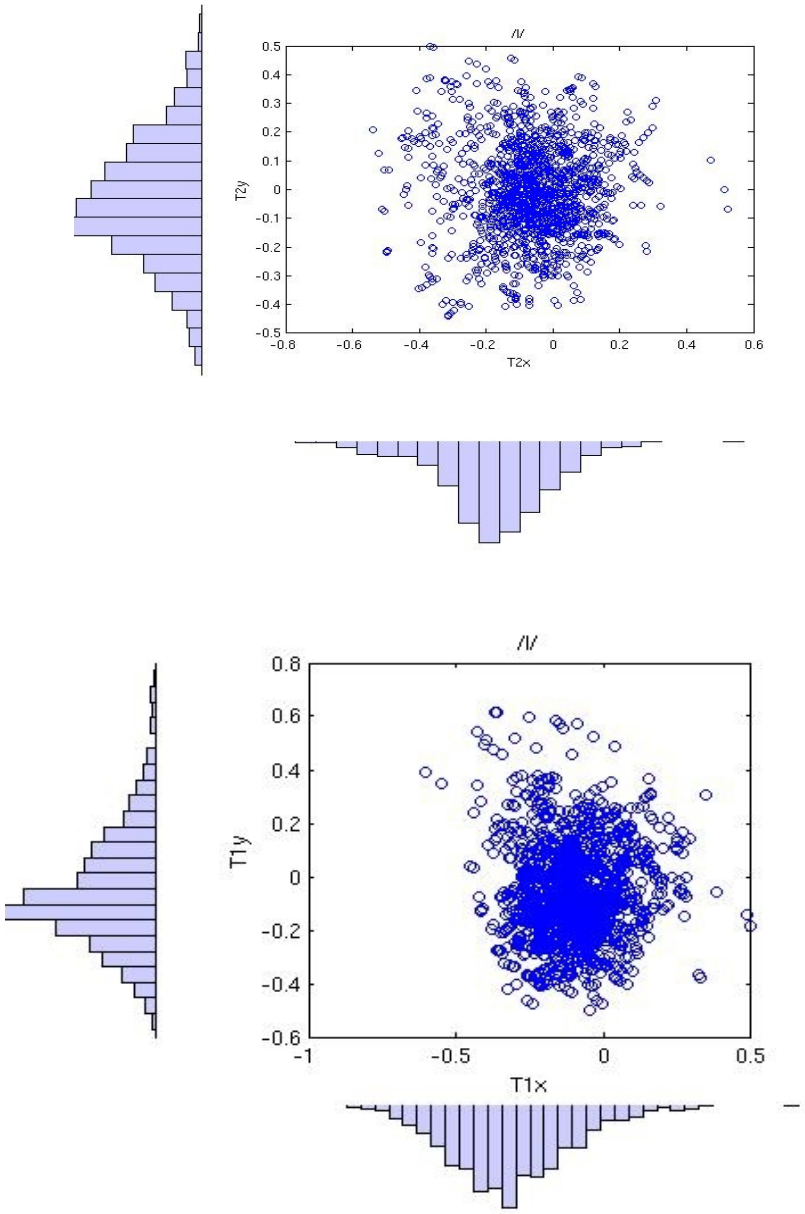
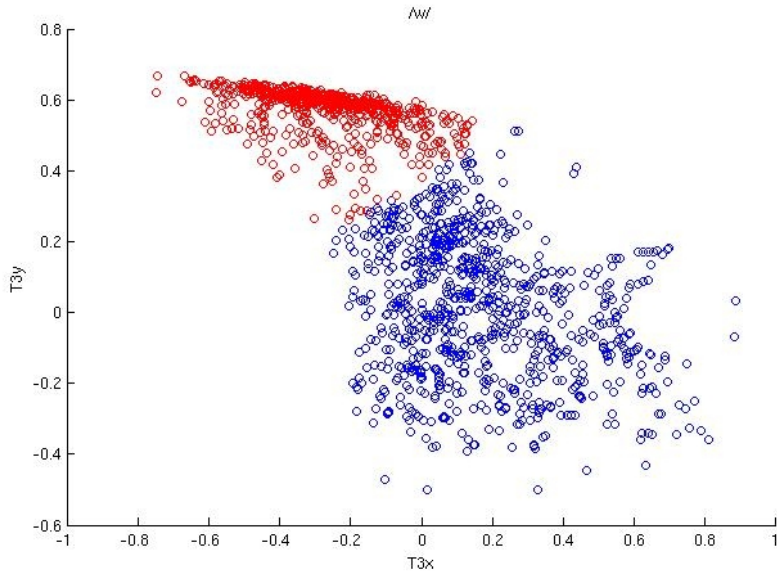


Figure 23 – *Top* - T2 coil locations for 1500 frames closest in acoustic space to lateral approximant /l/ and histograms for T2 x & y. *Bottom* - T1 coil locations for 1500 frames closest in acoustic space to lateral approximant /l/ and histograms for T1 x & y.

Finally, the labio-velar approximant /w/ shows no non-uniqueness for the lips (UL, LL), but shows a similar pattern of non-uniqueness to that of /k/ and /g/, as shown in Figure 24. The GBMS algorithm has clustered the data into 2 distinct clusters, with the actual query frame located at coordinates T3_x=0.14/T3_y=0.26, at the top of the blue cluster. In the production

of the voiced labio-velar approximant, the tongue-dorsum is raised to narrow the vocal-tract and produce a turbulent airflow. The tongue-dorsum does not make contact with the soft-palate, unlike /k/ and /g/ phones. Figure 24 (*bottom left*) demonstrates that in the inversion there is a significant distribution mass at T3_y=0.6 and a lesser sized distribution at T3_y=0.2. As before, a histogram of the distribution of 'w' in the entire data-set for T3_y (Figure 24 *bottom* right) shows only a single peak around position 0.2, which corresponds to the query point. Inspecting the frame-aligned labels for these 1500 frames shows that there is indeed mixing of other velar phone frames, with voiceless velar plosive /k/ in particular contributing 28.93% of the nearest neighbour frames. As with each example that has been found that demonstrates the mixing of acoustically similar phones, it is important to reiterate that this does not necessarily mean that /w/ and /k/ are acoustically equivalent. Rather, that these particular articulatory configurations are able to produce similar acoustic parameters.



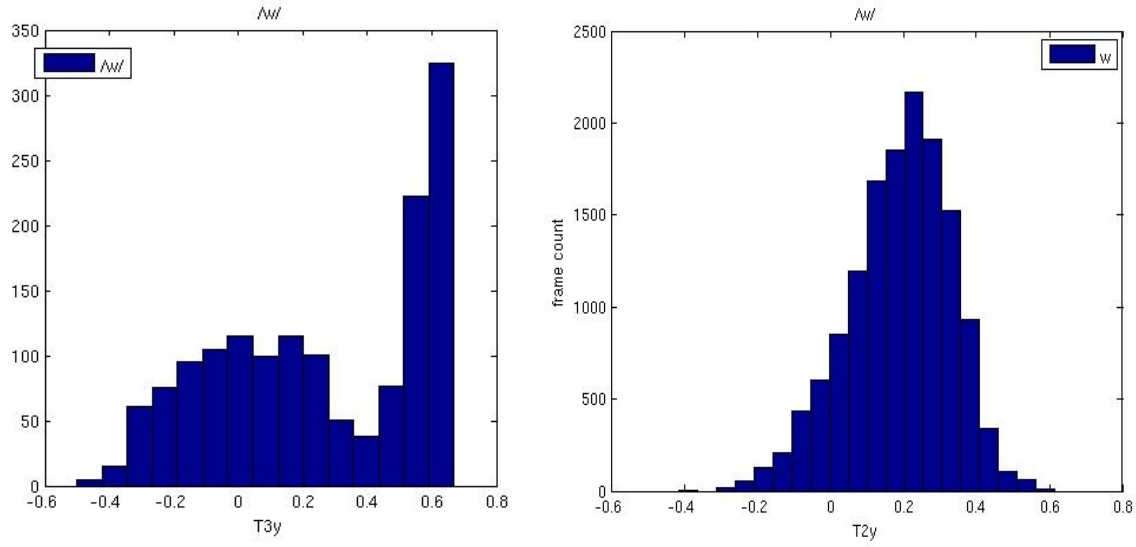


Figure 24 - *Top* - T3 coil locations for 1500 frames closest in acoustic space to labio-velar approximant /w/ - *Bottom* – A Histogram of the 1500 frames closest in LSF space to a centre frame of /w/ for the tongue-dorsum (T3) y coordinate (*left*) and a histogram for all frames of 'w' in the data-set (*right*).

Vowels

Of the vowels in the data-set that were not diphthongs, triphthongs or the mid-central unstressed schwa, the GBMS clustering algorithm found that 87.5% were demonstrably uni-modal in distribution. This is in line with Qin & Carreira-Perpiñán's study (2007), where /æ/, /i/ and /uː/ were shown to be uni-modal, yielding a single cluster. Of the 12.5% of vowels that yielded multiple clusters, such as /æ/ for the upper-lip, an examination of the point clouds shows that they do show a predominantly singular clustering, with a small number of data-points in each of remaining clusters. This can be seen in Figure 25 (*left*), which shows the upper-lip coordinates associated with the 1500 nearest neighbours to /æ/. In disambiguating the acoustically similar frames, I found frames from diphthongs (/aɪ/, /au/), as well as the occasional consonant (/f/,/k/). However, we can conclude that, as previous studies have shown, vowels are predominantly uni-modal in distribution, as with /ɪ/ in Figure 25 (*right*)

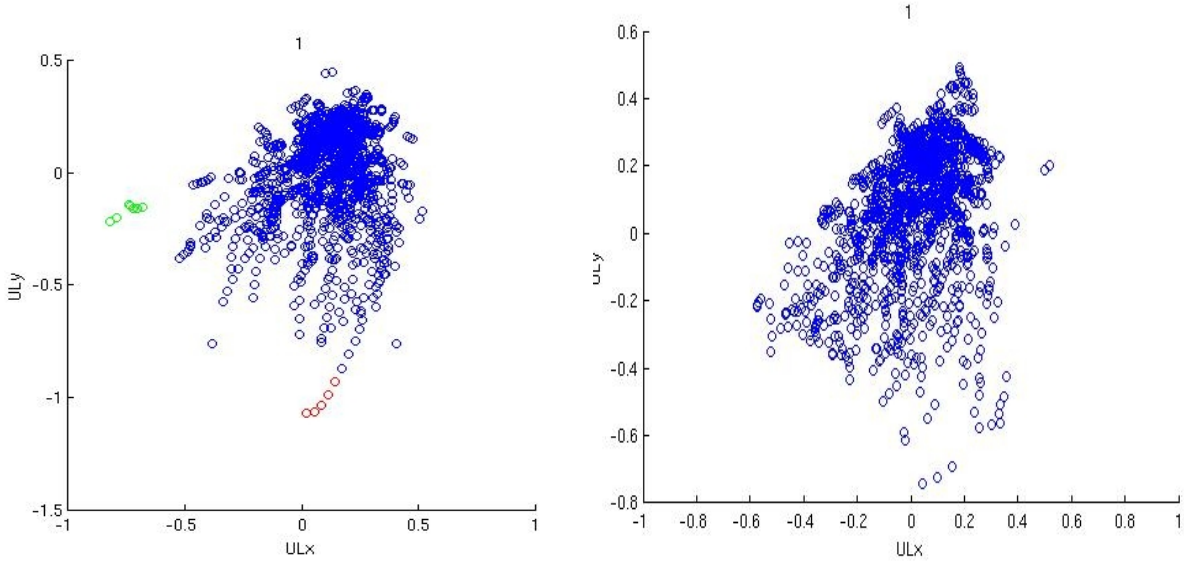


Figure 25 - UL coil locations for 1500 frames closest in acoustic space to the vowel /æ/ (*left*) and /ɪ/ (*right*).

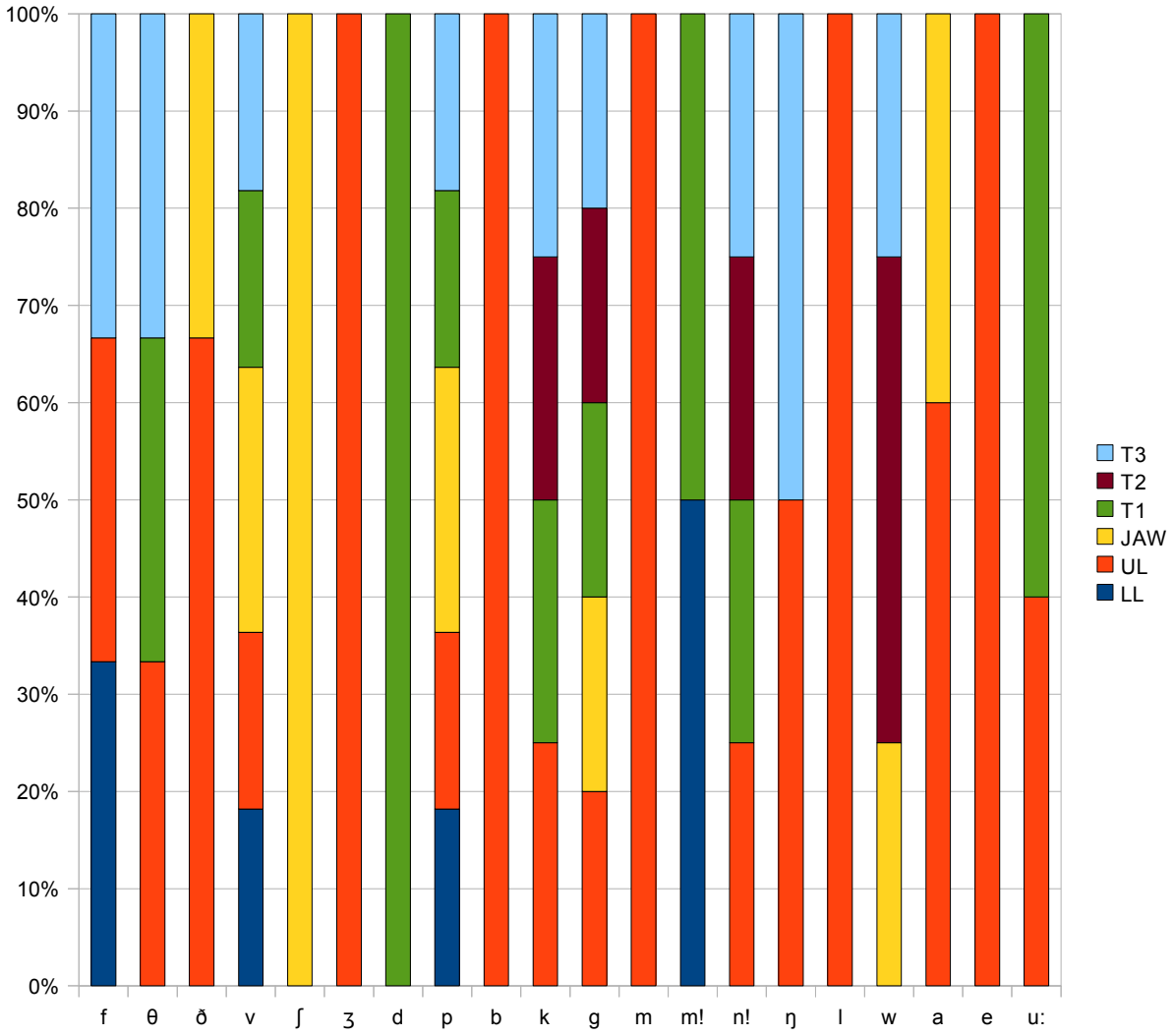


Figure 26 – A chart of all phones that demonstrated non-unique articulatory configurations from articulatory points associated with 1500 acoustically close frames to that particular phone, along with the percentage to which each articulator contributed to this articulatory non-uniqueness.

Conclusion & Further Work

Conclusion

Through this project, I have sought to quantify the extent to which different vocal-tract shapes can produce parametrically similar sounds in the *day one* subset of the *mngu0* EMA data-set. Through acoustic-articulatory inversion, subsequent clustering of the data using non-parametric methods, I found that 19.48% of the 799159 frames tested showed what could be considered a non-unique inversion from the acoustic domain to the articulatory domain,yielding multiple clusters, while the remaining 80.52% yielded a single cluster. In comparing the number of clusters found here to the proportion of multiple modes and clusters found in Qin & Carreira-Perpiñán's study (2007), my clustering algorithm found an average of approximately 14.48% more instances of inversion non-uniqueness in this data-set.

However, the size and type of the data-sets differed (Qin used X-Ray microbeam cinematography, whereas I used a much larger EMA corpus derived from a 3D EMA system which was more consistent) and so one would expect the larger data-base to present an increased proportion of non-uniqueness. Furthermore, as no frames of speech were omitted from this broad empirical study, such as diphthongs, triphthongs and transitional frames between fully realised phones, a proportion of the non-uniqueness found may be merely an attribute of continuous speech. For example, the schwa sound is attributed to many unstressed mid-central vowel sounds, but in the transition from one phone to another, different configurations of articulators in the vocal tract could produce an acoustically similar sound to the /ə/ vowel. Moreover, in investigating a selection of representative frames of phones in the data-set, it became clear that the level of non-uniqueness was not as easily quantifiable as initially thought.

Across different categories of phones, namely fricatives, plosives, nasals and approximants, some phones showing apparent articulatory non-uniqueness were revealed to produce

clustered point clouds that consisted of acoustically similar, but not equivalent, frames from other phones. Labiodental fricatives /f/ and /v/ were shown to contain clusters of articulatory frames from each other, which could suggest that unvoiced acoustic frames in the data-set are not being appropriately represented by the LSF acoustic parametrisation. Further examples of this are with bilabial plosives /p/ and /b/ and velar plosives /k/ and /g/, which all showed evidence of voiced and unvoiced phones produced by the same place of articulation being categorised as acoustically similar to a reference point and grouped into distinct clusters in the articulatory domain.

Across all 799159 frames, the upper-lip was the highest contributor of non-unique distributions of articulatory coordinates associated with 1500 acoustic vectors closest to a given acoustic reference frame, and this contribution increased when acoustic-articulatory inversion of individual phones was examined (50% of all phones tested demonstrated articulatory non-uniqueness in the distribution of 1500 acoustically similar frames). For example, the acoustic nearest neighbours to syllabic alveolar nasal stop ‘n!’ showed a non-unique articulatory distribution for the upper-lip through a substantial number of voiced bilabial plosive /m/ frames being mixed in the overall articulatory distribution. Though syllabic alveolar nasal stop ‘n!’ is not normally produced with a bilabial closure, in certain contexts alveolar consonants can be subject to anticipatory coarticulation towards initial consonant of another word (e.g. Utterance 580 in *mngu0* - ‘Prison Garth’ can sound more like ‘Prism Garth’). This could explain the acoustic similarity between a given acoustic frame of ‘n!’ and ‘m’, and thus the corresponding non-unique articulatory distribution.

The level of articulatory unique distributions of acoustically similar acoustic frames has been shown to be especially low for most articulators that are especially necessary to the production of a given phone. For example, fricatives such as unvoiced dental /t/ and unvoiced alveolar /s/ yield a single cluster for the tongue-tip, their critical articulator. However, as discussed, the upper-lip, lower-lip and tongue-dorsum differ in this respect. These articulators demonstrate, for a phone that uses that place of articulation critically, dense clusters as well as other clusters pertaining to phones that do not share the same articulatory distribution. From this, it becomes clearer that each articulator has a different range of possible articulations as well as different number of phones that use the articulator critically, and that this corresponds

to the number of clusters that can be found for any given articulator.

To summarise, the methodology outlined in this dissertation has allowed the superficial quantification of the level to which non-unique articulatory coordinate distributions, associated with 1500 acoustic vectors close to a given frame, exist in the *day one* subset of the *mngu0* EMA corpus. Phones such as bilabial plosives /p, b/, velar plosives /k, g/ and syllabic alveolar nasal stop 'n!', have been shown to contribute highly to the total non-uniqueness discovered. My results compare similarly to Qin & Carreira-Perpiñán's study (2007) with regards to vowels yielding uni-modal clouds, but differ in our findings for approximants, where I found that both lateral alveolar approximant /l/ and alveolar approximant /ɹ/ yielded only uni-modal distributions for the tongue articulators (T1, T2, T3). However, the real question is whether or not these clusters actually correspond to a phone being articulated in numerous ways, or whether, as is more likely shown in the results of this dissertation, that the parametrisation of the acoustic data allows multiple articulatory combinations to correspond to a single acoustic sound.

Further work

In many respects, the methodology put forth and tested in this paper is an outdated and insufficient way to model the inversion mapping, and thus non-uniqueness. More recent papers by Qin & Carreira-Perpiñán (2009) derived conditional densities for an articulatory configuration given an acoustic vector. Their data was modelled using a Gaussian mixture and an EM algorithm, which allows a more accurate description of the density of the data. They then used a true Gaussian mean-shift algorithm to find the modes of the dense regions, and filtered out spurious modes to find true multi-modal distributions. I did initially attempt to use a 2 component GMM along with a Gaussian mean-shift algorithm, but this resulted in two problems. Firstly, as previously discussed, I experienced difficulties in determining how many components to use for each articulator, leading to an otherwise uni-modal cloud to be clustered, and secondly, the computational complexity in performing the true Gaussian mean-shift over all 799159 frames in the corpus was too large, taking an average of 25 seconds to process each frame. However, there are a number of methods to accelerate this process.

First, as suggested in Qin & Carreira-Perpiñán (2010), thresholding out components that are beyond a certain distance from a given acoustic vector would reduce the computational bottleneck. Secondly, Carreira-Perpiñán (2006) proposed several acceleration strategies for Gaussian mean-shift, one of which used a spatial discretion strategy to divide the data into cells and make all points within a cell converge to the same mode, reducing iterations and computational complexity. Unfortunately, due to time-constraints, I could not implement an acceleration strategy here, and so used the GBMS algorithm to find clusters instead of modes (though the modes of the data will likely be within each cluster).

In further work, I would implement an accelerated mode-finding GMS algorithm but instead using Mixture-Density Networks (MDNs) as a trained model to provide a full frame-wise probability density function for each articulatory configuration conditioned on a corresponding acoustic vector. An MDN can model arbitrary distributions in data using numerous components, meaning articulatory configurations can be modelled far more accurately that other trainable inversion mapping systems (eg. MLPs). The conditional probability density outputted for each frame could then be used with a Gaussian mean-shift algorithm, using a fixed-point iteration at each centroid of the conditional mixture. From this, a richer description of the level of articulatory non-uniqueness in the *mngu0* corpus by way of mode-finding could be quantified. However, the question still remains as to whether articulatory non-uniqueness truly exists, and whether or not it is actually just a product of the parametrisation of the acoustic data.

References

1. Ananthakrishnan, G. and Engwall, O., ; “Resolving Non-Uniqueness in the Acoustic-to-Articulatory Mapping ” Centre for Speech Technology, KTH (Royal Institute of Technology), Stockholm, Sweden , ICASSP 2011, 4628-4631
2. Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W., “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique.” J. Acoust. Soc. Am. 63, 1535–1555., 1978
3. Banno, H. , Hata, H. , Morise, M. , Takahashi, T. , Irino, T. & Kawahara, H. “Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation ” The Acoustical Society of Japan , Acoust. Sci. & Tech. 28, 3 (2007)
4. Bishop, C., “Neural Networks for Pattern Recognition”. Oxford University Press, Oxford 1995
5. Carreira-Perpiñán, M. Á, “Mode-finding for mixtures of gaussian distributions,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1318–1323, 2000
6. Carreira-Perpiñán, M. Á. (2000): "Mode-finding for mixtures of Gaussian distributions". IEEE Trans. on Pattern Analysis and Machine Intelligence 22(11):1318-1323.
7. Carreira-Perpiñán, M. Á. (2006): "Acceleration strategies for Gaussian mean-shift image segmentation". IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2006), pp. 1160-1167.
8. Carreira-Perpiñán, M. Á. (2006): "Fast nonparametric clustering with Gaussian blurring mean-shift". 23rd Int. Conf. Machine Learning (ICML 2006), pp. 153-160.
9. Carreira-Perpiñan, M. Á. (2008): "Generalised blurring mean-shift algorithms for nonparametric clustering". IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2008)
10. Cheng, Y. “Mean shift, mode seeking, and clustering.” IEEE Trans. PAMI, 17, 790–799., 1995
11. Fukunaga, K., & Hostetler, L. D. “The estimation of the gradient of a density function, with application in pattern recognition.” IEEE Trans. Inf. Theory, 21, 32–40 , 1975

12. Jurafsky, D., and Martin, J. H., 2009. “Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.” 2nd edition. Prentice-Hall. 2009
13. Kobayashi, T., Yagya, M. & Shirai, K, “Application of neural networks to articulatory motion esimation” in ‘Proc ICASSP’, pp 489-492, 1991
14. Ladefoged, P. “Elements of Acoustic Phonetics” 2nd Edition, University of Chicago Press, Ltd, London, 1996
15. Lindblom, B., Lubker, J., Gay, T., “Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation”. J. Phonet. 7, 147–161. 1979
16. MacKay, D., "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp.284–292, 2003
17. Mermelstein, P., “Articulatory model for the study of speech production”. J. Acoust. Soc. Am. 53 (4), 1070–1082, 1973
18. Moore, A., “A tutorial on kd-trees”, University of Cambridge Computer Laboratory Technical Report No. 209, 1991.
19. Neiberg, D., Ananthakrishnan, G. and Engwall, O., “In search of Non-uniqueness in the Acoustic-to-Articulatory Mapping “ Interspeech 2009, 2799-2802
20. Neiberg, D., Ananthakrishnan, G. and Engwall, O., “The Acoustic to Articulation Mapping: Non-linear or Non-unique?”, in Proc. Interspeech, 1485-1488, 2008.
21. Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zachs, J. & Levy, S. (1992), “Inferring articulation and recognising gestures from acoustics with a neural network trained on X-ray microbeam data’, J Acoust Soc Am. 688-700, 1992
22. Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T. “Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements.” J Acoust Soc Am. : 3078–3096. 1992
23. Qin, C. and Carreira-Perpiñán, M. Á. (2007): "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping". Interspeech 2007, pp. 74-77.
24. Qin, C. and Carreira-Perpiñán, M. Á. (2010): "Articulatory inversion of American English /r/ by conditional density modes". Interspeech 2010, pp. 1998-2001
25. Rahim, M., Goodyear, C., Kleijn, W., Schroeter, J., Sondhi, M., “On the use of neural

networks in articulatory speech synthesis." J. Acoust. Soc. Am. 93 (2), 1109–1121., 1993

26. Rahim, M.G., Kleijn, W.B., Schroeter, J., Goodyear, C.C., "Acoustic-to-articulatory parameter mapping using an assembly of neural networks." In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 485–488, 1991
27. Richmond, K, Hoole, P. & King, S. "Announcing the Electromagnetic Articulography (Day 1) Subset of the *mngu0* Articulatory Corpus" in proc Interspeech, Florence, Italy, 2011.
28. Richmond, K., "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in Proc. Interspeech, Pittsburgh, USA, September 2006.
29. Richmond, K., "Estimating articulatory parameters from the acoustic speech signal." Ph.D. Thesis, The Centre for Speech Technology Research, Edinburgh University. 2002
30. Richmond, K., King, S., and Taylor, P., "Modelling the uncertainty in recovering articulation from acoustics." Computer Speech and Language, 17:153-172, 2003.
31. Richmond, K.. "Preliminary inversion mapping results with a new EMA corpus." In Proc. Interspeech, pages 2835-2838, Brighton, UK, September 2009.
32. Richmond, K.. "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion." In M. Chetouani, A.Hussain, B.Gas, M.Milgram, and J.-L. Zarader, editors, Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, volume 4885 of Lecture Notes in Computer Science, pages 263-272. Springer-Verlag Berlin Heidelberg, December 2007
33. Roweis, S., "Data driven production models for speech processing." Ph.D. Thesis, California Institute of Technology, Pasadena, California, 1999
34. Singampalli, V., Jackson, D., Philip, J. B. : "Statistical identification of critical, dependent and redundant articulators", In INTERSPEECH-2007, 70-73, 2007
35. Wakita, H., "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art." IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 281–285. 1979
36. Westbury J. R., "X-ray microbeam speech production database user's handbook."

Madison, WI: X-ray Microbeam Facility, 1994

37. Wrench, A., Hardcastle, W.J., "A multichannel articulatory speech database and its application for automatic speech recognition." In: Proc. 5th Seminar on Speech Production. Kloster Seeon, Bavaria, pp. 305–308., 2000
38. Zachs, J., Thomas, T.R., "A new neural network for articulatory speech recognition and its application to vowel identification." Computer Speech Language 8, 189–209., 1994

Illustrations

K-D Tree
http://en.wikipedia.org/wiki/File:3dtree.png
Accessed 18/10/2011