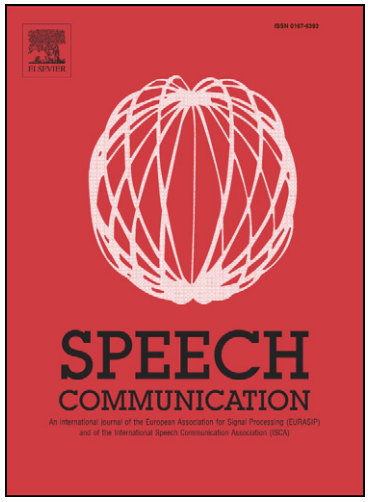
Mapping between Acoustic and Articulatory Gestures

G. Ananthakrishnan, Olov Engwall



Mapping between Acoustic and Articulatory Gestures

G. Ananthakrishnan, Olov Engwall

PII: S0167-6393(11)00016-1
DOI: 10.1016/j.specom.2011.01.009
Reference: SPECOM 1967

To appear in: *Speech Communication*

Received Date: 31 May 2010
Revised Date: 15 January 2011
Accepted Date: 21 January 2011

Please cite this article as: Ananthakrishnan, G., Engwall, O., Mapping between Acoustic and Articulatory Gestures, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.01.009

Mapping between Acoustic and Articulatory Gestures

G. Ananthakrishnan, Olov Engwall

Centre for Speech Technology (CTT), School of Computer Science and Communication, KTH (Royal Institute of Technology), SE–100 44 Stockholm, Sweden, Tel. +468-790 75 65

Abstract

This paper proposes a definition for articulatory as well as acoustic gestures along with a method to segment the measured articulatory trajectories and the acoustic waveform into gestures. Using an simultaneously recorded acoustic-articulatory database, the gestures are detected based on finding critical points in the utterance both in the acoustic and articulatory representations. The acoustic gestures are parameterized using 2-D cepstral coefficients. The articulatory trajectories are essentially the horizontal and vertical movements of Electromagnetic Articulagraphy (EMA) coils placed on the tongue, jaw and lips along the midsagittal plane. The articulatory movements are parameterized using 2D-DCT using the same transformation that is applied on the acoustics. The relationship between the detected acoustic and articulatory gestures in terms of the timing as well as the shape is studied. Acoustic-to-articulatory inversion is also performed using a GMM-based regression, in order to study this relationship further. The accuracy of predicting of the articulatory trajectories from the acoustic waveform are at par with state-of-the-art frame-based methods with dynamical constraints (with an average error of 1.45-1.55 mm for the two speakers in the database). In order to evaluate the acoustic-to-articulatory inversion in a more intuitive manner, a method based on the error in estimated critical points is suggested. Using this method, it was noted that the estimated articulatory trajectories using the acoustic-to-articulatory inversion methods were still not accurate enough to be within the perceptual tolerance of audio-visual asynchrony.

Email address: agopal@kth.se, engwall@kth.se (G. Ananthakrishnan, Olov Engwall)

Key words: Acoustic Gestures, Articulatory Gestures,
Acoustic-to-Articulatory Inversion, Critical Trajectory Error

Mapping between Acoustic and Articulatory Gestures

G. Ananthakrishnan, Olov Engwall

Centre for Speech Technology (CTT), School of Computer Science and Communication, KTH (Royal Institute of Technology), SE–100 44 Stockholm, Sweden, Tel. +468-790 75 65

1. Introduction

The relationship between an acoustic signal and the corresponding articulatory trajectories is of interest both for practical applications (such as speech coding, robust ASR, or feedback in computer-assisted pronunciation training) and on theoretical grounds, e.g. with respect to human speech perception and production. Among the different theories of speech perception, three main theories, namely the Motor theory (Liberman et al., 1967), the Direct realist theory (Diehl et al., 2004) and the Acoustic landmark theory (Stevens, 2002) claim that humans make use of articulatory knowledge when perceiving speech.

The motor theory of speech perception considers the perception of speech as a special phenomenon. According to the theory, speech perception is carried out by analyzing the signal based on the innate knowledge of the articulatory production of the particular sound. Because of the invariance in the production mechanism, signals that differ in acoustic properties by a large amount, can still be perceived as the same phonemic class. A classic example is that even though the acoustic properties of the initial segment /d/ in /da/, /di/ and /du/ are different, it is categorized into the same phonemic class.

The direct realist theory reasons along similar lines as the motor theory, but does not claim that speech perception is largely different from the

Email address: agopal@kth.se, engwall@kth.se (G. Ananthakrishnan, Olov Engwall)

perception of other kinds of sounds. The theory postulates that the objects of perception in case of speech are articulatory gestures, and not phonemic targets as proposed by the Motor theory. The gestures are inferred from evidence given in the acoustic signal.

The landmark based theory of speech perception also makes use of articulatory gestures in order to explain the phenomenon of speech perception. The theory claims that the segments in speech are encoded by different states of the articulators. Due to the quantal nature of the mapping between articulatory and acoustic parameters, when moving from a particular encoded configuration of articulators to the next, we can perceive distinct segments in the acoustics.

In this paper, we draw inspiration from the direct realist theory, in that we attempt to study the mapping between articulatory trajectories and acoustic segments of speech. For this we first define what we call acoustic and articulatory 'Gestures' and then propose a method to detect and segment these gestures. The analysis is further substantiated by performing acoustic-to-articulatory inversion, where the articulatory trajectories of an utterance are predicted from acoustic segments.

Acoustic-to-articulatory inversion has commonly been performed by applying an inversion-by-synthesis method, in which an articulatory model (such as Maeda's (Maeda, 1988)) is first used to build a codebook by synthesizing sounds from the entire articulatory space of the model (Atal et al., 1978). Inversion is then performed by a lookup in the codebook in combination with constraints on smoothness or entropy of the estimated trajectories. Recently, statistically based inversion methods have been able to provide further insight. These methods rely on databases of simultaneously collected acoustics and articulatory data, e.g., Electromagnetic Articulography (EMA) (Wrench, 1999; Toda et al., 2008; Richmond, 2002) or X-ray microbeam data (McGowan and Berger, 2009; Dusan and Deng, 2000). Some researchers have also employed visual information from the databases (such as videos or markers on the face) in order to make better predictions of the articulation (Katsamanis et al., 2008; Kjellström and Engwall, 2009). For a review of various data-driven methods, please refer to Toutios and Margaritis (2003). In brief, the problem of data-driven inversion is usually tackled using statistical regression methods, using different types of machine learning algorithms, e.g., Linear Regression (Yehia et al., 1998), Gaussian Mixture Model Regression (Toda et al., 2004a), Artificial Neural Network Regression (Richmond, 2006) and Hidden Markov Model (HMM) regression (Hiroya and Honda, 2004).

It is then assumed that the articulatory configuration, given the acoustics, is a random variable with as many dimensions as the number of measured articulator positions.

Most of the methods, both analytical and statistical, have tried to predict area functions or the position of discrete flesh points of the articulators at a particular time instant given the acoustics, rather than trying to predict the shape of the articulatory trajectory or the gestures using the acoustics of an utterance. Several researchers have used dynamic constraints on the articulatory parameters knowing that the movement is along a smooth trajectory (Ouni and Laprie, 2002; Richmond, 2006; Zhang and Renals, 2008). Özbek et al. (2009) augmented Mel Frequency Cepstral Coefficients (MFCC) with formant trajectories and showed that there is a slight improvement in the prediction of the articulator trajectories.

The above paradigm of predicting the articulator positions at each time instant can be said to draw its inspiration from the motor theory, in that it corresponds to the proposed innate mechanism of mapping the acoustics directly to the articulatory production. In contrast, we propose an inversion method that is closer to the direct realist theory, in that the units of inversion are acoustic gestures and their corresponding articulatory movements, rather than articulatory parameters at a single instance of time with smoothing constraints. Such a method of mapping gestures in the acoustic and the articulatory domains has not been tried with success before, because of two reasons. The first problem is that of segmentation. There are no clear or consistent ways of segmenting the acoustics into gestures, whereas segmenting into phonemes is deemed easier because it can be verified with our understanding of speech units. The second problem is parameterizing time-varying acoustic features. Most acoustical analyses deal with short windows of the signal where the signal is considered stationary. In order to map acoustic and articulatory gestures, a time varying parametrization is necessary.

We therefore propose a general segmentation algorithm for time-varying data, which can be applied to segment both acoustics and articulatory trajectories into units which we call 'gestures'. The segmentation is effected by finding the 'Critical points' in the acoustic and articulatory trajectories. The relationship between the acoustic and articulatory gestures, being interesting and quite complex is studied with some detail, especially the question of timing between the gestures made by different articulators with respect to the acoustics. The articulatory gestures occur at different times for the different articulators. While the acoustic gestures are likely to overlap with

the articulatory gestures made by some articulators, they may not overlap with other articulators.

When we perform inversion, no information about the articulator movements is available. So instead of using different articulatory gestures we parameterize the acoustic gestures and the corresponding movement of articulators that we observe in the training data. The corresponding articulator movements could span over one or more articulatory gestures. The acoustic gestures are parameterized using length independent time-frequency 2-D cepstral coefficients, obtained using a Two Dimensional Discrete Cosine Transform (2D-DCT). The 2D-cepstral coefficients give a time-frequency representation for these segments. The articulator movements during this acoustic gesture are also parameterized by the same function, the 2D-DCT. The parameterized gestures are then modeled as a joint distribution using the multivariate Gaussian Mixture Model (GMM). The correspondence between the acoustic and articulatory movements are learned using Gaussian Mixture Model Regression (GMMR) (Sung, 2004), which is used to predict the articulatory gestures corresponding to unseen acoustic gestures. Finally, in order to find smooth articulatory trajectories, between predictions from adjacent acoustic gestures, we perform a Minimum Jerk Smoothing.

We study the mapping between the acoustic and articulatory gestures as well as the inversion method using a corpus of simultaneous acoustic and articulatory measurements. The articulatory measurements are based on the positions of EMA coils on the tongue, lip, jaw and velum of two speakers. Based on the 'Critical points' we detect, we propose a new evaluation criterion which gives a more intuitive understanding about the errors made by acoustic-to-articulatory inversion. We also compare the proposed method against the standard frame-based inversion method where the acoustic-articulatory relationships are learned using the same machine learning technique, namely GMMR.

This article is structured as follows. In Section 2 we first describe the motivation and algorithms for segmenting the acoustic and articulatory gestures. We also describe the parameterization schemes and the machine learning techniques we use for acoustic-to-articulatory inversion. In Section 3 we detail the acoustic-articulatory data we use for our experiments as well as the methods adopted for evaluating the segmentation algorithms as well as the acoustic-to-articulatory inversion. Section 3 also describes the experiments we performed using the proposed techniques and data. We discuss the results we obtained from our experiments in Section 4, before concluding on

the findings of this study in Section 5.

2. Theory and Methods

2.1. About Gestures

Our use of the term 'Gestures' is not from a semiotic point of view, which requires that a gesture necessarily has a linguistically significant meaning. Here, a gesture is more from a phonological point of view. The gesture specifies a unit of production, such as the movement during the production of a phoneme or a syllable, as described by the direct realist theory of speech perception (Fowler, 1996).

Although it is quite clear what articulatory gestures are qualitatively, there is no clear quantitative method for defining them. The definition is even more vague when one refers to an acoustic gesture. It is especially unclear what the unit of the gesture within a sentence or a phrase is. Secondly, the notion of linear sequences of non-overlapping segments of speech has been criticized by some researchers (Browman and Goldstein, 1986; Keating, 1984). The organization of the temporal movements of different articulators may further differ for different speakers, languages or contexts. On the other hand, some studies have shown that the gestures may be controlled by invariant articulatory targets (MacNeilage, 1970) or acoustic targets (Miller, 1989) and thus, the gestures themselves may not be important and can be retrieved by applying constraints on the transitions between the acoustic or articulatory targets.

The problem of finding a correspondence between articulatory gestures and the acoustic signal thus makes it necessary to obtain a quantitative definition of what gestures imply. The same definition should be valid for both signals. Secondly, the definition should include an implicit method for segmenting individual 'gestures' from a sequence.

The notion behind our definition is that there is an innate correlation between targets and gestures, even though there may not be a one-to-one mapping between them. Each gesture has a minimum of two targets, because there must be some sort of motion involved. If there are only two targets, the object making the gesture starts at one target, move towards the second target and stops. If there are more than one target within a specified amount of time, then the object need not stop before it continues towards the next target. This is the case in the utterance of a sentence, consisting of several targets and several gestures. In theory, by controlling the curvature of the

trajectory, an object can move from one target to another via an in-between target without reducing its speed while approaching it. However, it has been found that human motor movements (especially the limbs and oculomotor systems) seem to follow the so called '$1/3^{rd}$ power law' (Viviani and Terzuolo, 1982) in the speed-curvature relationships. The velocity of motion in human motor movements is related to the curvature as

$$v(t) = kc(t)^{-1/3} \tag{1}$$

where v is the velocity and c is the curvature at time t and k is the velocity gain. This means that when the curvature is larger, the velocity is reduced to allow for greater precision (Schmidt et al., 1979). Thus reduction of velocity is a good indicator of the human motor object approaching a target. While Viviani and Terzuolo (1982) talked about motor movements in the context of hand and finger movements while writing or drawing, Perrier and Fuchs (2008) showed that even though the power law is valid in an overall sense for articulatory movements it may not hold for individual movements of the articulators, probably due to the high elasticity of their tissue. The relationship between an increase in curvature and a decrease in instantaneous velocity was however preserved. Viviani and Terzuolo (1982) also observed that the angle made by the trajectory with respect to the horizontal axis was a good indicator for segments in the motion. Points of inflection and cusps were characterized by a large change in angle made by the moving object. Thus those points where there is a drop in velocity and a large change in the angle can be considered as articulatory targets. Gestures are the motion through or towards such targets. The true targets may not be reached because of the time constraints while uttering a sentence and by how much they are missed depends on the velocity.

We propose a two-step approach in segmenting gestures. First we locate what we call the 'critical points' in the trajectory, which are the projections of the theoretical targets onto the trajectory. We then define a gesture as the motion through one such 'critical point'.

2.2. Articulatory Gestures

For an utterance with T time samples, let $\boldsymbol{\gamma}_a(t) \in \Re^n$ be the column vector corresponding to the position of the articulator a at time instant t. The absolute velocity (speed) $v_a(t) = |\boldsymbol{\gamma}_a(t) - \boldsymbol{\gamma}_a(t-1)|_2$ is calculated between the positions $\forall\, t : 2 \leq t \leq T$, where $|.|_2$ is the L-2 norm of the vector. The

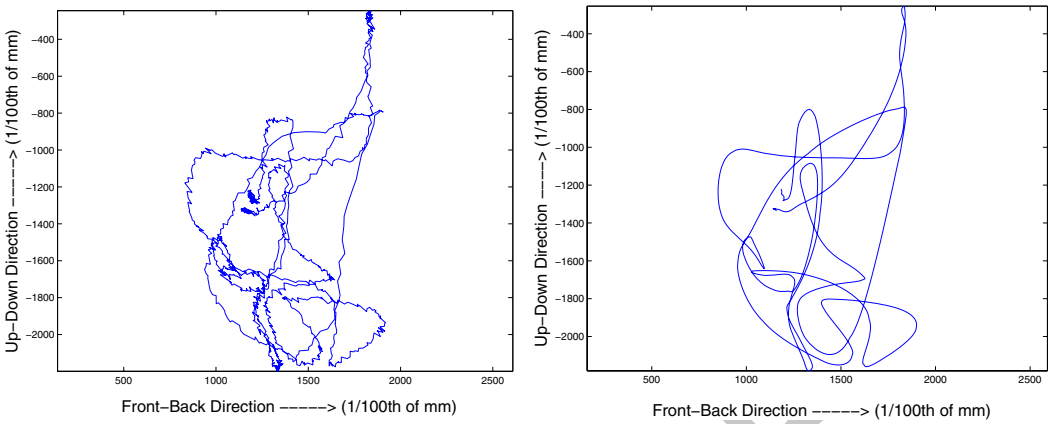
'Importance' function, which gives an indication of how close the position is to a target, $I_a(t)$ can be calculated as

$$I_a(t) = log\left(\frac{\theta_a(t)}{2\pi} - \frac{v_a(t)}{\max\limits_{1\leq i\leq T} v_a(i)}\right) \tag{2}$$

The angle $\theta_a(t)$ is the acute angle (in radians) between the vectors $\boldsymbol{\gamma}_a(t-1) - \boldsymbol{\gamma}_a(t)$ and $\boldsymbol{\gamma}_a(t) - \boldsymbol{\gamma}_a(t+1)$. A 'critical point' is a local maximum in this 'Importance' function. The Importance function needs to be smooth in order to find good local maxima, and a minimum jerk trajectory algorithm is therefore used for smoothing. A minimum jerk trajectory is the smoothest possible trajectory an object can take between two points with the minimum peak velocity during the trajectory. Since jerk is the third derivative of the position, setting the fourth derivative to zero would minimize the jerk. In order to fit the minimum jerk trajectory, we need to integrate the fourth order differential equation. Solving for each of the 4 derivatives as well as the constant of integration gives us a 5^{th} order polynomial equation. Given the noisy (jittery) trajectory of the object $\boldsymbol{\gamma}_a(t)$, a smoothed version $\boldsymbol{\gamma}_{sa}(t)$ can be obtained as

$$\boldsymbol{\gamma}_{sa}(t) = \begin{bmatrix} 1 & & 1 \\ t & & t \\ t^2 & & t^2 \\ t^3 \ldots n \text{ times} & & t^3 \\ t^4 & & t^4 \\ t^5 & & t^5 \end{bmatrix}^T \begin{bmatrix} \bar{\mathbf{1}} & \bar{\mathbf{t}} & \bar{\mathbf{t}}^2 & \bar{\mathbf{t}}^3 & \bar{\mathbf{t}}^4 & \bar{\mathbf{t}}^5 \\ & & & \vdots & & \\ & & & n \text{ times} & & \\ \bar{\mathbf{1}} & \bar{\mathbf{t}} & \bar{\mathbf{t}}^2 & \bar{\mathbf{t}}^3 & \bar{\mathbf{t}}^4 & \bar{\mathbf{t}}^5 \\ \bar{\mathbf{0}} & \bar{\mathbf{1}} & 2\bar{\mathbf{t}} & 3\bar{\mathbf{t}}^2 & 4\bar{\mathbf{t}}^3 & 5\bar{\mathbf{t}}^4 \\ & & & \vdots & & \\ & & & n \text{ times} & & \\ \bar{\mathbf{0}} & \bar{\mathbf{1}} & 2\bar{\mathbf{t}} & 3\bar{\mathbf{t}}^2 & 4\bar{\mathbf{t}}^3 & 5\bar{\mathbf{t}}^4 \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \mathbf{2} & 6\bar{\mathbf{t}} & 12\bar{\mathbf{t}}^2 & 20\bar{\mathbf{t}}^3 \\ & & & \vdots & & \\ & & & n \text{ times} & & \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \bar{\mathbf{2}} & 6\bar{\mathbf{t}} & 12\bar{\mathbf{t}}^2 & 20\bar{\mathbf{t}}^3 \end{bmatrix}^{\dagger} \begin{bmatrix} \boldsymbol{\gamma}_a(\bar{\mathbf{t}}) \\ d\boldsymbol{\gamma}_a(\bar{\mathbf{t}}) \\ d^2\boldsymbol{\gamma}_a(\bar{\mathbf{t}}) \end{bmatrix} \tag{3}$$

where † indicates the pseudo-inverse of a matrix. $\bar{\mathbf{t}}$ is a column vector of time instances from interval $[t-w_s, t+w_s]^T$, with $2w_s+1$ being the window length. $\bar{\mathbf{0}}, \bar{\mathbf{1}}$ and $\bar{\mathbf{2}}$ are column vectors the same size at $\bar{\mathbf{t}}$. Thus $\boldsymbol{\gamma}_a(\bar{\mathbf{t}})$ would



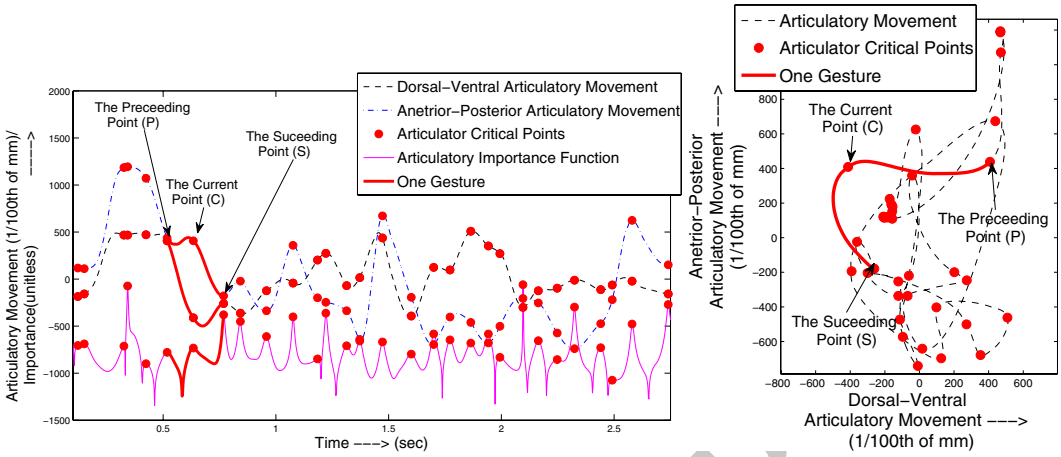
(a) Original Trajectory (b) Smoothed Trajectory

Figure 1: (a) The original recording of the trajectory of an EMA coil placed on the tongue tip along the mid-sagittal plane during the utterance of the sentence, "Jane may earn more money by working hard". (b) The smoothed version of the same trajectory using minimum jerk smoothing with smoothing window $w_s = 40$ ms.

be a column vector with $(2w_s + 1) * n$ rows[1]. The trajectory is expected to be smooth and following minimum jerk within this window. Figure 1 shows the original jittery trajectory and the smoothed version of an EMA coil placed on the tongue tip in the MOCHA-TIMIT recordings (Wrench, 1999). The jitter in the signal can probably be attributed to measurement errors of the EMA coil.

The Importance function, calculated on this smooth trajectory has more reliable local maxima than when calculated on the original trajectory, facilitating better detection of 'critical points'. The level of smoothing and thus the number of critical points depends on the window length. The larger the window, the finer transitions in the trajectory will be smoothed over, hence resulting in fewer gestures. Figure 2 shows the Importance function of the trajectory calculated using Equation 2 and the critical points obtained from its local maxima. Since a gesture was defined as the movement through at least one such critical point, we consider a gesture as the movement between two alternate critical points. That is, for every critical point C, the gesture starts from the preceding critical point P and lasts until the succeeding one S unless C is the first or the last critical point. Adjacent gestures overlap, since

[1] $*$ is the multiplication symbol



(a) Importance Function (b) Critical Points in 2D

Figure 2: (a) The trajectory of an EMA coil placed on the tongue tip along the mid-sagittal plane during the utterance of the sentence, "Jane may earn more money by working hard" along with the Importance function and 'critical points'. It can be noted that the absolute value of the Importance function is not crucial, but the relative importance for different parts of the articulatory trajectories is. Hence the y-axis denotes the scale only for the articulatory trajectories ($1/100^{\text{th}}$ of an mm). (b) The EMA trajectory along the vertical and horizontal axes. One such 'gesture' is also shown.

the trajectory PC of one gesture corresponds to CS for the previous one. The importance function as well as one such gesture is shown in Figure 2.

The application of the above method to motion, such as in articulatory data, is rather intuitive in view of the speed-curvature relationship. We propose to apply the same paradigm to acoustic signals, as outlined in the following subsection.

2.3. Acoustic Gestures

There are several automated methods to segment speech into small time units. Segmentation after counting the number of level-crossings in a region of the speech waveform (Sarkar and Sreenivas, 2005) is usually highly accurate. Methods using intra-frame correlation measures between spectral features to obtain the segments called the Spectral Transition method (STM) (Svendsen and Soong, 1987) is also a popular method. Statistical modeling using Autoregression (or ARMA) models (Van Hemert, 1991) and HMM based methods (Toledano et al., 2003) are often used to good effect. Many different features like amplitude (Farhat et al., 1993), short time energy in

different frequency sub-bands (Gholampour and Nayebi, 1998; Ananthakrishnan et al., 2006), fundamental frequency contour, (Saito, 1998), auditory models, (Zue et al., 1989) and Mel Frequency Cepstral Coefficients (MFCC) (Toledano et al., 2003) have also been tried. While most research is directed towards detecting boundaries, some algorithms, including the one presented in this article, are directed towards finding acoustic landmarks (Zue et al., 1989; Liu, 1996) in the stable regions of the speech signal. The landmarks have often been described as linguistically or phonetically motivated events. The approach we have used is following Ananthakrishnan et al. (2006) as we find the energy along different frequency sub-bands to give multi-dimensional acoustic trajectories along time, and then locate the landmarks by applying simple physical rules on these acoustic trajectories.

We represent the acoustic signal as a time-varying filter-bank based on the Equivalent Rectangular Band-width (ERB) scale (Moore and Glasberg, 1983) instead of the traditional 'Mel' scale. The advantage of using such a filter-bank is its relationship with the critical bands of hearing, in which the noise outside the critical band is suppressed. In contrast to the short-time segmental approach, the signal is filtered into frequency sub-bands. The k^{th} spectral component of the transform of the time signal $x(t) : 1 \leq t \leq T$ sampled at sampling frequency F_s is given by

$$\mathbf{X}_k(t) = \alpha(k) \sum_{m=1}^{L(k)} \mathbf{W}_k(m) x(t-m) \tag{4}$$

where, $L(k)$ is the length of the window corresponding to the k^{th} spectral component. $\alpha(k)$ is a weight that is set to 1 in the current experiments, but could correspond to the equal loudness weights or pre-emphasis.

The window function $\mathbf{W}_k : 1 \leq k \leq K$ is a set of Finite Impulse Response (FIR) linear phase band pass filters. Their Central Frequencies (C_F) are calculated by dividing the ERB scale into K equal parts, where K is the total number of filters (45 in our experiments). $C_F(K)$ must be less than $F_s/2$. Their Band-Widths (B_W) are calculated by Equation 5 which is the approximation of the ERB scale made by Moore and Glasberg (1983)

$$B_W = 6.23 * 10^{-6} * f^2 + 9.339 * 10^{-2} * f + 28.52 \tag{5}$$

where f is the frequency in Hz. The order, $L(k)$, depends on the pass band frequency and is calculated as $L(k) = 2/B_W(k)$. The order for the FIR

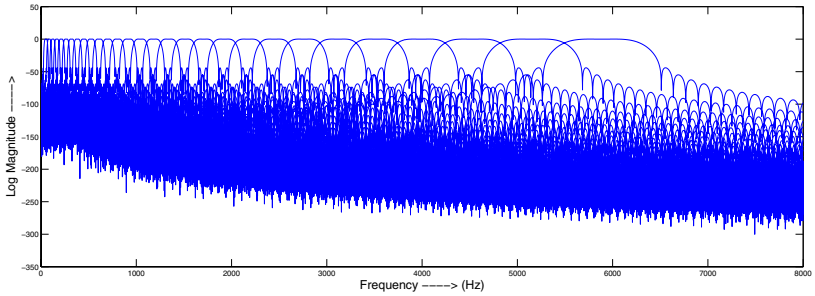


Figure 3: The frequency magnitude response of the ERB Filter-banks $\mathbf{W}_k$ with $B = 80$ Hz and 45 filters. One can see that the sub-band ripple is below 40 dB for all the filters.

filters also indicates the time resolution of the filters. One can see that these are dependent on the frequency giving higher temporal resolution to higher frequencies and higher frequency resolution to lower frequencies. Thus this sort of spectral modeling is expected to be an advantage over the traditional short-time analysis window methods. The filter $\mathbf{W}_k(t) : 1 \leq t \leq L(k)$ is calculated as follows

$$\mathbf{W}_k(t) = H(t) * \frac{\sin\left(\frac{(t-(L(k)/2))\,B_W(k)}{F_s}\right)}{\left(t - \frac{L(k)}{2}\right)} * \exp\left(\frac{-j2\pi t C_F(k)}{F_s}\right) \quad (6)$$

where $H(t)$ is the windowing function, in this case, the 'Hann' window. Figure 3 shows the frequency response of the designed ERB filter-bank. It is quite clear from this figure that while the main lobe (pass-band lobe) is quite flat, the sub-band ripple for all the filters is below 40 dB. This reduces the leakage from the higher frequency sub-bands to lower frequency ones. This property would not be exhibited by a uniform order filter-bank.

The complex signal $\mathbf{X}_k(t)$ is then converted to a real signal by finding its absolute value and compressing it using the log scale approximation of loudness, as

$$l\mathbf{X}_k(t) = 10\log_{10}(|\mathbf{X}_k(t)|^2) \quad (7)$$

where $|.|$ is the absolute value. The real signal $l\mathbf{X}_k(t)$ is used for further processing. In our experiments the minimum frequency of the filter-bank was 80 Hz, the maximum frequency was less than 6500 Hz and the total number of filters was 45. The configuration was not optimized for the task

at hand, but small changes in these numbers did not result in any larger differences in the experimental results.

Figure 4 shows the original output of the filter-banks and after smoothing with the minimum jerk formulation, which can be considered as a 5^{th} degree polynomial smoothing, with weighted coefficients. While this provides a smoothing for the frequency representation, it does not remove the salient features of the spectrogram, as is also illustrated in Figure 5

Applying the conditions for finding the Importance function as in Equation 2, the angle θ_a and speed v_a are calculated on a K dimensional vector $l\mathbf{X}_k$. The 'critical points' are detected as defined in Section 2.2. We hope the 'critical points' to correspond to the stable regions of the acoustics, which are the projections of the acoustic target onto the measured acoustic space. Thus this algorithm should be able to predict the salient landmarks in the speech signal.

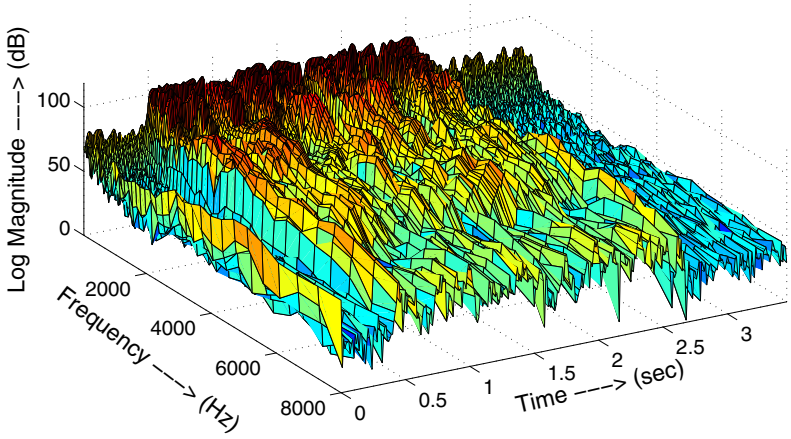
2.4. Two-Dimensional Discrete Cosine Transform

Mel Frequency Cepstral Coefficients (MFCC) are the most common acoustic parametrization for speech recognition and more recently synthesis. The cepstra are often calculated by taking the cosine transform of the short time log of the frequency warped spectrum of the acoustic signal. It is known that MFCCs of consecutive segments of speech are highly correlated. In order to use time-varying information, velocity (or acceleration) coefficients are also added in the parameterizations.

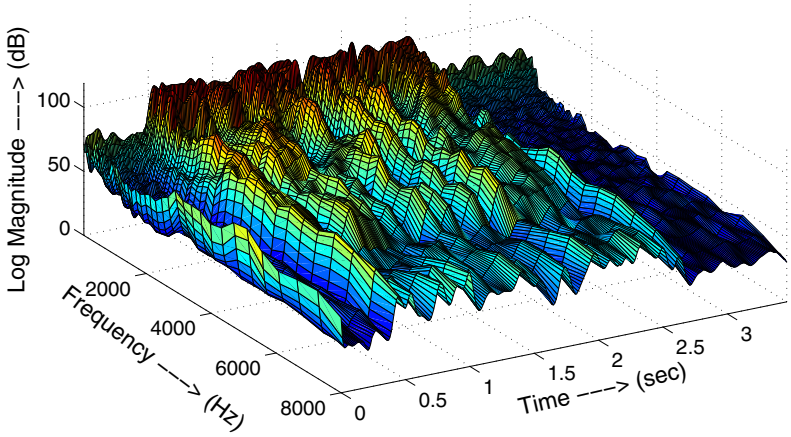
A two-dimensional cepstrum (2D-cepstrum) along the time and frequency dimensions was suggested by Ariki et al. (1989), with a linear frequency scale. It was later adapted to the Mel Frequency scale by Milner and Vaseghi (1995). Such a parametrization of speech is shown to be a time varying representation with parameters that are highly de-correlated with each other. Thus, by using 2-D cepstra, further feature reduction schemes such as Principal Component Analysis or Linear Discriminant Analysis need not be performed in order to reduce the correlation between the features.

In most previous studies, the 2D-cepstrum was calculated for a fixed duration window. In this study, it is instead calculated for segments of varying duration, since the duration of each gesture could vary greatly, and a length independent representation of the acoustic segment is hence required.

The 2D-cepstra are calculated by applying a 2-dimensional discrete cosine transform (2D-DCT), as follows. For $1 \leq p \leq P$ and $1 \leq q \leq Q$, (where P

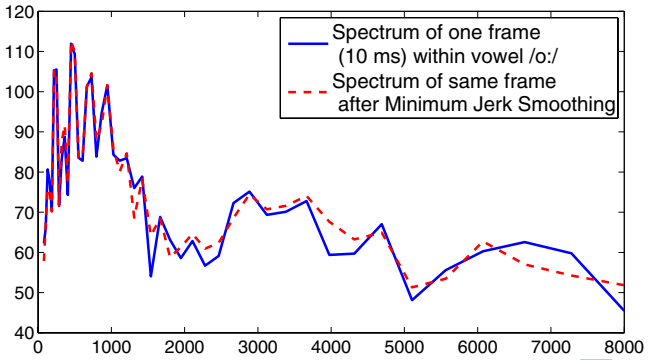


(a) ERB output

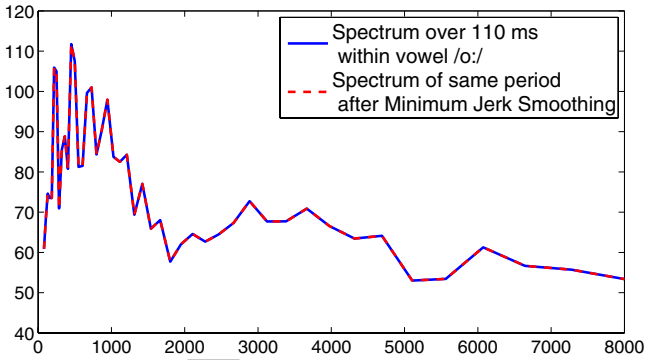


(b) Smoothed ERB output

Figure 4: A part of the spectrogram from ERB filter-bank outputs of an utterance of the sentence "Jane may earn more money by working hard" sub-sampled to 500Hz, before and after minimum jerk smoothing with $w_s = 40ms$.



(a) Spectrum of one frame



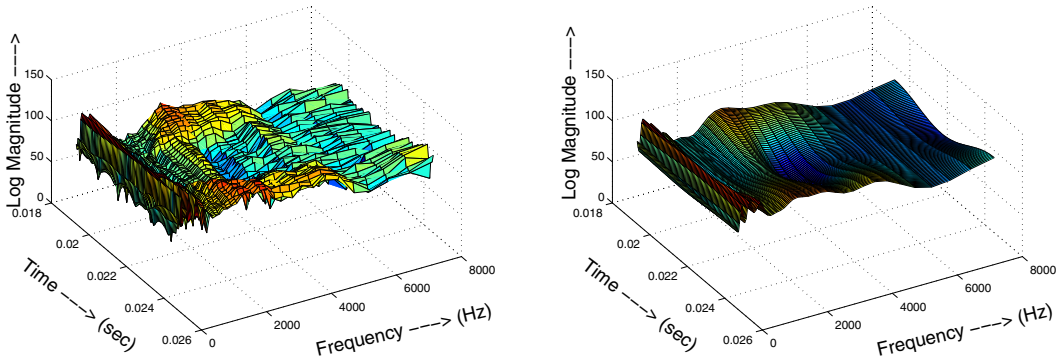
(b) Long-term average spectrum

Figure 5: The figure illustrates that although the Minimum Jerk Smoothing affects the spectrum of single frames (in this case a frame corresponding to the vowel /o:/ in the context of the word 'more'), there is no significant difference in spectral properties when averaged over time.

and Q are the number of cepstra in the frequency and time, respectively), the time-varying cepstral coefficients are

$$\boldsymbol{\tau}(p,q) = \sum_{t=1}^{T}\sum_{k=1}^{K} \frac{l\mathbf{X}_k(t)}{T} * cos\left(\frac{\pi(k-\frac{1}{2})(p-1)}{K}\right) * cos\left(\frac{\pi(t-\frac{1}{2})(q-1)}{T}\right) \tag{8}$$

where K is the total number of frequency components (or filters) as in Equation 4 and T is the length of the gesture in terms of number of samples. The axis along p is called the 'quefrency' and the axis along q is the corresponding parameter along time, which we call 'meti', following the tradition of flipping the first two syllables. Quefrency has the units of time and meti has the units of frequency. It should be noted that the 2D-DCT has been modified so that



(a) Original Segment

(b) Reconstructed Segment

Figure 6: (a) The original spectrogram segment (output of the ERB filter-bank) of the acoustic gesture during the sequence of phonemes /ʤ eɪ n/ in the context of the word 'Jane'. (b) Reconstructed spectrogram segment(from 2D-cepstrum with $P = 18$ and $Q = 3$)

this representation is length invariant, which means that the parameters are not affected by stretching or compression in time. In that sense, this representation is length-normalized. By selecting P and Q to be smaller than K and T respectively, this representation provides a compression of complexity, i.e. the representation is only an approximation of the original signal.

Along with the 2D-cepstra, which were normalized with respect to time, the actual duration of the gesture is taken as an additional feature, in case there were would be dependencies on duration. Figure 6 shows what features of the original spectrogram, obtained from the output of the ERB filter-banks for a gesture, are retained by the parameterized 2D-cepstrum.

The articulatory gestures were parameterized in the same manner, i.e. 2D-DCT coefficients, without allowing for compression in the number of articulatory parameters., i.e., P was the same as the number of articulatory parameters and Q was the same as the number chosen for the acoustic gestures. Thus 2D-DCT is a parametrization scheme which is applicable to time-varying segments of both acoustic signals and articulatory trajectories.

While the methods described in Section 2.2 and 2.3 have shown how it is possible to segment articulatory gestures, the correspondence between the acoustic and articulatory gestures is quite complex as will be discussed in Section 4.2. While we have simultaneous recordings of acoustics and articulation for training the regression for acoustic-to-articulatory inversion, we do not have knowledge of the articulation during the actual inversion process.

For this reason, we do not segment the articulatory gestures separately but apply the acoustic segmentation to the articulator movements. Instead of parameterizing the articulatory gestures, we parameterize the articulatory movement corresponding to the segmented acoustic gestures. This makes the task for acoustic-to-articulatory inversion more tractable. Secondly the matrix $\boldsymbol{\tau}$ is converted to a vector, in order to perform regression.

2.5. GMM Regression

When performing acoustic-to-articulatory inversion, the acoustic waveforms of the utterances are first segmented into overlapping gestures and each acoustic gesture is parameterized using 2D-cepstra. For each acoustic gesture, the corresponding articulatory movement, which may be a part of one or more gestures for different articulators is also parameterized using 2D-DCT. The the mapping between acoustic gestures and their corresponding articulatory movement is learnt using one of the state-of-the-art machine learning algorithms, Gaussian Mixture Model regression (GMMR) (Sung, 2004). It is a piece-wise linear space approximation and it can be used to calculate the regression in a probabilistic sense. The GMMR is explained briefly below, following the notation used by Toda et al. (2004b).

The conditional probability density of a variable $\mathbf{y}_t \in \Re^d$ (in this case the vectorized version of the 2D-DCT on articulatory movements) conditioned on variable $\mathbf{x}_t \in \Re^D$ (in this case the vectorized version of the 2D-cepstra), modeled as a GMM with M Gaussians for a given instance t (in this case one acoustic gesture), is represented as

$$P(\mathbf{y}_t|\mathbf{x}_t) = \sum_{m=1}^{M} P(m|\mathbf{x}_t) P(\mathbf{y}_t|\mathbf{x}_t, m) \tag{9}$$

where

$$P(m|\mathbf{x}_t) = \frac{\rho_m \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \mathbb{S}_m{}^{xx})}{\sum_{n=1}^{M} \rho_n \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n^x, \Sigma_n^{xx})} \tag{10}$$

is the conditional probability of each Gaussian. $\mathcal{N}$ represents the function of a Gaussian distribution. The parameters are $\boldsymbol{\mu}_m^x$ and $\mathbb{S}_m{}^{xx}$, the mean vector and covariance matrix respectively. ρ_m are the weights for the individual Gaussians.

$$P(\mathbf{y}_t|\mathbf{x}_t, m) = \mathcal{N}(\mathbf{y}_t; \mathbf{E}_{m,t}^y \mathbb{D}_m^{yy}) \tag{11}$$

is the conditional distribution of each Gaussian component with parameters $\mathbf{E}^{y}_{m,t}$, the mean vector and $\mathbb{D}^{yy}_{m}$, the covariance matrix. $\mathbf{E}^{y}_{m,t}$ is calculated as

$$\mathbf{E}^{y}_{m,t} = \boldsymbol{\mu}^{y}_{m} + \mathbb{S}_m{}^{yx}(\mathbb{S}_m{}^{xx})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}^{x}_{m}) \tag{12}$$

If the covariance matrices of the individual Gaussian components for the joint distribution of $[\mathbf{x}_t\, \mathbf{y}_t]^T$ is $\mathbb{S}_m$, then

$$\mathbb{S}_m = \begin{bmatrix} \mathbb{S}_m{}^{xx} & \mathbb{S}_m{}^{xy} \\ \mathbb{S}_m{}^{yx} & \mathbb{S}_m{}^{yy} \end{bmatrix} \tag{13}$$

Thus the covariance matrix $\mathbb{D}^{yy}_{m}$ is calculated as follows.

$$\mathbb{D}^{yy}_{m} = \mathbb{S}_m{}^{yy} - \mathbb{S}_m{}^{yx}(\mathbb{S}_m{}^{xx})^{-1}\mathbb{S}_m{}^{xy} \tag{14}$$

The Minimum Mean Square Estimate (MMSE) for the regression, $\widehat{\mathbf{y}}_t$, given $\mathbf{x}_t$, is calculated as

$$\widehat{\mathbf{y}}_t = E[\mathbf{y}_t|\mathbf{x}_t] = \sum_{m=1}^{M} P(m|\mathbf{x}_t)\mathbf{E}^{y}_{m,t} \tag{15}$$

where $E[.]$ is the expectation of the distribution. The GMM on the joint space (xy) is obtained using the Expectation Maximization (EM) algorithm (Bilmes, 1998). The estimated vector is the weighted average of the different conditional means estimated over individual Gaussian components.

2.6. Minimum Jerk Smoothing with multiple weighted hypotheses

In our method, the estimates of the articulatory trajectories are parameterized and are hence calculated by the inverse transform of Equation 8, taking care of the length of the required articulatory segments. Due to overlapping acoustic gestures, there is a corresponding overlap of trajectory estimates at the critical point (c.f. Section 2.1). The predicted articulatory movement corresponding to adjacent acoustic gestures may not form a smooth articulator path. There are at least 2 hypotheses about the estimated articulatory positions at every point from the preceding and succeeding acoustic gestures and in fact 3 hypotheses at the critical point. This overlap in information is handled using a minimum jerk smoothing with multiple weighted hypotheses as shown in Figure 7. In the current implementation, the weights for each hypothesis are set to be equal, but they could be optimized through further

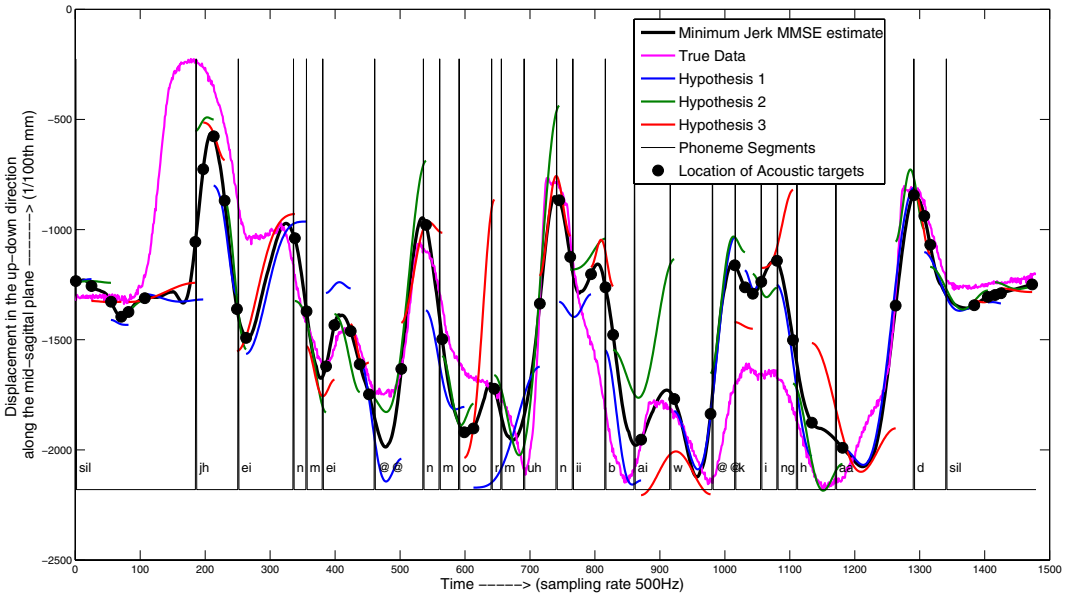


Figure 7: Due to overlapping acoustic gestures, multiple hypothesis about the articulatory gestures are generated. For every critical point, Hypothesis 1 is the articulatory segment predicted corresponding to the previous critical point in the acoustics, Hypothesis 2 is the predicted segment corresponding to the current critical point and Hypothesis 3 is the predicted segment corresponding to the succeeding critical point. This example is shows the acoustic-to-articulatory inversion for the tongue-tip (TT) during the sentence, "Jane may earn more money by working hard"

.

experimentation. The minimum jerk smoothing for a time column vector $\overline{\mathbf{t}}$ of time interval $[t - w_s, t + w_s]^T$, with multiple hypotheses at each time instant, is the minimum mean square error (MSE) solution to the optimization function J,

$$J(\boldsymbol{\beta}_t) = (\boldsymbol{\varphi} - \mathbb{G} * \boldsymbol{\beta}_t)^T * diag(\boldsymbol{\Phi}) * (\boldsymbol{\varphi} - \mathbb{G} * \boldsymbol{\beta}_t) \tag{16}$$

$\boldsymbol{\beta}_t^{5\times 1}$ are the parameters of the minimum jerk trajectory. The vector $\boldsymbol{\varphi}^{3*(2*w_s+1)*h\times 1}$ is given by

$$\boldsymbol{\varphi} = \begin{bmatrix} \begin{bmatrix} H_1(\overline{\mathbf{t}})^T & dH_1(\overline{\mathbf{t}})^T & d^2H_1(\overline{\mathbf{t}})^T \end{bmatrix}^T \\ \begin{bmatrix} H_2(\overline{\mathbf{t}})^T & dH_2(\overline{\mathbf{t}})^T & d^2H_2(\overline{\mathbf{t}})^T \end{bmatrix}^T \\ \vdots \\ \begin{bmatrix} H_h(\overline{\mathbf{t}})^T & dH_h(\overline{\mathbf{t}})^T & d^2H_h(\overline{\mathbf{t}})^T \end{bmatrix}^T \end{bmatrix} \tag{17}$$

where $[H_1 \quad H_2 ... \quad H_h]^T$ are the h hypotheses predictions of the values of one articulatory parameter from the acoustic-to-articulatory inversion based on 3 consecutive acoustic gestures. If the t corresponds to a critical point, then $h = 3$, otherwise $h = 2$. dH and d^2H denote the corresponding velocity and the acceleration parameters. Matrix $\mathbb{G}^{3*(2*w_s+1)*h\times 6}$ is

$$\mathbb{G} = \begin{bmatrix} 1 & \bar{\mathbf{t}} & \bar{\mathbf{t}}^2 & \bar{\mathbf{t}}^3 & \bar{\mathbf{t}}^4 & \bar{\mathbf{t}}^5 \\ 0 & 1 & 2\bar{\mathbf{t}} & 3\bar{\mathbf{t}}^2 & 4\bar{\mathbf{t}}^3 & 5\bar{\mathbf{t}}^4 \\ 0 & 0 & 2 & 6\bar{\mathbf{t}} & 12\bar{\mathbf{t}}^2 & 20\bar{\mathbf{t}}^3 \\ \text{repeat} & h & \text{times} & \vdots & \vdots & \vdots \\ 1 & \bar{\mathbf{t}} & \bar{\mathbf{t}}^2 & \bar{\mathbf{t}}^3 & \bar{\mathbf{t}}^4 & \bar{\mathbf{t}}^5 \\ 0 & 1 & 2\bar{\mathbf{t}} & 3\bar{\mathbf{t}}^2 & 4\bar{\mathbf{t}}^3 & 5\bar{\mathbf{t}}^4 \\ 0 & 0 & 2 & 6\bar{\mathbf{t}} & 12\bar{\mathbf{t}}^2 & 20\bar{\mathbf{t}}^3 \end{bmatrix} \tag{18}$$

The weight vector,$\mathbf{\Phi}^{3*(2*w_s+1)*h\times 1}$, is the weight for each hypothesis for each time instance in $\bar{\mathbf{t}}$. The velocity and acceleration parameters can also be weighted independently. Using the parameters, $\boldsymbol{\beta}_t$, the new smoothed trajectory $\widehat{\gamma}(t)$ can be found by

$$\widehat{\gamma}(t) = \begin{bmatrix} 1 & t & t^2 & t^3 & t^4 & t^5 \end{bmatrix} * \boldsymbol{\beta}_t \tag{19}$$

3. Data and Experiments

3.1. Data

We used two sets of data for running our experiments. The detection of acoustic gestures required an acoustic database with highly accurate transcription and segmented data, in order to assess how well the acoustic critical points are detected. For this purpose we used the TIMIT database (Seneff and Zue, 1988). The test set contained sentences spoken by 168 speakers in 8 American dialects with a total of 1344 sentences. Since the method did not use any training, the experiments were run directly on the test corpus.

The rest of the experiments were conducted on the simultaneously recorded Acoustic-EMA data from the MOCHA database (Wrench, 1999) consisting of 460 TIMIT sentences spoken by two speakers, one male (msak) and one female (fsew). The sentences had a total number of 46 phonemes including silence and breath. The sentences were more or less balanced in covering the phoneme set. The sentences were recorded along with 14 articulatory

channels consisting of the X- and Y-axis trajectories of 7 EMA coils, namely Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (VE). The position of the coils were selected to represent the position of the important articulators that have often been described in previous studies (Perkell et al., 1992; Hoole, 1996). It has also been shown that the whole tongue contour (or at least the oral region) can be predicted from knowing the positions of these coils (Qin et al., 2008). Since the magnet for recording the EMA coils was head mounted, the positions were normalized for head movement. The co-ordinates were rotated so that the origin coincided with another coil placed at the upper jaw, which was invariant to the articulation. The articulatory trajectories were processed as described by Richmond (2002) in order to remove the drift. The system recorded the data with a resolution of 0.01 mm, but the effective resolution was 0.43 mm on an average (Hoole, 1996).

3.2. Description of the Experiments

We used the TIMIT database containing only acoustic data in order to verify whether the gesture segmentation and critical points detection algorithm could detect acoustically relevant segments and salient landmarks respectively. We calculated the number of phonemes that were represented by at least one gesture and the number of phonemes that were represented by more than one gesture. Thus, the accuracy of the segmentation would be indicated by how many times at least one critical point is detected within the duration of a phoneme. An insertion denotes whether a phoneme was segmented into more than one gesture. More than one gesture per phoneme may be suitable for diphthongs or aspirated stop consonants, but may not be appropriate for other stop consonants, fricatives and vowels.

By increasing the smoothing (larger w_s), the number of insertions were expected to decrease but at the cost of not detecting all the phonemes. One must note here that the focus of this segmentation scheme is not on getting highly accurate acoustic segments, but to have a scheme which is also compatible with segmenting articulatory trajectories in order to explain correspondences between acoustic and articulatory gestures. The results are detailed in Section 4.1

It is more difficult to judge whether the articulatory gestures are detected correctly and meaningfully. Critical points constituted around 1 to 4% of the trajectory lengths depending on the articulator and the content of the sentence. By performing minimum jerk interpolation between the critical

points, the entire trajectories were estimated. These were compared with the original trajectories. The error is expected to increase with a larger value of w_s. As a comparison, we also interpolated and estimated trajectories from randomly selected points on the original trajectories. We used the articulatory measurements from the MOCHA-TIMIT database in order to evaluate the detection of articulatory gestures.

Once the acoustic and articulatory gestures are obtained, one needs to ascertain if these gestures had any relevance to the study in terms of correspondences between acoustics and articulatory trajectories. We therefore made several analyses to compare the obtained acoustic and articulatory critical points as well as the corresponding gestures. We first performed a visual assessment comparing the timing between the acoustic gesture and articulatory gestures, explained in Section 4.2. We then performed a clustering of all the articulatory gestures, which were parameterized using the 2D-DCT. Based on the clustering, we looked at phonetic labels of the acoustic segments corresponding to the critical point in the articulatory trajectory. We expected that the clustering would reveal a relationship between the articulatory gestures and the role the gestures play in producing the particular phoneme.

Finally, we performed acoustic-to-articulatory inversion in order to assess whether the gesture detection and parametrization schemes were effective or not. This was compared to a baseline method using the same machine learning algorithm (GMMR), but using point-by-point inversion, i.e. inversion from every frame of the acoustics to the corresponding articulator position. This is described in further detail in Section 3.5.

3.3. Evaluation Criteria for the Inversion

Since the predictions of the articulation are trajectories, a commonly used evaluation criterion in acoustic-to-articulatory inversion is the Root Mean Square Error ($RMSE$) between the measured and estimated trajectories of every articulator a.

$$RMSE_a = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\gamma}_a(t) - \widehat{\boldsymbol{\gamma}}_a(t))^2} \tag{20}$$

where $\boldsymbol{\gamma}_a$ is the measured trajectory and $\widehat{\boldsymbol{\gamma}}_a$ is the estimated trajectory of length T. The mean $RMSE$ ($mRMSE$) is the mean across all the A artic-

ulators, calculated as

$$mRMSE = \sum_{a=1}^{A} RMSE_a \tag{21}$$

The second standard evaluation criterion is the Correlation Coefficient (CC_a) between the measured and the estimated trajectories calculated as

$$CC_a = \left| \frac{\sum_{t=1}^{T} (\boldsymbol{\gamma}_a(t) - \widehat{E}[\boldsymbol{\gamma}_a]) * (\widehat{\boldsymbol{\gamma}}_a(t) - \widehat{E}[\widehat{\boldsymbol{\gamma}}_a])}{\sqrt{\sum_{t=1}^{T} (\boldsymbol{\gamma}_a(t) - \widehat{E}[\boldsymbol{\gamma}_a])^2 * \sum_{t=1}^{T} (\widehat{\boldsymbol{\gamma}}_a(t) - \widehat{E}[\widehat{\boldsymbol{\gamma}}_a])^2}} \right| \tag{22}$$

where $\widehat{E}[.]$ is the estimate of the expected value, i.e. the sample mean. The mean Correlation Coefficient mCC is calculated by averaging over all articulatory trajectories.

Both these criteria, although used quite often, may not really be effective in determining where or what the error really is. The estimated trajectory may simply be out of phase with the true trajectory, which, depending on the articulator in question, is not as much a problem as making a different trajectory all together. Besides, the error made for different parts of the trajectory (for different phonemes) may not be of equal importance. Another issue is that the $RMSE$ error would be lower for smoother trajectories. This means that gestures without much movement (which then are not as important) would be predicted better than gestures with more movements. Most of the drawbacks associated with $RMSE$ are also applicable to CC. Additionally, calculating CC gives no intuitive idea about the location of the error and about how significant the error is. It is generally known that a low $RMSE$ and a high CC is good, but they do not indicate whether the performance of the state-of-the-art systems are good enough for their purpose.

One evaluation method would be to use these estimates in an articulatory synthesis model and see whether the estimates are able to produce intelligible speech. The quality of the sound produced by the synthesizers is however highly dependant on the vocal tract excitation function (or glottal source modeling) (Childers, 1995). Since these factors are unknown, synthesized speech hence may not make a fair comparison when the articulatory features are estimated by other techniques than inversion-by-synthesis.

Another method of evaluating the overall goodness of the estimates is to use the estimated trajectories to enhance speech recognition. Several studies

(Wrench and Richmond, 2000; Zlokarnik, 1993; Stephenson et al., 2000) have shown that *measured* articulatory data improves the performance of speech recognition systems significantly. However, almost none of the studies that tried to enhance speech recognition performances with *estimated* trajectories (or probabilistic models of the estimates) were successful in improving speech recognition significantly (Stephenson et al., 2000; Markov et al., 2006; Neiberg et al., 2009).

Engwall (2006) and Katsamanis et al. (2008) have suggested two alternative evaluation schemes for acoustic-to-articulatory inversion based on a classification task and a weighted $RMSE$, respectively. The first method attempts to determine if the important articulatory features are correctly recovered, while the second gives more importance to errors that were found to be statistically important for a given articulator and phoneme.

The evaluation method proposed in this article relies on the critical points and thus depends on the reliability of the method to obtain the critical points. If the critical points are calculated reliably, then the rest of the trajectory can be obtained by interpolating between the critical points (described later in Section 4.1). However, the estimated critical points may not just be misplaced in position, but may also be misplaced in time. Secondly, a very jittery movement which is able to predict the critical points is not adequate, which means that erroneous insertion of critical points needs to be penalized. Similarly, a smooth prediction may give a high CC and $RMSE$ but may not have enough critical points. Thus the proposed error measure which we call 'Critical Trajectory Error'(CTE) finds the displacement both in space and time, and returns a quantity which gives an indication of how unsynchronized the estimated trajectory is. The units of this error measure is a unit of time, typically seconds or milliseconds.

3.4. Algorithm to Find CTE

Consider the measured trajectory $\boldsymbol{\gamma}_a$ and the estimated trajectory $\widehat{\boldsymbol{\gamma}_a}$

1. Find the measured critical points $[C_p \quad C_t]$ on $\boldsymbol{\gamma}_a$. C has two dimensions, position p (units mm) and time t (units ms). Say there are M critical points.
2. Find the average velocity, ν, of the gesture associated with each critical point m.
   $\forall m$

$$\nu(m) = \frac{\sum_{k=C_t(m-1)}^{C_t(m+1)} |\boldsymbol{\gamma}_a(k) - \boldsymbol{\gamma}_a(k-1)|}{C_t(m+1) - C_t(m-1)} \tag{23}$$

3. Find the estimated critical points $[\widehat{C}_p \quad \widehat{C}_t]$ on $\widehat{\boldsymbol{\gamma}}_a$.
   Say there are N estimated critical points.
4. Initiate N flags, $F(1 \leq n \leq N)$, required to know whether all the estimated critical points find their correspondences.
5. Initialize the total error for the entire trajectory, $tCTE_a = 0$ for articulator a.
6. $\forall m : 1 \leq m \leq M$
   (a) The nearest critical point in the estimated trajectory to the m^{th} critical point in the measured trajectory is found as
   $\widehat{N}_m = \underset{1 \leq n \leq N}{\arg\min} (C_t(m) - \widehat{C}_t(n))^2$
   (b) set $F(\widehat{N}_m)$ to indicate that the critical point in the estimated trajectory has a correspondence in the measured trajectory
   (c)
   $$tCTE_a = tCTE_a + \left(\left(\frac{C_p(m) - \widehat{C}_p(\widehat{N}_m)}{\nu(m)} \right)^2 + (C_t(m) - \widehat{C}_t \widehat{N}_m))^2 \right)^{1/2} \tag{24}$$
   This equation adds to the total error, $tCTE_a$, the Euclidean distance caused by the displacement of the estimated critical point in the time-position, 2-D space. In order the normalize between the two axes, i.e. time and position, the displacement in position is normalized by the velocity. Thus importance to the position with respect to time depends on the velocity. The resulting value has the same units as that of time.
7. In order to penalize all the extra critical points in the estimated trajectory without a corresponding critical point in the measured trajectory, $\forall n \in \sim F$, (where $\sim$ is an unset flag),
   (a) The nearest critical point in the measured trajectory to the n^{th} critical in the estimated trajectory is found as
   $\widehat{M}_n = \underset{1 \leq m \leq M}{\arg\min} (C_t(m) - \widehat{C}_t(n))^2$

(b)

$$tCTE_a = tCTE_a + \left(\left(\frac{C_p(\widehat{M}_n) - \widehat{C}_p(n)}{\nu(\widehat{M}_n)} \right)^2 + (C_t(\widehat{M}_n) - \widehat{C}_t(n))^2 \right)^{1/2} \quad (25)$$

This equation adds to the total critical error the displacement at the extra critical point in comparison to the closest critical point in the original trajectory. Thus all the excess estimated critical points are penalized.

8. $tCTE_a$ now contains the total critical error for an articulatory channel a over an utterance. This is averaged over the true number of critical points in the utterance, and the final average critical error is $CTE_a = \frac{tCTE_a}{M}$.

This method weighs the displacement in position error by the inverse of the average speed during the gesture. So if the gesture is very slow, a larger penalty is given to the difference in position, while if the gesture is fast, a lower penalty is given. For missing critical points, the error would be quite large because the closest estimated critical may be highly displaced in time. For inserted critical points, the error is calculated with respect to a closest critical point in the measured trajectory, as shown in Figure 8.

This error measure thus gives a better idea about how well the algorithm performs in terms of how far the estimated trajectory is from being perfectly synchronized with the measured trajectory. The drawback, however, is the reliance on a method to find these critical points.

3.5. Inversion Experiments

The inversion experiments were conducted on the MOCHA-TIMIT database. For comparison with our gesture based method, we follow Toda et al. (2004b), who applied the GMMR technique to frame-based acoustic to articulatory inversion. They used 11 consecutive frames of 24 MFCC coefficients as acoustic parameters and the positions of the articulators corresponding to the central acoustic frame as the articulatory features to be detected. The training samples were corresponding acoustic-articulatory frames from a part of the data. Articulation prediction was made based on every instance of the acoustic data in the testing set. They performed regression based on two methods, namely the Minimum Mean Square Error Estimate (MMSE) and

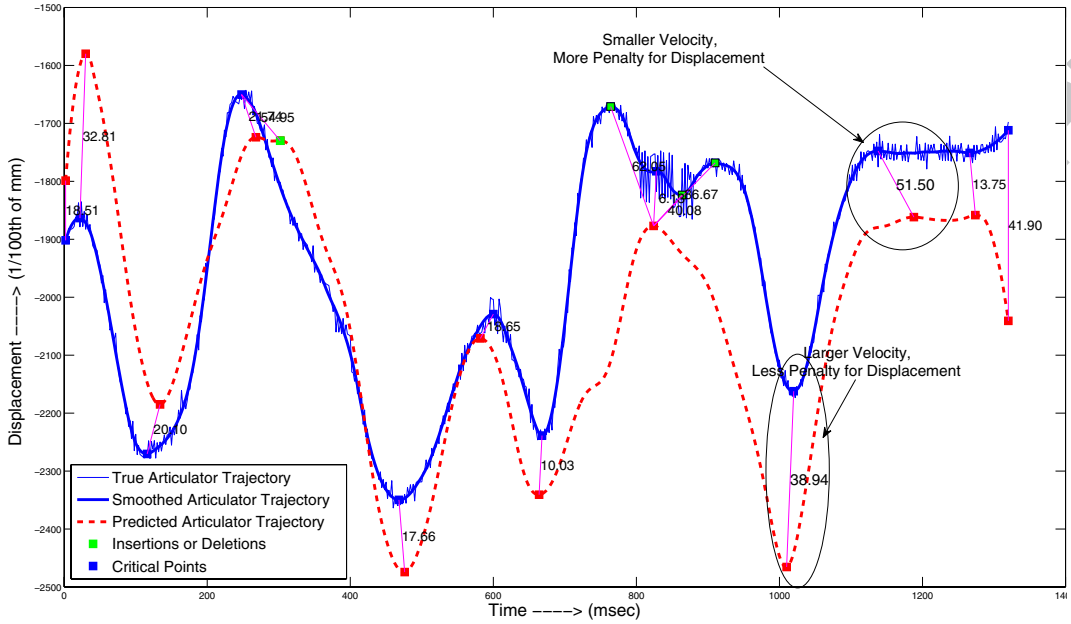


Figure 8: Plot showing how the Critical Trajectory Error (CTE) measures are calculated. One can see that this CTE error measure gives an idea about how unsynchronized the estimate is with respect to the original trajectory. The time scales are in milliseconds.

the Maximum Likelihood Trajectory Estimate (MLTE). The former method simply considered the positions of the articulators, while the latter considered the velocity of the articulators, in order to improve the estimation. We replicated their experiments with as much fidelity as was possible, in order to have a baseline for evaluating the gesture mapping.

3.5.1. Frame-based Inversion

The EMA data was low-pass filtered and down-sampled to 100 Hz, in order to correspond to the acoustic frame shift rate. Each acoustic frame was parameterized by 24 MFCC coefficients (including the 0^{th}), and 11 adjacent acoustic frames each of duration 25 ms (at a frame rate of 100 Hz) were considered. The features were reduced using Principal Component Analysis (PCA) such that all components that contributed to less than 2% of the variation was removed. Thus each acoustic frame had between 64 to 69 (different for each cross-validation set) acoustic features and contained information from 125 ms of the signal. The delta features for the articulatory measurements were

also computed with a look-ahead and lag of 30 ms for the MLTE estimation, giving 28 articulatory features corresponding to the central frame of the acoustic features. A ten-fold cross-validation was performed where 90% (314 sentences, around 94,100 data-frames) of each of the speaker data was used for training the GMMR models and 10% (46 sentences, around 10,400 data-frames) of the same speaker data was used for testing each speaker model's performance. The MFCC and the articulatory trajectory vectors of the training data were normalized to zero mean with a Standard Deviation (SD) of 1. The parameters were optimized on the male speaker. The number of Gaussians that gave the best results was 64 when using the entire training set. The MMSE and MLTE estimates were then filtered using the cut-off frequencies that were suggested by Toda et al. (2008) for each articulator trajectory.

3.5.2. Gesture based Inversion

For the method proposed in this article, we performed segmentation of the acoustic data into acoustic gestures as described in Section 2.3. Since no information about the articulation would be available while performing inversion, we applied the same segmentation to the articulatory movements and parameterized articulatory movements corresponding to the acoustic gestures. These movements may not have been complete articulatory gestures, but would include parts of one or more consecutive gestures. The acoustic gestures and their corresponding articulatory movements were encoded using the 2D-cepstra and 2D-DCT respectively, as described in Section 2.4. After segmentation, we had an average of around 26,450 samples of acoustic-articulatory pairs for training and an average of around 2,430 pairs for testing. Each acoustic gesture was parameterized by $P \times Q$ quefrency and meti parameters plus the actual duration of the gesture. So with P equal to 18 and Q equal to 3, there would be 55 parameters (18×3+1).

The ten-fold cross-validation was performed for this method similarly as for the frame-based method. The encoded parameters of the training were normalized to have a zero mean and an SD of 1, before the training the GMMR with 64 Gaussians. The articulatory trajectories were not filtered or down-sampled as was the case in the frame-based method. The test sentences in both the cases were normalized according to the mean and SD, calculated on the training set. All evaluations were performed against the drift-corrected articulatory trajectories at the original sampling rate rather than the down-sampled version of the trajectories.

Three main parameters may influence the results of the gesture-based acoustic-to-articulatory inversion, namely, the level of smoothing for segmenting the acoustic gestures, and the number of 2-D cepstral coefficients for parameterizing the acoustic space along quefrency and meti (P and Q respectively). We assumed that the same number of meti components are sufficient for parameterizing the articulatory trajectories, although in principle this could be another parameter to optimize. In speech recognition, 12 to 20 MFCC are typically considered, and in this study $P \in \{12, 15, 18, 20\}$ were tested. The order for Q should typically be quite small, between 3 and 5. The higher the number of coefficients Q, the lower the compression and the more the variations in the trajectories and noise in the acoustic signals as well as articulatory trajectories are captured. Thus $Q \in \{3, 4, 5\}$ were tested. The sizes of the window for smoothing that were tested were $w_s \in \{30, 40, 50\}$ ms (*c.f.* Equation 3). We conducted a $4 \times 3 \times 3$ grid search over the possible parameter choices.

4. Results and Discussion

4.1. Detecting Critical Points and Gestures

Table 1 shows the results for detecting acoustic gestures on the TIMIT database. Most of the deletion errors occurred when the phoneme duration was less than 10 milliseconds, while most of the insertion errors occurred for phonemes with durations longer than 200 ms. Several diphthongs and aspirated stops (/p, t, k/) had insertions too.

Window Length (w_s)	Accuracy (%)	Insertions (%)
30 ms	83.16	23.98
40 ms	76.5	11.67
50 ms	69.48	5.91

Table 1: Performance of the acoustic gesture detection algorithm. Accuracy indicates how many times at least one critical point was detected within the duration of a phoneme. Insertions denotes how many times a phoneme was segmented into more than one gesture.

The results are comparable with other segmentation methods described in the literature (e.g. Sarkar and Sreenivas (2005), Svendsen and Soong (1987)). However, the focus and evaluation strategies of the two methods were different. While the proposed method tries to find critical points, which

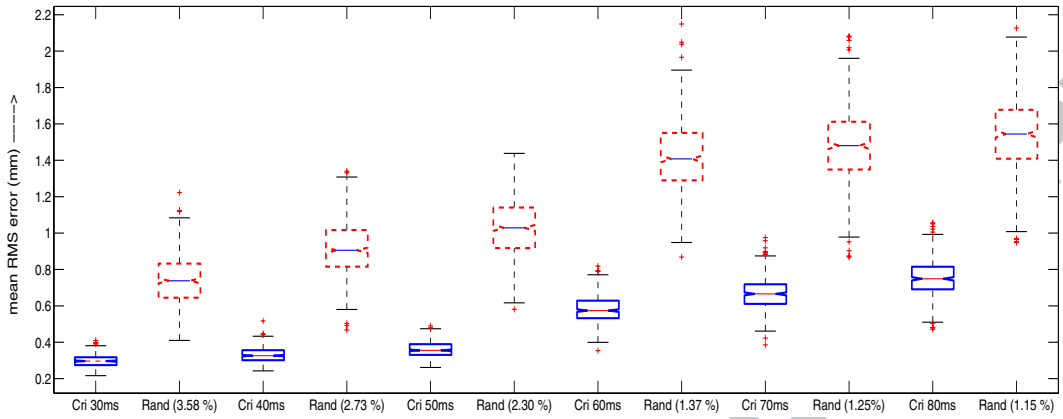


Figure 9: Comparison of the mean RMSE (mm) of the trajectory reconstruction by interpolation using only the critical points. For comparison, the reconstruction error when interpolating between the same number of randomly chosen points over the trajectory is also shown. By chosing critical points, more information about the trajectory can be recovered as compared to chosing points randomly on the trajectory.

are likely to occur somewhere in the middle of the segment, the other methods tried to detect phonemic boundaries.

Figure 9 shows that in spite of knowing the positions of only 1 to 4% of the points in the trajectory, i.e. the positions of the critical points, the $RMSE$ for reconstruction is as low as 0.33 mm. This error is in fact lower than the average resolution of the recording system. For comparison, the reconstruction error for interpolating between the same percentage of randomly selected points on the trajectory is also shown. The error for the same percentage sample of the trajectories is much higher, around 0.8 mm to 1.5 mm.

The results described in Table 1 show that the critical point detection algorithm is able to detect phonetically relevant units in the acoustic signals. Figure 10 illustrates that the detected critical points for most of the phonemes lie close to the center of the segmented phonemes. The corresponding acoustic gestures are usually found to be triphones or longer phonetic units, but diphones and monophones are also common, whenever there are insertion errors. The next question is whether these detected gestures can be used for acoustic-to-articulatory inversion, which is investigated in Section 4.3. The results in Figure 9 show that the critical points in the articulator are good representative points along the trajectory. The relevance of

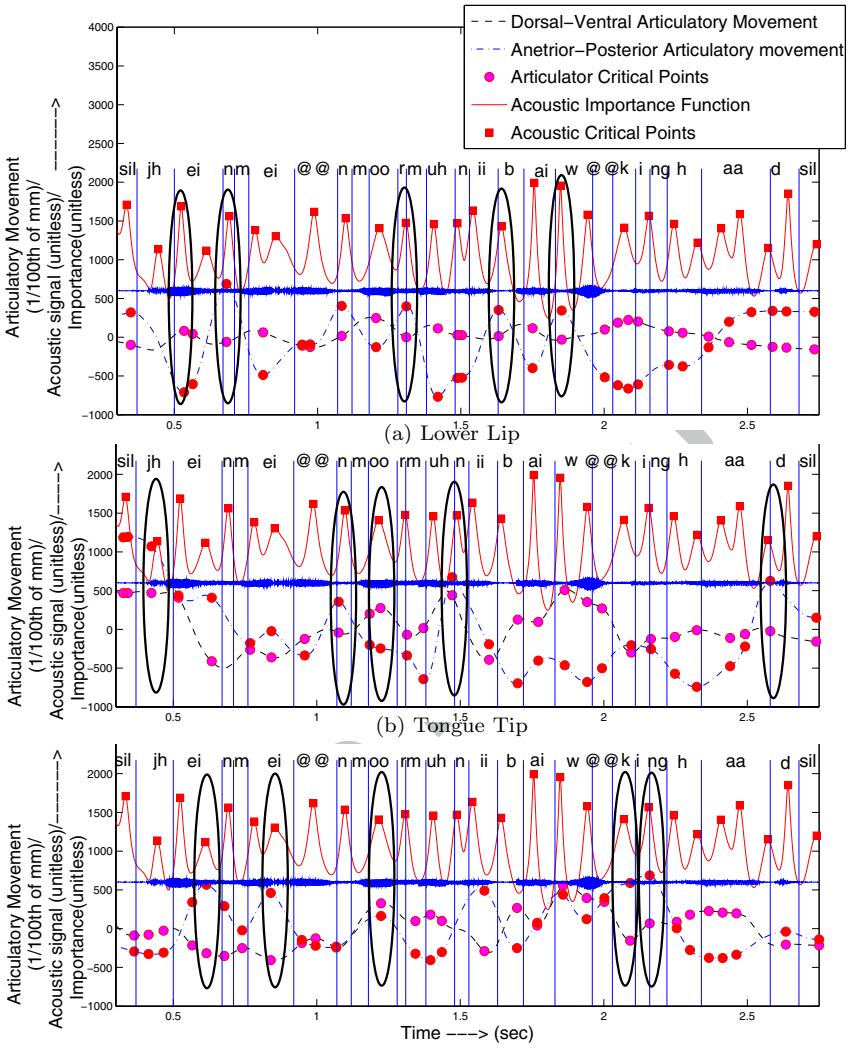
these critical points, and the correspondences of the same in two modalities (i.e. acoustics and articulation) is studied in the following section.

4.2. Relationship Between Acoustic and Articulatory Critical Points

The critical points detected using the acoustic signal and the articulatory trajectories have a very complex relationship. There is a high correlation between the critical points when the particular articulator is important for the acoustics, but low when it is not so, as can be seen in Figure 10. The IPA symbols corresponding to the ASCII characters denoting phoneme labels, displayed on the figure, are detailed in Table 4.

The critical points on the lower lip (LL) are synchronized with the acoustic critical points for the phoneme sequences 'r-m, b and w' (/ɹ-m, b, w/). We see synchronization between the critical points on the tongue tip (TT) and the acoustics for phonemes 'jh and n' and the silence before the 'd' (/dʒ, n, d/). One find synchrony between the tongue dorsum (TD) and acoustics for phonemes 'ei, oo, k and ng' (/eɪ, oː, k, ŋ/). On the other hand, the many critical points are either absent or not synchronized with the acoustic critical point for articulators which are not important for the production. However, all such relationships are not straightforward, since what the important articulators for a certain phoneme are, is not easily known. Many phonemes, have more than one important articulator, while for some phonemes, especially vowels, the important articulator depends on the context. Some other phonemes, such as diphthongs, may have more than one critical point for more than one articulator. In order to study this further, we analyzed the co-occurrences between the articulator critical point and acoustic critical point, within a window of 40 ms. The results are displayed in Figures 11 and 12. The articulatory gestures corresponding to the detected critical points are first parameterized using 2D-DCT and then clustered using k-means clustering. For each critical point in the cluster, the probability that it coincides with an acoustic critical point is calculated. An acoustic critical point is considered coinciding if it falls within 40 ms of a detected articulatory critical point.

From Figure 11*a*, it is clear from the shape of the centroid of the cluster that this cluster represents a sharp opening of the jaw. The phonemes with the highest probability of acoustic-articulatory co-occurrence for the Lower Jaw (LJ) are open vowels and diphthongs. However, some consonants like 'th, g and w' (/θ, g, w/) too have a sharp jaw opening. The second cluster, shown in Figure 11*b* shows a jaw opening along with movement in the horizontal



(a) Lower Lip

(b) Tongue Tip

(c) Tongue Dorsum

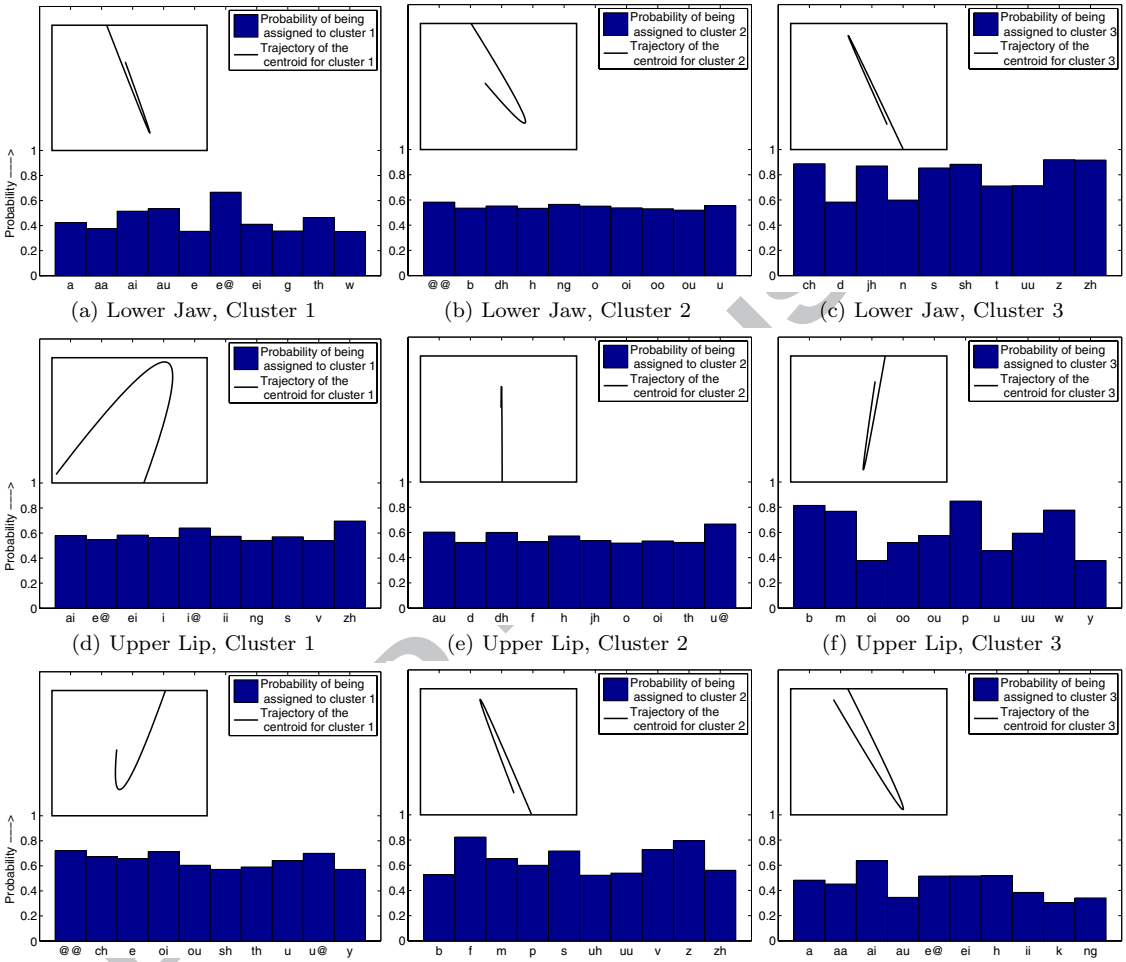
Figure 10: Illustration of the relationship between the critical points in the the acoustic signal and the different articulatory channels for the sentence "Jane may earn more money by working hard". In all figures, the upper part shows the acoustic importance function and critical points and the speech waveform, while the bottom part shows the articulatory trajectories. For the y-axis the scales of the acoustics and articulatory trajectories are not maintained because the illustration indicates the relative changes in acoustics, articulation and the Importance. The vertical lines represent the phoneme boundaries marked by forced aligned annotations. The ellipses indicate the regions where synchrony is observed between the the acoustic and articulatory critical points. In most cases, the critical point in the trajectory of the articulator that is expected to be important for the production of the corresponding sound is synchronous with the acoustic critical point. The critical points for the unimportant articulators may not co-occur with the acoustic critical point.

plane, but not as sharp as in cluster 1. The maximum co-occurrence is for back vowels, diphthongs as some consonants like 'b, dh, h and ng' (/b, ð, h, ŋ/). The third cluster shown in Figure 11*c* is for gestures with a closing of the jaw. These are mostly for fricatives and affricates such as 's, z, sh, zh, ch and jh' (/s, z, ʃ, ʒ, ʧ, ʤ/), where the jaws tend to close in synergy with the tongue tip. What is interesting here, is that although one would consider the opening of the jaw as an important articulatory gesture for open and back vowels, the probability of acoustic-articulatory co-occurrence is lower than for phonemes where the jaw closing occurs. This shows that the timing of the articulatory critical point for jaw opening (which is typically the extreme end of the gesture) is not as important and consistent as that of the jaw closing.

Figure 11*d* shows the cluster where the upper lip (UL) is retracted upwards and inwards. Vowels such as 'i and ii' (/ɪ, iː/) where the lips are drawn sideways are often found to co-occur in the acoustics and articulatory critical points. Besides these, the consonant 'zh' (/ʒ/) is seen to have a high probability of co-occurrence for cluster one with upper lip retraction. However, since there are very few instances of this phoneme, it could be an artifact of the limited context available. The second cluster shown in Figure 11*e* indicated a sharp upper lip raising, with no horizontal movement. There is a mix of vowels, diphthongs and consonants, showing no strong pattern. The diphthong 'u@' (/ʊə/) shows the highest probability of synchrony in this cluster, which is intuitive. The third cluster, with a sharp downward and forward movement, show in Figure 11*f* is dominated by phonemes which are usually attributed to being produced by the lips, such as the bi-labials 'p, b and m' and phonemes such as 'w and uu' (/p, b, m, w, uː/). Here it is clear that the timing of the upper lip lowering is more important and correlated to that of the acoustics than that of the upper lip raising. As expected, the cluster for lip lowering corresponds to the place of articulation of the represented phonemes.

Figures 11*h* to 11*i* show the three clusters for the lower lip (LL) gestures. Cluster 3 for the lower lip has similar phonemes as the cluster 1 for the lower jaw in Figure 11*a*, which shows the synergy between the lower lips and the jaws. Cluster 2, indicating the lower lip raising gesture is dominated by similar phonemes as in the Cluster 3 of the upper lip, i.e., 'b, p, m, uu' (/b, p, m, uː/) etc. as well as the labio-dental fricatives 'f and v' (/f, v/). Again it is clear that the timing for lip closure is more important that lip opening.

Figures 12*a* to 12c represent the correspondences of the three clusters



(a) Lower Jaw, Cluster 1 (b) Lower Jaw, Cluster 2 (c) Lower Jaw, Cluster 3

(d) Upper Lip, Cluster 1 (e) Upper Lip, Cluster 2 (f) Upper Lip, Cluster 3

(g) Lower Lip, Cluster 1 (h) Lower Lip, Cluster 2 (i) Lower Lip, Cluster 3

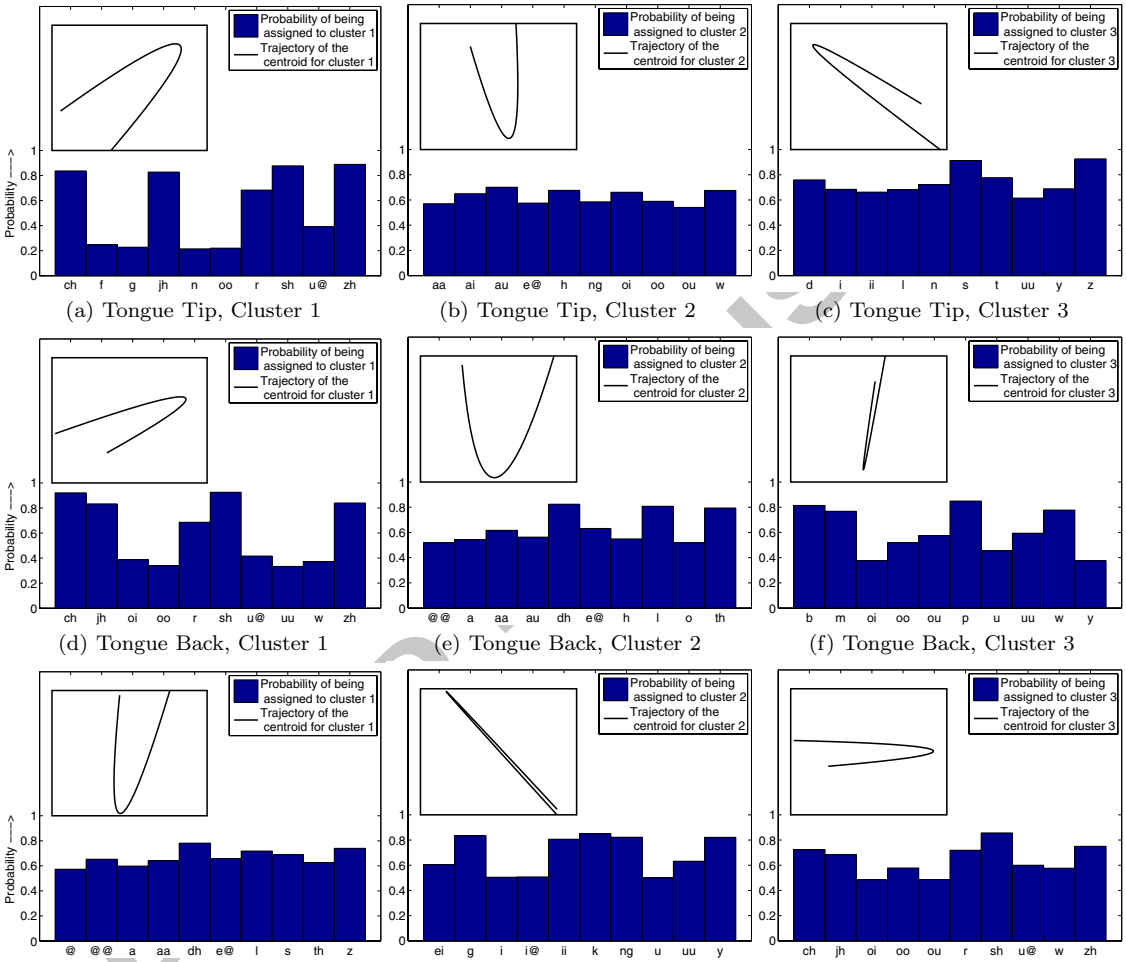
Figure 11: The articulatory gestures corresponding to the critical points are parameterized using 2D-DCT (P=2, Q=3) and then clustered into 3 clusters using k-means clustering. For each cluster, the probability of the articulatory critical point coinciding (falls within 20 ms) with the acoustic critical point, of each phoneme is calculated. Phonemes with the top 10 probabilities of the co-occurrence between the articulatory critical point and the acoustic critical point are displayed. The normalized shape of the mean articulatory gestures along the midsagittal plane for each of the clusters is displayed in the inset. The results shown are for the male speaker, who is recorded facing towards the left hand side.

made by the tongue tip (TT) gestures. The first cluster consists of tongue tip raising gestures, but slightly backwards. These are represented by post-alveolar consonants such as 'r, sh, zh, ch and jh' (/ɹ, ʃ, ʒ, ʧ, ʤ/). The third cluster, consisting of a tongue tip raising slightly forward, is represented by alveolar consonants like 't, d, n, s, z' (/t, l, d, n, s, z/) as well as vowels like 'i and ii' (/ɪ, iː/). Cluster 2, indicating the tongue tip lowering gesture, contains phonemes where the important articulator is likely to be one other than the tongue-tip, but have critical points for the tongue tip due to synergy with either the lips or the jaw. Some open vowels and diphthongs like 'aa, ai and au' (/aː, aɪ, aʊ/) are also represented in the tongue tip lowering cluster.

Cluster 1 of the tongue back (TB) gestures shown in Figure 12*d* nearly duplicates the first cluster for the tongue tip, both in the shape as well as the representative phonemes. Cluster 2 which represents the lowering of the tongue back is dominated by the dental fricative 'th and dh' (/θ, ð/) as well as the lateral 'l' (/l/) and open vowels such as 'a, o and aa' (/æ, ɒ, aː/). The third Cluster of sharp lowering of the tongue back, consists of phonemes where the place of articulation is usually the lips. This may be in synergy with the lips, which is not very obvious initially.

Lowering of the tongue dorsum (TD) is captured by cluster 1, shown in Figure 12*g*, and is important for of open and central vowels such as '@, @@, a and aa' (/ə, ɜː, æ, aː/) as well as alveolar and dental fricatives, 's z, th and dh' (/s, z, θ ð/) and lateral 'l'(/l/). The upward and forward movement of the tongue dorsum is indicated in Figure 12*h* which is dominated by the velar consonants 'g, k and ng' (/ɡ, k, ŋ/) and palatal approximant 'y' (/j/). The third cluster in Figure 12*i* shows a backward movement of the tongue dorsum for post-alveolar phonemes like 'ch, jh, sh, zh and r' (/ʧ, ʤ, ʃ, ʒ, ɹ/), showing a synergy with the front region of the tongue. However, back vowels and diphthongs such as 'oo, oi and ou' (/ɔː, ɔɪ, oʊ/) and the velar approximant 'w' (/w/) also show a backward tongue dorsum gesture, although not as frequently.

In most of the cases, the critical point in the articulatory gesture that is important for the production of the phoneme is co-incident with the acoustic critical point. There are also different types of synchrony and synergy between the different articulators, depending on the gesture involved. For example, the tongue dorsum backward movement is in synergy with the tongue back raising, but the closure or protrusion of the lips is in synergy with the tongue back lowering. This result may not have been obvious by simple correlation analysis because we now observe that the correlation depends on



(a) Tongue Tip, Cluster 1 (b) Tongue Tip, Cluster 2 (c) Tongue Tip, Cluster 3

(d) Tongue Back, Cluster 1 (e) Tongue Back, Cluster 2 (f) Tongue Back, Cluster 3

(g) Tongue Dorsum, Cluster 1 (h) Tongue Dorsum, Cluster 2 (i) Tongue Dorsum, Cluster 3

Figure 12: The articulatory gestures corresponding to the critical points are parameterized using 2D-DCT (P=2, Q=3) and then clustered into 3 clusters using k-means clustering. For each cluster, the probability of the articulatory critical point coinciding (falls within 20 ms) with the acoustic critical point, of each phoneme is calculated. Phonemes with the top 10 probabilities of the co-occurrence between the articulatory critical point and the acoustic critical point are displayed. The normalized shape of the mean articulatory gestures along the midsagittal plane for each of the clusters is displayed in the inset. The results shown are for the male speaker, who is recorded facing towards the left.

the type of gesture involved. The observations indicate that the tongue back lowering gesture is as important as the lip closure for the production of the bi-labials 'b and p' such that the probability of co-incidence between acoustic and articulatory critical points is around the same for both lips and tongue. For each cluster, the probability of the articulatory and acoustic critical point co-occurring is around 80% for several phonemes. But for many clusters and phonemes, the probability is around 50%. Under these circumstances, it is difficult to decide whether the gesture is important for production or not. This, however, indicates that there is a lot of variation, not only in the shape of gestures corresponding to the phoneme, but also in the timing of the gestures. These variations could be attributed to the co-articulation effects of the phoneme occurring in different contexts.

While this study details the issue of the co-occurrences of the acoustic and articulatory critical points as well as the clustering of articulatory gestures, it says little about how the acoustic gestures are related to the articulatory movements. The results from the inversion experiments in the next section, throw some light on this aspect.

4.3. Results from the Inversion Experiments

Table 2 shows the partial optimization table, i.e., the result of variation over one parameter at a time while keeping the other parameters to the optimal ones. The level of smoothing does not affect the performance of the algorithm substantially, but the best performance was for a detection with a balance between number of insertions and deletions (c.f. Table 1). The largest effect is seen by the number of meti parameters Q, as the performance decreases with more than 3 parameters.

Figure 13 compares the performances of the traditional frame based acoustic to articulatory inversion methods with the gesture based method proposed in this article. The Gesture based method shows an $mRMSE$ of 1.45mm (0.63 of Standard Deviation) and 1.55mm (0.64 of Standard Deviation) for the male subject (msak) and female subject (fsew), respectively. The figure shows that there is no statistical difference between the gesture-based method and the frame-based one using dynamical constraints. However there is a statistically significant difference ($p < 0.05$) between the methods using dynamic features and the MMSE based method. This shows that modeling of dynamics of the articulatory trajectories is important for the inversion.

The different methods showed an asynchrony (based on CTE) in the range of 48-50 ms. Earlier research (Reeves and Voelker, 1993) based on

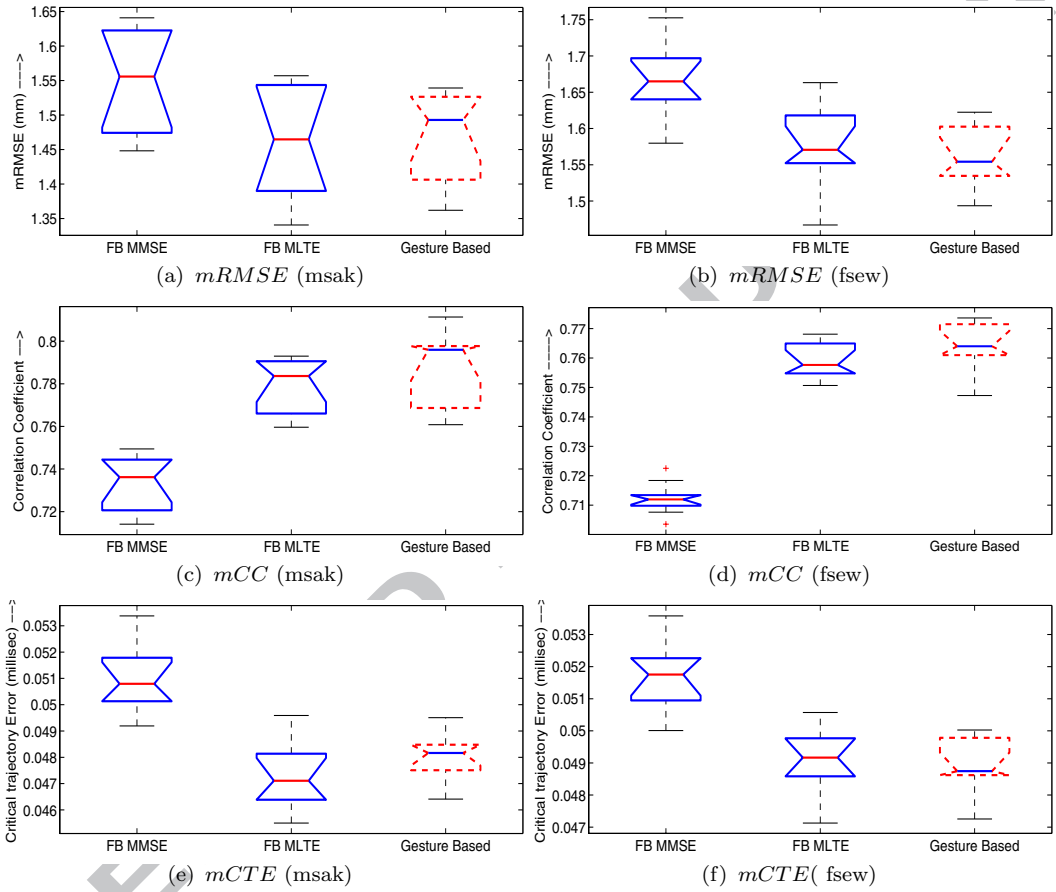
w_s	$RMSE$ (mm)	CC	CTE (msec)
30 ms	1.47	0.78	50.1
40 ms	1.45	0.79	48.4
50 ms	1.49	0.75	51.3
P	$RMSE$	CC	CTE
12	1.49	0.78	50.3
15	1.47	0.79	49.6
18	1.45	0.79	48.4
20	1.46	0.79	49.3
Q	$RMSE$	CC	CTE
3	1.45	0.79	48.4
4	1.5	0.74	50.1
5	1.55	0.71	52.3

Table 2: Table comparing the performance of the proposed method for different window lengths (w_s), number of 'quefrency' components (P) and number of 'meti' components (Q). When one parameter was being optimized, the default setting for the remaining parameters were $w_s = 40ms$, $P = 18$ and $Q = 3$. The results presented are the average over the ten-fold cross-validation on the male speaker (msak) using 64 Gaussian GMMR.

asynchrony between audio and video has shown that an asynchrony of around 40 ms cannot be detected easily by human subjects, but affects their performance in retrieving information from the audio. Thus we can say that the current methods for statistical inversion are close to the point where the error may not be detectable, but will definitely degrade the performance. This effect has been observed in experiments on enhancing speech recognition with estimated articulator trajectories (Wrench and Richmond, 2000; Neiberg et al., 2009). As mentioned in Section 3.3, measured trajectories could enhance the speech recognition accuracy, but the same was not true when estimated trajectories were used.

Figures 14 and 15 show the $RMSE$ estimates from the gesture based inversion algorithm for different phonemes. One can also observe that the largest error in terms of $RMSE$ is for the tongue tip which has the maximum variance among the different articulators, in accordance with previous observations (Richmond, 2002; Toda et al., 2008).

Table 3 shows that that the both the inversion methods are better for vowels, fricatives and diphthongs than for stop-consonants, nasals, liquids and approximants. The performance of the 'Gesture' based method is better



(a) $mRMSE$ (msak) (b) $mRMSE$ (fsew)

(c) mCC (msak) (d) mCC (fsew)

(e) $mCTE$ (msak) (f) $mCTE$(fsew)

Figure 13: Comparisons of the $mRMSE$, correlation coefficients (mCC) and critical trajectory error ($mCTE$) over a ten-fold cross-validation for different methods. The left column is for the male speaker (msak) and the right column is for the female speaker (fsew). The MMSE and MLTE methods are the traditional Frame Based (FB) methods, without and with dynamic features respectively, while the Gesture based method uses the same GMMR regression, but has gesture based features.

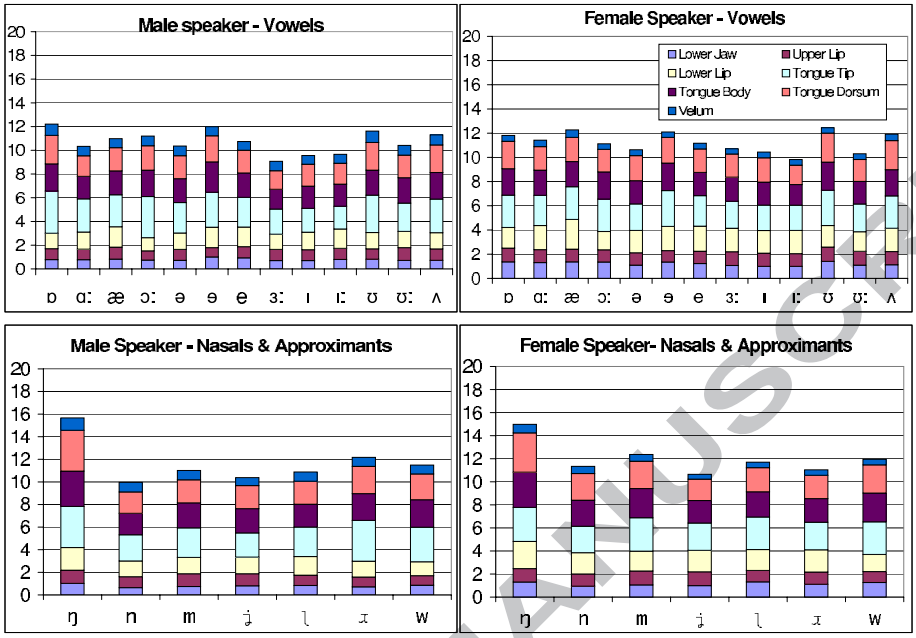


Figure 14: The *RMSE* in mm for individual phonemes and different articulators, from bottom to top: Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (VE)

for fricatives, stop consonants and diphthongs, while the frame-based method performs better for vowels, nasals and liquids. This is probably due to the better modeling of transients like diphthongs and stop consonants by the 'Gesture' based method. Insertion errors can cause problems to the 'Gesture' based method during essentially static segments of speech such as in longer vowels.

5. Conclusions and Future Work

There are three main contributions and insights on two aspects from the paper. The first contribution is a definition for acoustic and articulatory gestures along with a method of unsupervised segmentation of these gestures (or critical point detection) which can be applied in the same way on both the articulatory and acoustic spaces. This draws inspiration from the Direct Realist theory which supposes a direct correspondence between acoustic and articulatory gestures. The gestures are detected based on finding critical points in the trajectory of the articulators or in the acoustic waveform.

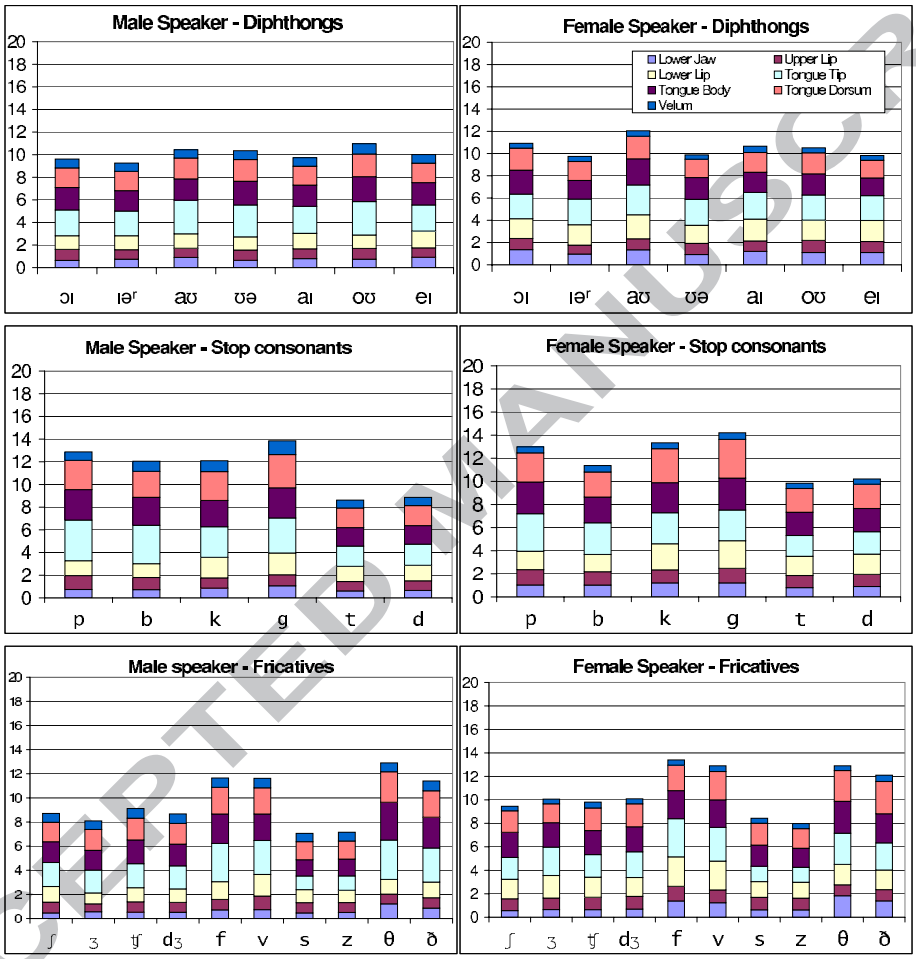


Figure 15: The *RMSE* in mm for individual phonemes and different articulators, from bottom to top: Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (VE)

Phoneme Type	$mRSME$ in mm (Std. Dev.)			
	'Gesture' based		Frame-based	
	male speaker	female speaker	male speaker	female speaker
Vowels	1.53 (0.13)	1.60 (0.11)	1.48 (0.10)	1.55 (0.17)
Diphthongs	1.43 (0.08)	1.50 (0.11)	1.46 (0.09)	1.52 (0.2)
Stop consonants	1.63(0.30)	1.71 (0.25)	1.86 (0.41)	1.77 (0.28)
Nasals and Liquids	1.66 (0.27)	1.71 (0.20)	1.62 (0.21)	1.65 (0.18)
Fricatives	1.38 (0.28)	1.57 (0.3)	1.52 (0.29)	1.65 (0.3)

Table 3: Table comparing the performance of the 'Gesture' based and frame-based algorithms in terms of $mRMSE$ and its standard deviation for different phoneme classes.

Based on these detected gestures, the relationship between the gestures and their timings was studied. Several observations regarding the importance of timing the articulators in synchrony, in order to effect the specific acoustic landmark were made. Studying the mapping between acoustics and articulation in terms of gestures, rather than the instantaneous position or acoustics of one frame, provides insights about two aspects, namely the timing between the acoustics and articulatory parameters and correlations between different articulators, depending on the type of gesture involved. While there seems to be a high correlation between gestures in the articulatory domain and gestures in the acoustic domain, the study also finds a high degree of variability in the types of gestures. This shows that the acoustic-articulatory gesture relationship is more complex than one may initially assume. While this paper does not claim to prove the superiority of the direct realist theory as against other theories of speech perception, it provides a basis for pursuing research in this direction.

The second contribution is the parametrization of acoustic segments using length-independent 2D-cepstral coefficients. This form of parametrization using 2D-DCT is suitable for both acoustics and articulatory trajectories, which is proved by the performing acoustic-to-articulatory inversion based on mapping acoustic gestures to their corresponding articulatory trajectories. The gesture based method follows a different paradigm from the traditional frame-based method. The machine learning algorithm used was exactly the same as the traditional frame-based methods. The only difference was in the types of units used for mapping and their parametrization. The frame-based method made use of every single frame of corresponding acoustic features and articulatory positions for making the inversion, while the Gesture based

method made use of longer segments of acoustics and articulatory movements, thereby reducing the load on the machine learning algorithm. There was a 4-fold reduction in the number of instances used for training in the Gesture based method which correspondingly reduces the training time for the machine learning based regression models.

While the overall performance of the Gesture based method was comparable with the the frame-based method with dynamic features, the performance over different phoneme classes was found to be slightly different in the Gesture based method. The frame-based methods were found to be partial to vowels and liquids which, being longer and static segments, contribute to a larger percentage of the frames in the database. The Gesture based method tries to provide only one sample of correspondences for every occurrence of a phoneme, while modeling the dynamic movement of the phoneme. This method shows a slight preference towards transients.

In spite of different types of parameters selected for the gesture detection and their differences in performance, the inversion results were largely unaffected. This may be attributed to the definition where adjacent gestures overlap with each other. Due to this, small errors in gesture detection may not have a large contribution to the inversion. So in principle, any segmentation algorithm may work equally well for Gesture based inversion as long as there is sufficient overlap between adjacent segments. The minimum jerk smoothing with multiple hypothesis could thus be an important contribution to the overall performance, although it is not easy to speculate on the extent of the importance.

The final contribution from the paper is the critical trajectory error measure CTE which could project the error of the estimation in terms of asynchrony between the trajectories, thus giving a more intuitive idea about the level of errors made. The paper shows that the present error of 48 to 50 ms of asynchrony may not be sufficient to drive oro-facial agents. We propose an error of 40 ms at most in the $mCTE$ for perceptually suitable use in automatic oro-facial simulations.

It would be interesting to see the scalability of the Gesture based method towards speaker adaptation. In addition to different sizes and shapes, different speakers may have different strategies in co-articulation. The Gesture based method may be more suitable to model various co-articulation strategies than the frame-based method.

The proposed method may also be an alternative to traditional short-time stationary (frame-based) approaches towards speech signal processing

in general. Earlier studies (Ananthakrishnan et al., 2009; Neiberg et al., 2008) have shown a non-uniqueness in the mapping between acoustic frames and positions of the articulators in continuous (natural) speech, which may be treated as evidence against the motor-theory of speech perception. It remains to be seen whether this sort of non-uniqueness can be observed even at the gestural level, thus either corroborating or contradicting the direct realist theory.

The gesture based method may be more useful than the frame-based one while driving virtual oro-facial agents (avatars) with articulatory or visual features in cases where the speed of the animation needs to be changed. The different articulatory gestures with varying speeds can be independently controlled quite easily. For example, a gesture corresponding to a particular phoneme may be made slower than others in order to stress on a particular aspect of the utterance.

Comparing the gestures made by the EMA coils with the gestures made by articulatory parameters derived from PCA is another direction that would be interesting.

Future work will also be directed towards implementation of a system which can be used for pronunciation feedback in the form of articulatory gestures.

6. Acknowledgements

This work is supported by the grant 621-2008-4490 from the Swedish Research Council.

References

Ananthakrishnan, G., Neiberg, D., Engwall, O., 2009. In search of Non-uniqueness in the Acoustic-to-Articulatory Mapping, in: Proc. Interspeech, Brighton, UK. pp. 2799 – 2802.

Ananthakrishnan, G., Ranjani, H., Ramakrishnan, A., 2006. Language Independent Automated Segmentation of Speech using Bach scale filter-banks, in: Proc. International Conference on Intelligent Sensing and Information Processing, Bangalore, India. pp. 115–120.

Ariki, Y., Mizuta, S., Nagata, M., Sakai, T., 1989. Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum. IEE Proceedings in Communications, Speech and Vision 136, 133–140.

Atal, S., Chang, J., Mathews, J., Tukey, W., 1978. Inversion of articulatory-to-acoustic information in the vocal tract by a computer-sorting technique. Journal of the Acoustical Society of America 63, 1535–1555.

Bilmes, J., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute 4, 1–13.

Browman, C., Goldstein, L., 1986. Towards an articulatory phonology. Phonology yearbook 3, 219–252.

Childers, D., 1995. Glottal source modeling for voice conversion. Speech Communication 16, 127–138.

Diehl, R., Lotto, A., Holt, L., 2004. Speech perception. Annual Review of Psychology 55, 149–179.

Dusan, S., Deng, L., 2000. Acoustic-to-articulatory inversion using dynamical and phonological constraints, in: Proc. 5th Seminar on Speech Production, Kloster Seeon, Germany. pp. 237–240.

Engwall, O., 2006. Evaluation of speech inversion using an articulatory classifier, in: Proceedings of the 7th International Seminar on Speech Production, Ubatuba-SP, Brazil. pp. 469–476.

Farhat, A., Perennou, G., Andre-Obrecht, R., 1993. A segmental approach versus a centisecond one for automatic phonetic time-alignment, in: Proc. European Conference on Speech Communication and Technology, Berlin, Germany. pp. 657–660.

Fowler, C., 1996. Listeners do hear sounds, not tongues. Journal of the Acoustical Society of America 99, 1730–1741.

Gholampour, I., Nayebi, K., 1998. A new fast algorithm for automatic segmentation of continuous speech, in: Proc. International Conference on Spoken Language Processing, Sydney, Australia. pp. 1555–1558.

Hiroya, S., Honda, M., 2004. Estimation of articulatory movements from speech acoustics. IEEE Trans. Speech and Audio Processing 12, 175–185.

Hoole, P., 1996. Issues in the acquisition, processing, reduction and parameterization of articulographic data. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München 34, 158–173.

Katsamanis, A., Ananthakrishnan, G., Papandreou, G., Maragos, P., Engwall, O., 2008. Audiovisual speech inversion by switching dynamical modeling governed by a hidden Markov process, in: Proc. European Signal Processing Conference, Lausanne, Switzerland.

Keating, P., 1984. Phonetic and phonological representation of stop consonant voicing. Language 60, 286–319.

Kjellström, H., Engwall, O., 2009. Audiovisual-to-articulatory inversion. Speech Communication 51, 195–209.

Liberman, A., Cooper, F., Shankweiler, D., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychological Review 74, 431–461.

Liu, S., 1996. Landmark detection for distinctive feature-based speech recognition. Journal of the Acoustical Society of America 100, 3417–3430.

MacNeilage, P., 1970. Motor control of serial ordering of speech. Psychological Review 77, 182–196.

Maeda, S., 1988. Improved articulatory models. Journal of the Acoustical Society of America 84, S146.

Markov, K., Dang, J., Nakamura, S., 2006. Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. Speech Communication 48, 161 – 175.

McGowan, R., Berger, M., 2009. Acoustic-articulatory mapping in vowels by locally weighted regression. The Journal of the Acoustical Society of America 126, 2011.

Miller, J., 1989. Auditory-perceptual interpretation of the vowel. Journal of the Acoustical Society of America 85, 2114–2134.

Milner, B., Vaseghi, S., 1995. An analysis of cepstral-time matrices for noise and channel robust speech recognition, in: Proc. European Conference on Speech Communication and Technology, Madrid, Spain. pp. 519–522.

Moore, B., Glasberg, B., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. Journal of the Acoustical Society of America 74, 750–753.

Neiberg, D., Ananthakrishnan, G., Blomberg, M., 2009. On acquiring speech production knowledge from articulatory measurements for phoneme recognition, in: Proc. Interspeech, Brighton, UK. pp. 1387–1390.

Neiberg, D., Ananthakrishnan, G., Engwall, O., 2008. The Acoustic to Articulation Mapping: Non-linear or Non-unique?, in: Proc. Interspeech, Brisbane, Australia. pp. 1485–1488.

Ouni, S., Laprie, Y., 2002. Introduction of constraints in an acoustic-to-articulatory inversion method based on a hypercubic articulatory table, in: Proc. International Conference on Spoken Language Processing, Denver, USA. pp. 2301–2304.

Özbek, I.Y., Hasegawa-Johnson, M., Demirekler, M., 2009. Formant Trajectories for Acoustic-to-Articulatory Inversion, in: Proc. Interspeech, Brighton, UK. pp. 2807–2810.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., Jackson, M., 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. The Journal of the Acoustical Society of America 92, 3078–3096.

Perrier, P., Fuchs, S., 2008. Speed-curvature relations in speech production challenge the one-third power law. Journal of Neurophysiology 100, 1171–1183.

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

Reeves, B., Voelker, D., 1993. Effects of audio–video asynchrony on viewers memory, evaluation of content and detection ability. Research Report Prepared for Pixel Instruments, Los Gatos, California, USA .

Richmond, K., 2002. Estimating articulatory parameters from the speech signal. Ph.D. thesis. The Center for Speech Technology Research, Edinburgh.

Richmond, K., 2006. A trajectory mixture density network for the acoustic-articulatory inversion mapping, in: Proc. Interspeech, Pittsburgh, USA. pp. 577–580.

Saito, T., 1998. On the use of F0 features in automatic segmentation for speech synthesis, in: Proc. International Conference on Spoken Language Processing, Sydney, Australia. pp. 2839–2842.

Sarkar, A., Sreenivas, T., 2005. Automatic speech segmentation using average level crossing rate information, in: Proc. International Conference on Acoustics, Speech, and Signal Processing, Philedelphia, USA. pp. 397–400.

Schmidt, R., Zelaznik, H., Hawkins, B., Frank, J., Quinn, J., 1979. Motor-output variability: A theory for the accuracy of rapid motor acts. Psychological Review 86, 415–451.

Seneff, S., Zue, V., 1988. Transcription and alignment of the timit database. TIMIT CD-ROM Documentation .

Stephenson, T.A., Bourlard, H., Bengio, S., Morris, A.C., 2000. Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables, in: Proc. International Conference on Spoken Language Processing, Beijing, China. pp. 951–954.

Stevens, K., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. Journal of the Acoustical Society of America 111, :1872–1891.

Sung, H., 2004. Gaussian mixture regression and classification. Ph.D. thesis. Rice University,Houston.

Svendsen, T., Soong, F., 1987. On the automatic segmentation of speech signals, in: Proc. International Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland. pp. 77–80.

Toda, T., Black, A., Tokuda, K., 2004a. Acoustic-to-articulatory inversion mapping with Gaussian mixture model, in: Proc. Interspeech, Jeju Island, Korea. pp. 1129–1132.

Toda, T., Black, A., Tokuda, K., 2004b. Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis, in: Fifth ISCA Workshop on Speech Synthesis, Pittsburgh, USA. pp. 31–36.

Toda, T., Black, A., Tokuda, K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Communication 50, 215–227.

Toledano, D., Gomez, L., Grande, L., 2003. Automatic phonetic segmentation. IEEE Transactions on Speech and Audio Processing 11, 617–625.

Toutios, A., Margaritis, K., 2003. A rough guide to the acoustic-to-articulatory inversion of speech, in: 6th Hellenic European Conference of Computer Mathematics and its Applications, Athens, Greece. pp. 1–4.

Van Hemert, J., 1991. Automatic segmentation of speech. IEEE Transactions on Signal Processing 39, 1008–1012.

Viviani, P., Terzuolo, C., 1982. Trajectory determines movement dynamics. Neuroscience 7, 431–437.

Wrench, A., 1999. The MOCHA-TIMIT articulatory database. Queen Margaret University College, Tech. Rep .

Wrench, A., Richmond, K., 2000. Continuous speech recognition using articulatory data, in: Proc. International Conference on Spoken Language Processing, Beijing, China. pp. 145–148.

Yehia, H., Rubin, P., Vatikiotis-Bateson, E., 1998. Quantitative association of vocal-tract and facial behavior. Speech Communication 26, 23–43.

Zhang, L., Renals, S., 2008. Acoustic-Articulatory Modeling With the Trajectory HMM. IEEE Signal Processing Letters 15, 245–248.

Zlokarnik, I., 1993. Experiments with an articulatory speech recognizer, in: Proc. European Conference on Speech Communication and Technology, Berlin, Germany. pp. 2215–2218.

Zue, V., Glass, J., Philips, M., Seneff, S., 1989. Acoustic segmentation and phonetic classification in the SUMMIT system, in: Proc. International

Conference on Acoustics, Speech, and Signal Processing, Glasgow, Scotland. pp. 389–392.

A. List of Phonemes

ASCII Symbol	IPA representation	ASCII Symbol	IPA representation
Vowels and Diphthongs			
o	ɒ	e	e
aa	ɑː	@@	ɜː
a	æ	i	ɪ
oo	ɔː	ii	iː
@	ə	u	ʊ
iy	ɘ	uu	uː
uh	ʌ	u@	ʊə
oi	ɔɪ	ai	aɪ
i@	ɪə^{r}	ou	oʊ
ow	aʊ	ei	eɪ
Stop Consonants			
p	p	b	b
t	t	d	d
k	k	g	ɡ
Nasals, approximats and other sonorants			
m	m	l	l
n	n	r	ɹ
ng	ŋ	w	w
y	j		
Fricatives			
f	f	v	v
ch	ʧ	j	ʤ
s	s	z	z
sh	ʃ	zh	ʒ
th	θ	dh	ð
h	h		

Table 4: The list of phonemes used in this study along with the ASCII symbols and corresponding IPA symbols