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

We propose a method for Acoustic-to-Articulatory Inversion based on acoustic and articulatory 'gestures'. A definition for these gestures along with a method to segment the measured articulatory trajectories and the acoustic waveform into gestures is suggested. The gestures are parameterized by 2D DCT and 2D-cepstral coefficients respectively. The Acoustic-to-Articulatory Inversion is performed using a GMM-based regression and the results 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).

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. 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

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

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 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 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 predicting the shape of articulatory trajectories ('gestures') from acoustic segments of speech. We use a database with simultaneous recordings of acoustics and articulatory trajectories to perform a statistically based acoustic-to-articulatory inversion.

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) or X-ray microbeam (Yehia et al., 1998; Toda et al., 2008; Richmond, 2002). 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). Toutios and Margaritis (2003) have reviewed various data-driven methods. 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 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 and articulatory gestures, 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 to perform acoustic-to-articulatory inversion by the following method. The utterance is first segmented into acoustic and articulatory units, which we call 'gestures'. The segmentation algorithm we propose is a general algorithm that can be applied to any time-varying data, so both articulatory and acoustic gestures can be detected by the same method. The detected acoustic gestures are then parameterized using length independent time-frequency 2-D cepstral coefficients, which give a time-frequency representation for these segments. The corresponding movement made by the articulators during this acoustic gesture are also parameterized by the same function, which is a Two Dimensional Discrete Cosine Transform (2D-DCT). The corresponding articulatory and acoustic gestures are then modeled as a joint distribution using the multivariate Gaussian Mixture Model (GMM). The correspondence between the acoustic and articulatory gestures are learned using GMM regression (Sung, 2004) which is used to predict the articulatory

gestures corresponding to unseen acoustic gestures. We evaluate this inversion method using an EMA corpus of simultaneous acoustic and articulatory measurements. We also compare the proposed method against the standard frame-based inversion method with the same machine learning technique. We found that the proposed gesture based method performed as well as the conventional frame-based inversion method.

This article is structured as follows. First our definition of 'gestures' along with the segmentation strategy is described in Section 2. We then evaluate the segmentation strategy with experiments on the detection of gestures in Section 3, before turning to the main focus of the article, namely the relationship between acoustic and articulatory gestures. The description of the parametrization of the gestures is first given in Section 4. The regression technique used in the acoustic-to-articulatory inversion and a method to evaluate the results are outlined in Sections 5 and 6. The regression experiments performed and their results are then presented in Sections 7 and 8, before concluding with a discussion on the findings made in this study.

2 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. 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) and thus, the gestures themselves may not be important and can be retrieved by applying physical 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. 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. Because of the time constraints while uttering a sentence, the true targets may not be reached, and how closely the articulator comes depends on its 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'.

For an utterance with T time samples, let $\gamma_a(t)$ be the vector corresponding to the position of the articulator a at time instant t. The absolute velocity (speed) $d\gamma_a(t)$ is calculated between the positions $\gamma_a(t)$ and $\gamma_a(t-1)$ for all time instants 2 to T. The 'Importance' function, which gives an indication of how close the position is to a target, $I_a(t)$ can be calculated as

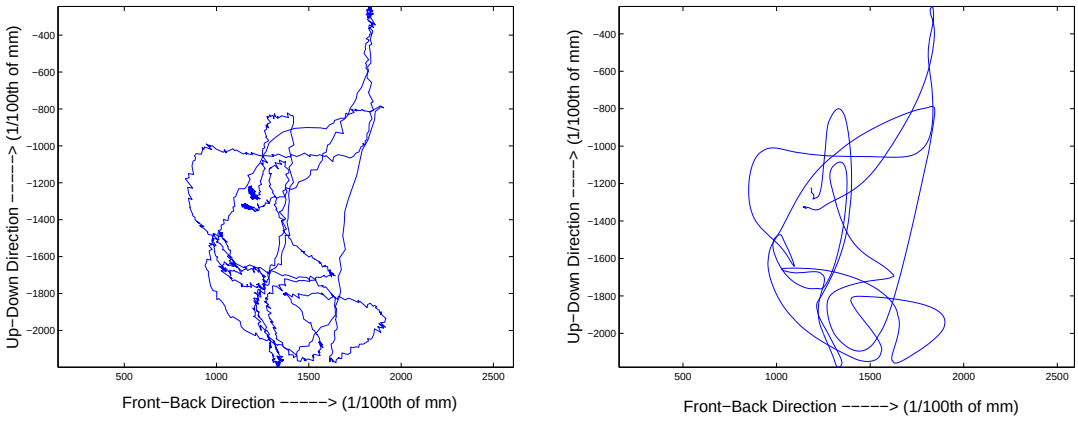
$$I_a(t) = log\left(\frac{\theta_a(t)}{2\pi} - \frac{d\gamma_a(t)}{\max\limits_{1\leq i\leq T} v_a(i)}\right) \tag{2}$$

The angle $\theta_a(t)$ is the acute angle (in radians) between the vectors $[\gamma_a(t-1), \gamma_a(t)]$ and $[\gamma_a(t), \gamma_a(t+1)]$. 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} polynomial equation. Given the noisy (jittery) trajectory of the object $\gamma_a(t)$, a smoothed version $\gamma_{sa}(t)$ can be obtained as

$$\gamma_{sa}(t) = \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \\ t^4 \\ t^5 \end{bmatrix}^T \begin{bmatrix} 1 & \bar{t} & \bar{t}^2 & \bar{t}^3 & \bar{t}^4 & \bar{t}^5 \\ 0 & 1 & 2\bar{t} & 3\bar{t}^2 & 4\bar{t}^3 & 5\bar{t}^4 \\ 0 & 0 & 2 & 6\bar{t} & 12\bar{t}^2 & 20\bar{t}^3 \end{bmatrix}^\dagger \begin{bmatrix} \gamma_a(\bar{t}) \\ d\gamma_a(\bar{t}) \\ d^2\gamma_a(\bar{t}) \end{bmatrix} \tag{3}$$

where † indicates the pseudo-inverse of a matrix and $\bar{t}$ is the vector of time instances from interval $[t - w_s, t + w_s]$, with $2 * w_s + 1$ being the window length. 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, facilitating better detection of 'critical points'. The level of smoothing and thus the number of critical points depends on the window



(a) Original Trajectory (b) Smoothed Trajectory

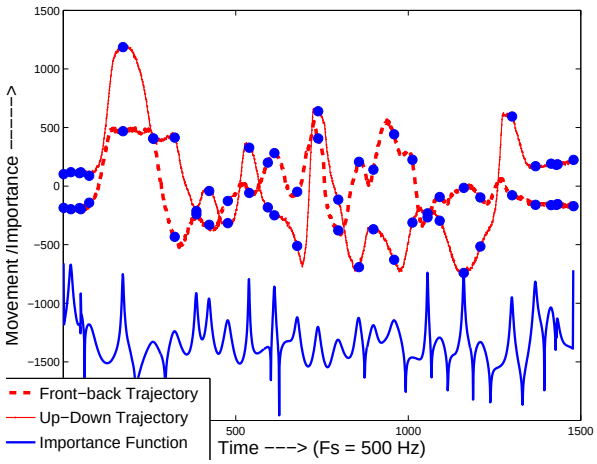
Fig. 1. (a) The original recording of the trajectory of an EMA coil placed on the tongue tip along the mid-sagittal plane during the utterance of the sentence, "Jane may earn more money by working hard". (b) The smoothed version of the same trajectory using minimum jerk smoothing.

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 the trajectory PC of one gesture corresponds to CS for the previous one. One such gesture is shown in Figure 2.

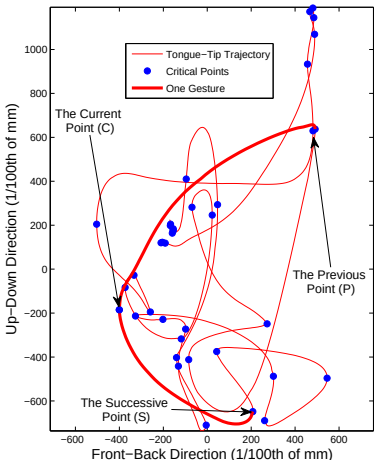
These critical points constitute around 1 to 4% of the trajectory lengths depending on the articulator and the content of the sentence. By performing minimum jerk interpolation between the critical points, the original trajectories can be estimated with a Root Mean Square Error ($RMSE$) of less than 0.4 mm (less than 15% of the standard deviation). The error increases with a larger amount of smoothing. 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.

2.2 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.



(a) Importance Function



(b) Critical Points in 2D

Fig. 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.

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), Mel Frequency Cepstral Coefficients (MFCC) (Toledano et al., 2003) etc. 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

$$X(k,t) = \alpha(k) \sum_{m=1}^{L(k)} W_k(m) x(t-m) \exp\left(\frac{-j2\pi t C_f(k)}{F_s}\right) \tag{4}$$

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 windows function $W_k(t)$ are Finite Impulse Response (FIR) linear phase low 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 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 $W_k(n) : 1 \leq n \leq L(k)$ is calculated as follows.

$$W_k(t) = H(t) * \frac{\sin\left(\frac{(t - (L(k)/2))\, B_W(k)}{F_s}\right)}{\left(t - \frac{L(k)}{2}\right)} \tag{6}$$

where $H(n)$ 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 $X(k,t)$ is then converted to a real signal by finding its absolute value and compressing it using the log scale approximation of loudness, as

$$lX(k,t) = 10 \log_{10}(|X(k,t)|^2) \tag{7}$$

The real signal $lX(k,t)$ is used for further processing. In our experiments the

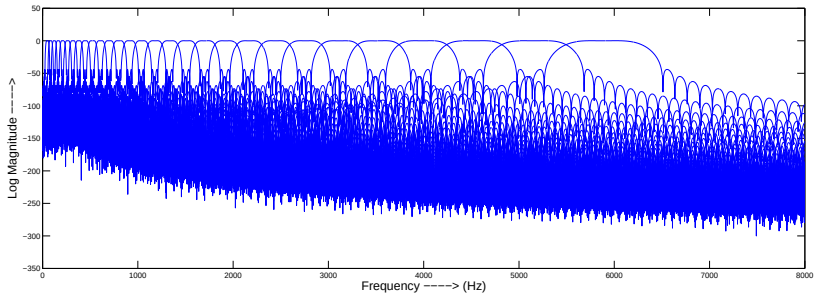


Fig. 3. The frequency magnitude response of the ERB Filter-banks with $B = 80$ Hz and 45 filters. One can see that the sub-band ripple is below 40 dB for all the filters.

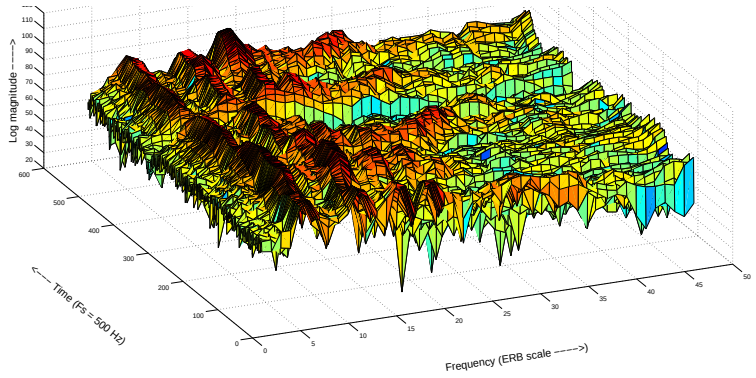
minimum frequency of the filter-bank was 80 Hz, the maximum frequency was less than 6500 Hz and the total number of filters was 45. The configuration was not optimized for the task at hand, but small changes in these numbers did not result in any larger differences in the experimental results.

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.

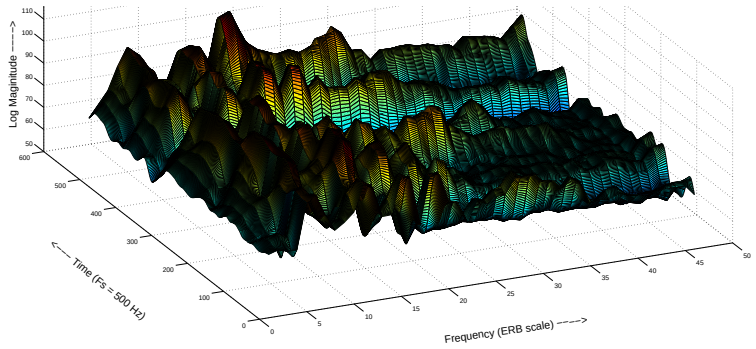
Applying the conditions for detecting the 'critical points' as defined in Section 2.1, we hope to detect the stable regions of the acoustics, which are the projections of the target acoustics onto the true acoustics. Thus this algorithm should be able to predict the salient landmarks in the speech signal. Figure 5 illustrates that the algorithm find critical points for most of the phonemes and that they lie close to the centre of the manually segmented phonemes. A more detailed analysis of the performance of this algorithm is discussed in Section 3.

2.3 Relationship Between Acoustic and Articulatory Gestures

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 5. The critical points on the lower lip (LL) are synchronized with the acoustic ones for the phonemes /m,b,w/. We see synchronization between the critical points on the tongue tip (TT) and the acoustics for phonemes /t,d/ and between the tongue dorsum (TD) and acoustics for phonemes /dʒ,ŋ/.



(a) ERB output



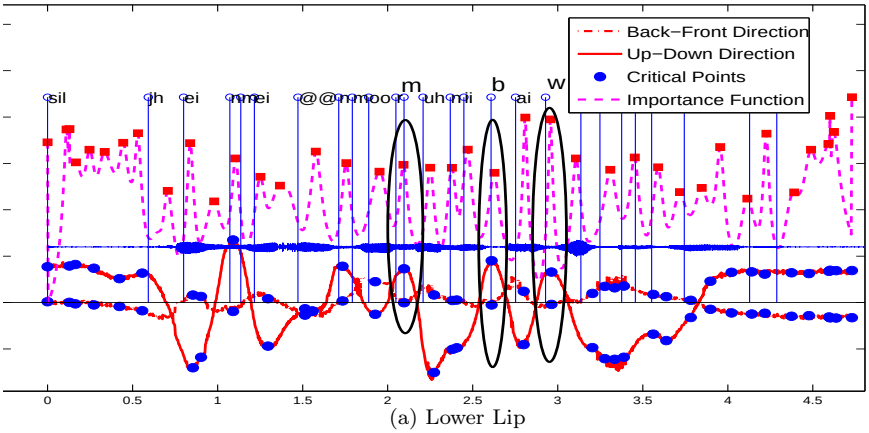
(b) Smoothed ERB output

Fig. 4. A part of the ERB scale log spectrogram from 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.

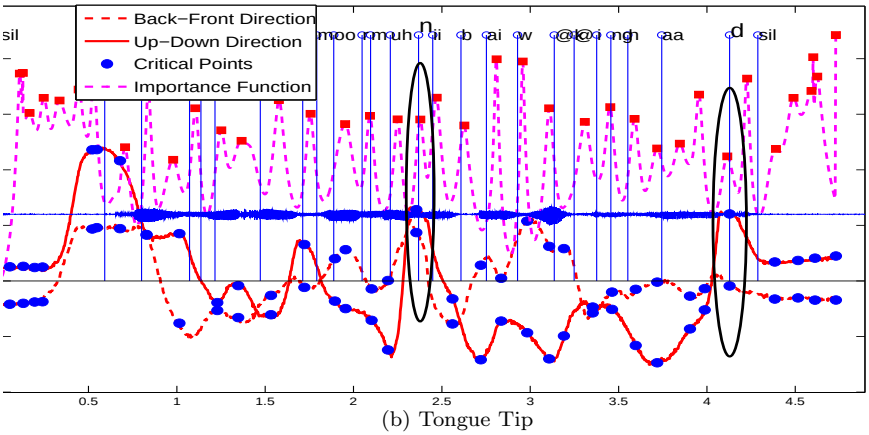
3 Gesture Detection Experiments

A set of experiments was made to estimate the accuracy of the acoustic gesture segmentation (or critical points detection) algorithm. A highly accurately transcribed and aligned data was required, and 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 results are presented directly on the test corpus. Note that these results are not optimized for the purpose of segmentation, which would then be done in the training set.

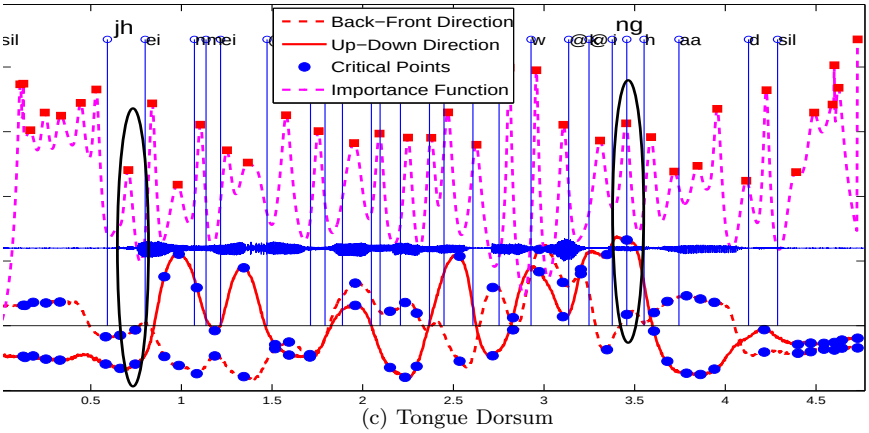
In order to get an estimate of the performance of the gesture detection algorithm, we calculated the number of phonemes that were represented by at least one gesture and the number of phonemes that were represented by more than one gesture. More than one gesture per phoneme may be adequate for diphthongs or aspirated stop consonants, but it was found that some long fricatives and vowels were also broken into more than one acoustic gesture. Table 1 shows the results on the TIMIT database. Most of the deletion errors



(a) Lower Lip



(b) Tongue Tip



(c) Tongue Dorsum

Fig. 5. 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. The x-axis are the time samples at a sampling frequency of 16000 Hz. 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 manual annotations.

occurred when the phoneme duration was less than 10 milliseconds.

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.

By increasing the smoothing (larger w_s), the number of insertions decrease but at the cost of not detecting all the phonemes. These numbers are comparable with most automatic segmentation schemes suggested in the literature which do not rely on extensive training based on orthogonal transcriptions. 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 in acoustic-to-articulatory inversion.

It is more difficult to judge whether the articulatory gestures are detected correctly. The method of evaluation is as follows: we interpolate between the critical points and compare the $RMSE$ between the interpolated trajectories and the measured ones. We used the articulatory measurements from the MOCHA-TIMIT database (Wrench, 1999) in order to evaluate our method. The data is described with further detail in Section 7. Figure 6 shows that in spite of having just 1 to 4% of the points in the trajectory, the $RMSE$ for reconstruction is as low as 0.33 mm. For comparison, the reconstruction error for interpolating between the same percentage of randomly selected points on the trajectory is shown.

These experiments thus show that the gesture detection algorithm is able to detect phonetically relevant units in the acoustic and articulatory signals. It remains to be seen whether these detected segments can be used for acoustic-to-articulatory inversion, which is the focus of the remainder of this article. A time-varying, but length independent, parametrization scheme is needed in order to be able to find a mapping between these acoustic and articulatory gestures. Section 4 describes 2-Dimensional Discrete Cosine Transforms (2D-DCT), a parametrization scheme which can be applied to both acoustic gestures and their corresponding articulatory movements.

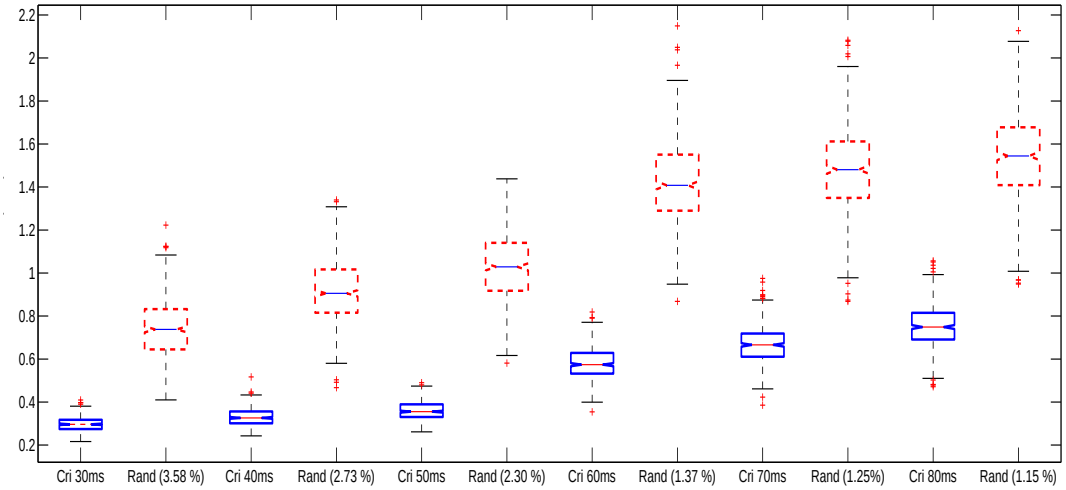


Fig. 6. Comparison of the mean RMSE (mm) of the trajectory reconstruction by interpolation using only the critical points. For comparison, the reconstruction error when interpolating between the same number of randomly chosen points over the trajectory is also shown.

4 2-Dimensional Discrete Cosine Transform

The acoustic and articulatory gestures are parameterized as two-dimensional cepstral coefficients and discrete cosine transforms, respectively. 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 MFCC 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.

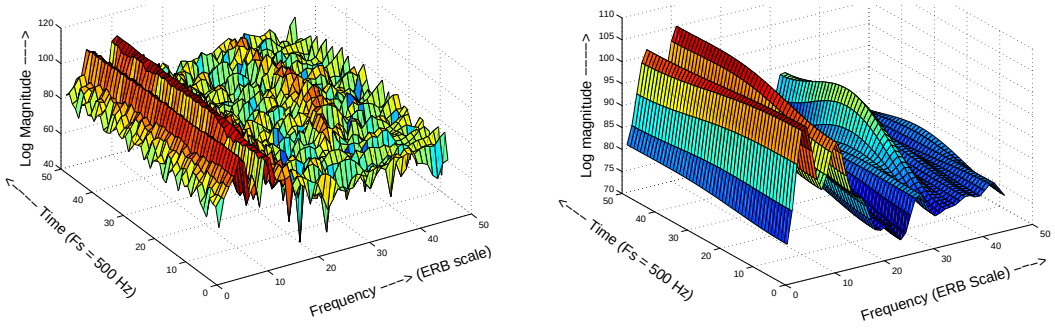
In most previous studies, the 2D-cepstrum was calculated for a fixed duration window. In this study, they are instead calculated for segments of varying duration, since the duration of each gesture could vary greatly, and a length independent representation of the acoustic segment is hence required.

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

$$\tau(p,q) = \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{lX(k,t)}{T} * cos\left(\frac{\pi(k-\frac{1}{2})(p-1)}{K}\right) * cos\left(\frac{\pi(t-\frac{1}{2})(q-1)}{T}\right) \quad (8)$$

where K are 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 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.

In speech recognition, 12 to 20 MFCC are typically considered, and in this study $P = \{12, 15, 18, 20\}$ was 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 are captured. The size of the window for smoothing was $w_s = \{30, 40, 50\}$ ms (*c.f.* Equation 3). 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 7 shows how the original ERB log spectrogram obtained from the output of the ERB filter-banks for a gesture is parameterized as a 2D-cepstrum.



(a) Original Segment (b) Reconstructed Segment

Fig. 7. (a) The original ERB log spectrogram segment of the acoustic gesture during the sequence of phonemes /niːb/ in the context of the words 'money back'. (b) Reconstructed spectrogram segment(from 2D-cepstrum with $P = 18$ and $Q = 3$)

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 parameters and Q was the same as the number chosen for the acoustic gestures. Thus 2D-DCT are a parametrization scheme which is applicable to time-varying segments of both acoustic signals and articulatory trajectories.

5 Regression

In order to evaluate the use of gesture mapping for articulatory inversion, we decided to compare it with the standard frame-based approaches 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. Toda et al. (2004b) applied this technique to acoustic-to-articulatory inversion using 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. Both the acoustics and articulatory data was first sub-sampled to 100 Hz. The training samples were corresponding frames of a part of the acoustic and articulatory 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 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. In our case, we performed a GMMR between the 2D-cepstra of the acoustic gestures and the 2D-DCT encoded articulatory gestures. The GMMR is explained briefly below following the notation used by Toda et al. (2004b).

The conditional probability density of a variable y_t conditioned on variable x_t, modeled as a GMM is represented as

$$P(y_t|x_t) = \sum_{m=1}^{M} P(m|x_t, \lambda) P(y_t|x_t, m, \lambda) \tag{9}$$

where

$$P(m|x_t, \lambda) = \frac{\rho_m \mathcal{N}(x_t; \mu_m^x, \Sigma_m^{xx})}{\sum_{n=1}^{M} \rho_n \mathcal{N}(x_t; \mu_n^x, \Sigma_n^{xx})} \tag{10}$$

and

$$P(y_t|x_t, m, \lambda) = \mathcal{N}(y_t; E_{m,t}^y D_m^{yy}) \tag{11}$$

and λ is the model in the joint space xy. The mean vector $E^y_{m,t}$ and the covariance matrix D^{yy}_m of the conditional probability distribution are

$$E^y_{m,t} = \mu^y_m + \Sigma_m(yx)(\Sigma^{xx}_m)^{-1}(x_t - \mu^x_m) \tag{12}$$

$$D^{yy}_m = \Sigma_m(yy) - \Sigma_m(yx)(\Sigma^{xx}_m)^{-1}\Sigma_m(xy) \tag{13}$$

The MMSE estimate for the regression, $\widehat{y_t}$, given x_t, is calculated as

$$\widehat{y_t} = E[y_t|x_t] = \sum_m^M P(m|x_t, \lambda)E^y_{m,t} \tag{14}$$

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. After regression, the estimates are often smoothed using dynamic information, such as the MLTE employed by Toda et al.(2004b).

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 gesture. Since there is an overlap of trajectory estimates at every critical point due to overlapping gestures (c.f. Section 2 and Figure 8), this overlap in information is handled using a minimum jerk smoothing with multiple weighted hypotheses. In the current implementation, the weights for each hypothesis are set to be equal, but they could be optimized through further experimentation. The minimum jerk smoothing for a time vector $\bar{t}$ 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 following optimization function J.

$$J(\beta_t) = (\Xi - \Gamma * \beta_t)^T * diag(\Phi) * (\Xi - \Gamma * \beta_t) \tag{15}$$

$\beta^{5\times1}_t$ are the parameters of the minimum jerk trajectory. $\Xi^{3*(2*w_s+1)*h\times1}$ is

$$\Xi = \begin{bmatrix} \begin{bmatrix} H_1(\bar{t})^T & dH_1(\bar{t})^T & d^2H_1(\bar{t})^T \end{bmatrix}^T \\ \begin{bmatrix} H_2(\bar{t})^T & dH_2(\bar{t})^T & d^2H_2(\bar{t})^T \end{bmatrix}^T \\ \vdots \\ \begin{bmatrix} H_h(\bar{t})^T & dH_h(\bar{t})^T & d^2H_h(\bar{t})^T \end{bmatrix}^T \end{bmatrix} \tag{16}$$

where $[H_1 \quad H_2 ... \quad H_h]^T$ are the h hypotheses and dH and d^2H denote the

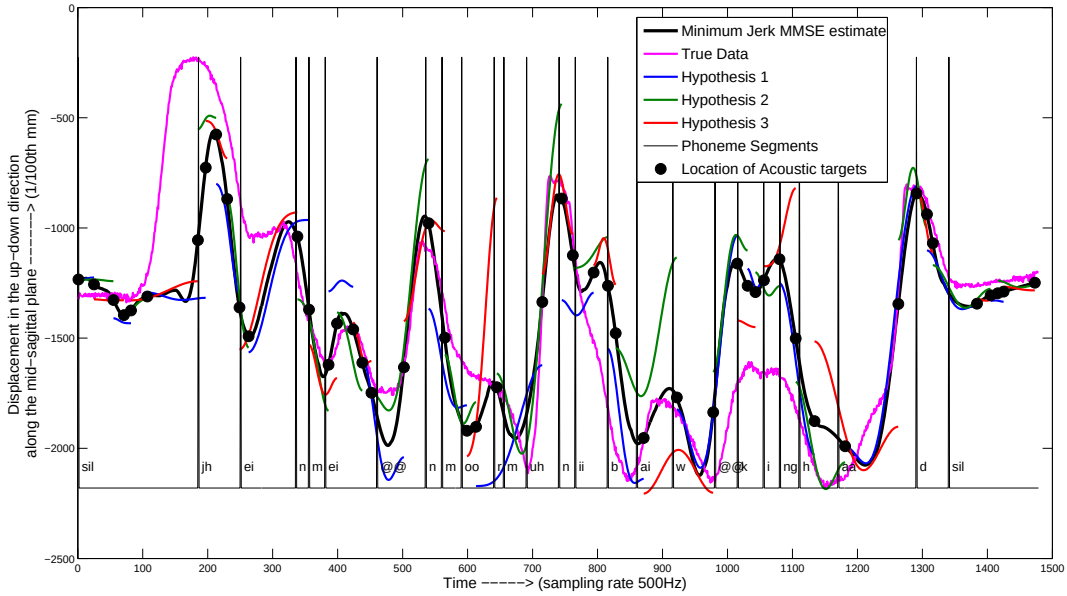


Fig. 8. Multiple hypotheses predicted at every critical point from overlapping gestures made by the tongue-tip (TT) for the sentence, "Jane may earn more money by working hard"

corresponding velocity and the acceleration parameters. $\Gamma^{3*(2*w_s+1)*h\times 6}$ is

$$\Gamma = \begin{bmatrix} 1 & \bar{t} & \bar{t}^2 & \bar{t}^3 & \bar{t}^4 & \bar{t}^5 \\ 0 & 1 & 2\bar{t} & 3\bar{t}^2 & 4\bar{t}^3 & 5\bar{t}^4 \\ 0 & 0 & 2 & 6\bar{t} & 12\bar{t}^2 & 20\bar{t}^3 \\ \text{repeat} & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \bar{t} & \bar{t}^2 & \bar{t}^3 & \bar{t}^4 & \bar{t}^5 \\ 0 & 1 & 2\bar{t} & 3\bar{t}^2 & 4\bar{t}^3 & 5\bar{t}^4 \\ 0 & 0 & 2 & 6\bar{t} & 12\bar{t}^2 & 20\bar{t}^3 \end{bmatrix} \tag{17}$$

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

$$\hat{\gamma}(t) = \begin{bmatrix} 1 & t & t^2 & t^3 & t^4 & t^5 \end{bmatrix} * \beta_t \tag{18}$$

6 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. The mean $RMSE$ ($mRMSE$) is the mean across all the A articulators.

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

$$CC_a = \left| \frac{\sum_{t=1}^{T} (\gamma_a(t) - \widehat{E}[\gamma_a]) * (\widehat{\gamma_a}(t) - \widehat{E}[\widehat{\gamma_a}])}{\sqrt{\sum_{t=1}^{T} (\gamma_a(t) - \widehat{E}[\gamma_a])^2 * \sum_{t=1}^{T} (\widehat{\gamma_a}(t) - \widehat{E}[\widehat{\gamma_a}])^2}} \right| \tag{19}$$

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 is not as much a problem as making a different trajectory. 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, as shown in Section 3. 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 milliseconds.

6.1 Algorithm to Find CTE

Consider the measured trajectory γ_a and the estimated trajectory $\widehat{\gamma_a}$

(1) Find the measured critical points $[C_p \quad C_t]$ on γ_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.
$\forall m$

$$\nu(m) = \frac{(\sum_{k=C_t(m-1)}^{C_t(m+1)} |\gamma_a(k) - \gamma_a(k-1)|)}{(C_t(m+1) - C_t(m-1))} \tag{20}$$

(3) Find the estimated critical points $[\widehat{C}_p \quad \widehat{C}_t]$ on $\widehat{\gamma_a}$.
Say there are N estimated critical points.
(4) Initiate N flags $F(1 \leq n \leq N)$, required to know whether all the critical points find the correspondences.
(5) Initialize the error value, $CTE_a = 0$ for articulator a.

(6) for $\forall m \in C_t$
   (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 $F(N_m)$ to indicate that the critical point in the estimated trajectory has a correspondence in the measured trajectory
   (c)

$$CTE_a = CTE_a + \left(\left(\frac{C_p(m) - \widehat{C}_p(N_m)}{\nu(m)} \right)^2 + (C_t(m) - \widehat{C}_t(N_m))^2 \right)^{1/2} \tag{21}$$

(7) In order to find all the 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,
   $CTE_a = CTE_a + |\widehat{C}_p(n) - \gamma_a(\widehat{C}_t(n))|$
(8) The final error for articulatory channel a, is the mean error for each critical point, $CTE_a = CTE_a/M$

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 9.

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.

7 Inversion Experiments

The inversion experiments were conducted using 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 14 articulatory channels consisted 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 trajectories were processed as described by Richmond (2002) in order to remove the drift.

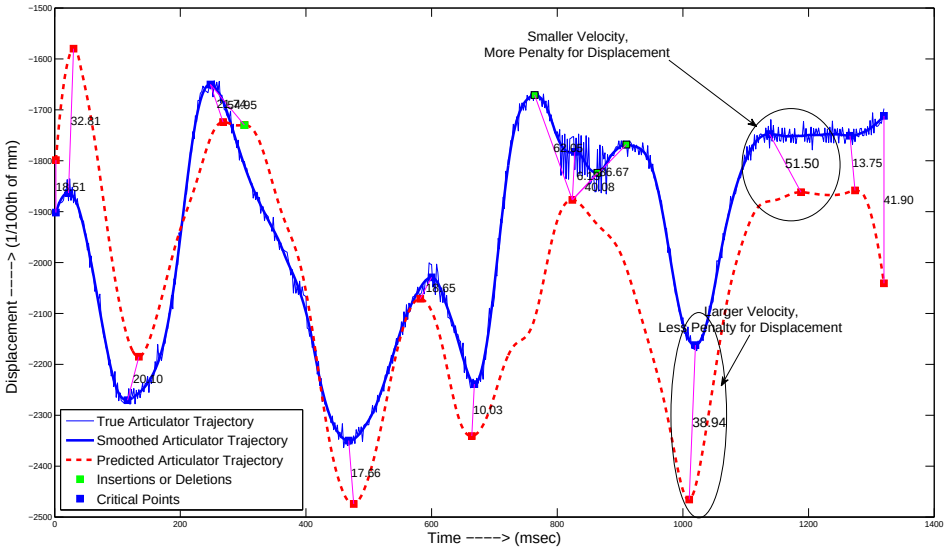


Fig. 9. Plot showing how the Critical Trajectory Error (CTE) measures are calculated. One can see that this CTE error measure gives an idea about how unsynchronized the estimate is with respect to the original trajectory. The time scales are in milliseconds.

For the baseline method, the EMA data was low-pass filtered and downsampled 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.

For the method proposed in this article, the segmentation into acoustic gestures and their corresponding articulatory movements were encoded using the 2D-cepstra and 2D-DCT respectively, as described in Section 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 further processed trajectories.

8 Results

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. 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. We conducted a $3 \times 4 \times 3$ grid search over the possible parameter choices. 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 10 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

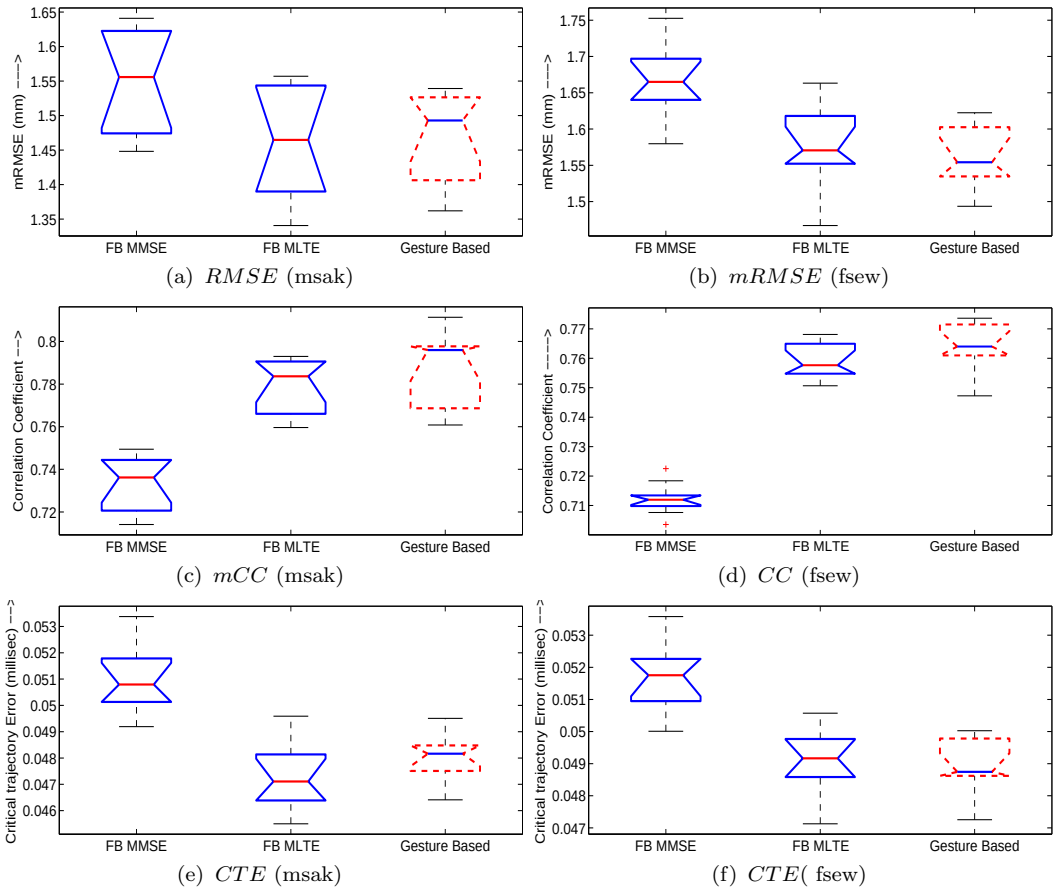
w_s	$RMSE$ (mm)	CC	CTE (msec)
30 ms	1.47	0.78	50.1
40 ms	1.45	0.79	48.4
50 ms	1.49	0.75	51.3
P	$RMSE$	CC	CTE
12	1.49	0.78	50.3
15	1.47	0.79	49.6
18	1.45	0.79	48.4
20	1.46	0.79	49.3
Q	$RMSE$	CC	CTE
3	1.45	0.79	48.4
4	1.5	0.74	50.1
5	1.55	0.71	52.3

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.

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 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). While measured trajectories could enhance the speech recognition accuracy, the same was not true when estimated trajectories were used.

Figure 11 shows the $RMSE$ estimates from the gesture based inversion algorithm for different phonemes. The error is quite balanced for the different classes of phonemes which is different from the usually reported observations (Richmond, 2002; Hiroya and Honda, 2004) that the inversion is better for



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

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

(e) CTE (msak) (f) CTE(fsew)

Fig. 10. 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.

vowels and diphthongs than for stop-consonants, nasals and approximants. This is probably be due to the better modeling of transients by the Gesture based method. 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 which is in accordance with previous observations.

9 Discussion and Conclusions

This article proposes a method of acoustic-to-articulatory inversion based on mapping acoustic gestures to their corresponding articulatory trajectories. The 'Gesture' based method follows a different paradigm than the traditional frame-based method. It draws its inspiration from the Direct Realist theory which supposes a direct correspondence between acoustic and articulatory

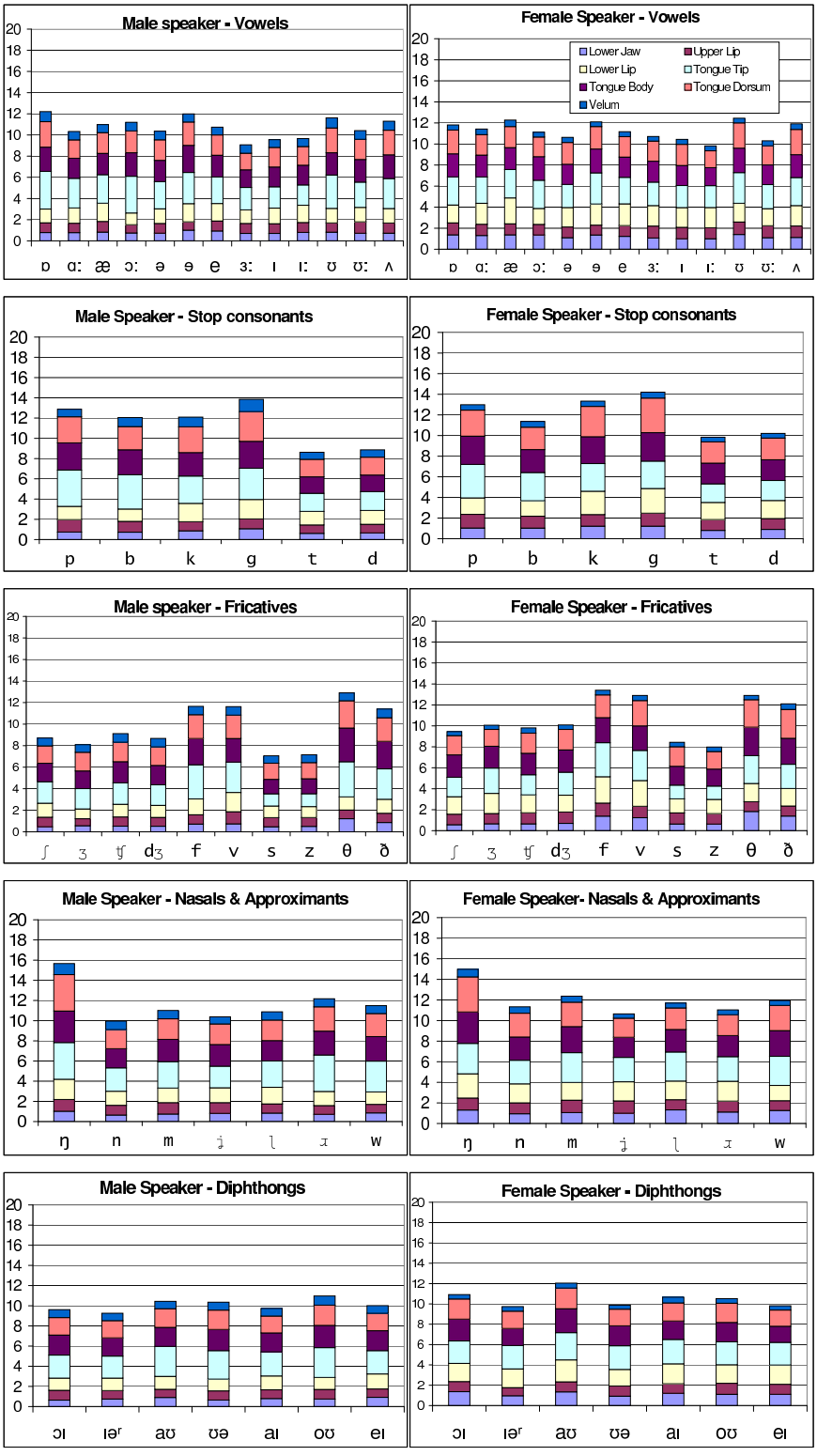


Fig. 11. The *RMSE* in mm for individual phonemes and different articulators, from bottom to top: Lower Jaw (LJ), Upper Lip (UL), Lower Lip (LL), Tongue Tip (TT), Tongue Body (TB), Tongue Dorsum (TD) and Velum (V)

gestures. 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 more or less even in the Gesture based method. The frame-based methods were found to be partial to vowels and diphthongs which, being longer, 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 thereby avoiding any bias towards particular types of phonemes.

In spite of different types of parameters selected for the gesture detection and their differences in performance, the inversion results were more or less 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 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 be an important contribution to the overall performance, although it is not easy to speculate on the extent of the importance.

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.

While this paper does not claim to prove the direct realist theory conclusively as against other theories of speech perception, it provides a basis for pursuing research in this direction. This may be an alternative to traditional short-time stationary (frame-based) approaches towards speech signal processing. 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. 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 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.

There are three main contributions from the paper. The first is a method of unsupervised segmentation of gestures (or critical point detection) which can be applied in the same way on both the articulatory and acoustic spaces. 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. The final contribution 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 also provides insights on two aspects, i.e. the relationship between the critical points in the articulatory and acoustic spaces and also between the gestures. Finally, the paper shows that there is no statistical significance between performing acoustic-to-articulatory inversion using the traditional frame-based method with dynamic constraints on the articulation and the gesture based method using the same machine learning algorithm (GMMR).

Future work will be directed towards speaker adaptation, verifying whether non-uniqueness exists between the mapping of Gestures and implementation of a system which can be used for pronunciation feedback in the form of articulatory gestures.

10 Acknowledgements

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

References

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

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

Ariki, Y., Mizuta, S., Nagata, M. and Sakai, T.: 1989, Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum, *IEEE Proceedings in Communications, Speech and Vision* **136**(2), 133–140.

Atal, S., Chang, J., Mathews, J. and 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. and Goldstein, L.: 1986, Towards an articulatory phonology, *Phonology yearbook* **3**, 219–252.

Childers, D.: 1995, Glottal source modeling for voice conversion, *Speech Communication* **16**(2), 127–138.

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

Engwall, O.: 2006, Evaluation of speech inversion using an articulatory classifier, *Proceedings of the* 7^{th} *International Seminar on Speech Production*, pp. 469–476.

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

Fowler, C.: 1996, Listeners do hear sounds, not tongues, *Journal of the Acoustical Society of America* **99**(3), 1730–1741.

Gholampour, I. and Nayebi, K.: 1998, A new fast algorithm for automatic segmentation of continuous speech, *Proc. International Conference on Spoken Language Processing*, Vol. 4, pp. 1555–1558.

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

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

Keating, P.: 1984, Phonetic and phonological representation of stop consonant voicing, *Language* **60**(2), 286–319.

Kjellström, H. and Engwall, O.: 2009, Audiovisual-to-articulatory inversion, *Speech Communication* **51**(3), 195–209.

Liberman, A., Cooper, F., Shankweiler, D. and Studdert-Kennedy, M.: 1967, Perception of the speech code, *Psychological Review* **74**(6), 431–461.

Liu, S.: 1996, Landmark detection for distinctive feature-based speech recognition, *Journal of the Acoustical Society of America* **100**(5), 3417–3430.

MacNeilage, P.: 1970, Motor control of serial ordering of speech, *Psychological Review* **77**(3), 182–196.

Maeda, S.: 1988, Improved articulatory models, *Journal of the Acoustical Society of America* **84**(S1), S146.

Markov, K., Dang, J. and Nakamura, S.: 2006, Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework, *Speech Communication* **48**(2), 161 – 175.

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

Moore, B. and Glasberg, B.: 1983, Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, *Journal of the Acoustical Society of America* **74**(3), 750–753.

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

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

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

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

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

Reeves, B. and 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*, PhD thesis, PhD thesis, The Center for Speech Technology Research, Edinburgh.

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

Saito, T.: 1998, On the use of F0 features in automatic segmentation for speech synthesis, *Proc. International Conference on Spoken Language Processing*, Vol. 7, Citeseer, pp. 2839–2842.

Sarkar, A. and Sreenivas, T.: 2005, Automatic speech segmentation using average level crossing rate information, *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 397–400.

Schmidt, R., Zelaznik, H., Hawkins, B., Frank, J. and Quinn, J.: 1979, Motor-output variability: A theory for the accuracy of rapid motor acts, *Psychological Review* **86**(5), 415–451.

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

Stephenson, T. A., Bourlard, H., Bengio, S. and Morris, A. C.: 2000, Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables, *Proc. International Conference on Spoken Language Processing*, Vol. 2, 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**(4), :1872–1891.

Sung, H.: 2004, *Gaussian mixture regression and classification*, PhD thesis, Rice University,Houston.

Svendsen, T. and Soong, F.: 1987, On the automatic segmentation of speech signals, *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Vol. 12, pp. 77–80.

Toda, T., Black, A. and Tokuda, K.: 2004a, Acoustic-to-articulatory inversion mapping with Gaussian mixture model, *Proc. International Conference on Spoken Language Processing*, ISCA, pp. 1129–1132.

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

Toda, T., Black, A. and Tokuda, K.: 2008, Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model, *Speech Communication* **50**(3), 215–227.

Toledano, D., Gomez, L. and Grande, L.: 2003, Automatic phonetic segmentation, *IEEE Transactions on Speech and Audio Processing* **11**(6), 617–625.

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

Van Hemert, J.: 1991, Automatic segmentation of speech, *IEEE Transactions on Signal Processing* **39**(4), 1008–1012.

Viviani, P. and Terzuolo, C.: 1982, Trajectory determines movement dynamics, *Neuroscience* **7**(2), 431–437.

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

Wrench, A. and Richmond, K.: 2000, Continuous speech recognition using articulatory data, *Proc. International Conference on Spoken Language Processing*, Vol. 4, Beijing, China, pp. 145–148.

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

Zhang, L. and 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, *Proc. European Conference on Speech Communication and Technology*, Vol. 3, Berlin, pp. 2215–2218.

Zue, V., Glass, J., Philips, M. and Seneff, S.: 1989, Acoustic segmentation and phonetic classification in the SUMMITsystem, *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 389–392.