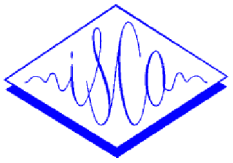
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Statistical and
Perceptual Audio Processing
Brisbane, Australia
September 21, 2008

Data-driven articulatory inversion incorporating articulator priors

Adam Lammert[1], *Daniel P. W. Ellis*[2], *Pierre Divenyi*[1] *

[1] EBIRE, Martinez, CA 94553
[2] Columbia University, New York, NY 10027
{alammert,pdivenyi}@ebire.org, dpwe@ee.columbia.edu

Abstract

Recovering the motions of speech articulators from the acoustic speech signal has a long history, starting from the observation that a simple concatenated tube model is a reasonable model for the origin of formant resonances. In this work, we take a different approach making minimal assumptions about the interdependence of acoustics and articulators by estimating the full joint distribution of the two spaces based on a corpus of paired data, derived from an articulatory synthesizer. This approach allows us to estimate posterior distributions of articulator state as well as finding the maximum-likelihood trajectories. We present examples comparing this approach to a related, earlier approach that did not incorporate prior distributions over articulator space, and demonstrate the advantages of learning the models from realistic utterances. We also indicate benefits available from jointly estimating particular pairs of articulators that have high mutual dependence.

Index Terms: articulatory inversion, speech acoustics

1. Introduction

Over the years, there have been many attempts to recover the motions of speech articulators from the acoustic speech signal. This type of recovery is an instance of the often challenging inverse problems. As such, the task is to infer model parameters from observed data. Early attempts emphasized analytical approaches, most of which sought a unique solution using acoustic models of the vocal tract [1, 2]. However, these approaches were quickly stunted, primarily by the fact that this inverse problem, like many others, suffers from non-uniqueness, and is therefore ill-posed [3]. Specifically, it is possible for a variety of articulator configurations to produce the same acoustic result an obvious challenge for inversion. When given an acoustic signature, there will almost certainly be some ambiguity as to which of several articulator configurations produced it. This non-uniqueness has been demonstrated by many researchers over the last few decades, both from computational approaches [4, 5] and experimental data [6, 7].

Despite the challenges, many researchers have nonetheless forged ahead. Indeed, there is a lot of motivation to provide satisfactory estimates of speech articulators from acoustics; such results would be useful in a variety of applications [8]. Most acknowledge the trouble presented by non-uniqueness, and attempt to overcome those difficulties by using probabilistic approaches. Specifically, a variety of machine learning techniques has been applied in hopes of capturing a mapping, both forward and inverse, between articulator motions and acoustic observations. This kind of research has made extensive use of articulatory codebooks [9, 10, 11], as well as neural networks [12, 13, 14, 15, 16, 17]. Other studies have reported the application of dynamical models [18, 19] and stochastic techniques [8, 20, 21], including Hidden Markov Models [22]. All have seen moderate success, usually with some variability depending on the type of speech sounds being inverted. For instance, vowels and other quasi-stationary portions of speech tend to produce more successful estimations than do transitional sounds, exemplified by stop consonants.

Worth noting, additionally, are some approaches inspired by human language development [23, 24, 25]. Most of these do not explicitly address the inverse problem, but they nonetheless contribute some useful ideas. Indeed, we all know that the inverse problem is solved regularly by human beings in the first years of life, as they learn to speak. Children commonly learn to imitate speech sounds by processing acoustic inputs, manipulating their own articulatory parameters and by drawing a mapping between the two domains. Thus, it must be possible; how can a computer be instructed to do it?

Presently, our approach uses an articulatory codebook. Codebooks are built from large sets of articulatory data with matching acoustic data, representing some sort of relation between the articulatory and acoustic domains. One of the first applications of codebooks to this problem used data gathered from Electromagnetic Articulography (EMA), which had been vector-quantized. A codebook assembled from this data was then used as a simple one-to-one lookup table [9]. However, using the codebook this way ignores the many-to-one mapping which makes articulatory inversion non-unique. A somewhat more successful attempt assumed a many-to-one mapping, by compiling a codebook that exhaustively covered the feasible articulatory space, using a speech synthesizer to construct the acoustic correlates of each configuration. Consequently, multiple articulator configurations could be associated with the same or very similar acoustic realizations. After compiling these options, they used dynamic programming to construct the most likely path through the field of possibilities [11]. This type of approach produced moderately successful results. A similar method was later tried [10], using data also gathered from EMA. However, to show any appreciable improvement over [11], they were required to augment their distance metrics a priori with phonemic information about the utterances.

We propose a new codebook method, which extends some previous ideas and constitutes a generalization of prior attempts. We assume that a very complex mapping may exist, and that no a priori knowledge about the utterance is available. Section 2 describes the theoretical foundation of our approach, and in

* This work was supported by the National Science Foundation (NSF) under Grant No. IIS-0535168 and by the VA Biomedical Laboratory Research. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or the VA.

section 3 we describe our experimental implementation. We discuss the implications of these initial results in section 4.

2. Approach

In this prior work, the most common assumption is of a deterministic acoustic system – which is to say if the articulator positions $\mathbf{A}$ are known, the acoustic observations $\mathbf{O}$ can be directly and unambiguously determined through some (nonlinear) function, $\mathbf{O} = f(\mathbf{A})$. Observations may not completely determine the articulators i.e. there may be several values of $\mathbf{A}$ that result in the same $\mathbf{O}$, but the forward acoustics are unambiguous i.e. there is only one $\mathbf{O}$ for a given $\mathbf{A}$. If, however, we have a system where the articulator state $\mathbf{A}$ is incomplete, then we may have a doubly-ambiguous situation, where a single $\mathbf{A}$ can result in multiple values for $\mathbf{O}$, as well as vice-versa. In this case, a more appropriate way to describe the relationship between articulator state and acoustic observations is as a joint probabilistic distribution $p(\mathbf{A}, \mathbf{O})$, which simply describes the absolute likelihood of any combination of articulator state and acoustic observation. Such a probabilistic description could also incorporate a number of other aspects of the problem, including measurement uncertainty, unmodeled variability in the system, and the *a priori* probabilities of particular acoustics and particular articulatory configurations (independent of each other). Given a suitable approximation to $p(\mathbf{A}, \mathbf{O})$, the inverse acoustics problem of inferring articulators $\mathbf{A}$ from acoustic observations $\mathbf{O}$ amounts to calculating the posterior distribution of the articulators given the observations i.e.:

$$\begin{aligned} p(\mathbf{A}|\mathbf{O}) &= p(\mathbf{A}, \mathbf{O})/p(\mathbf{O}) \\ &= \frac{p(\mathbf{A}, \mathbf{O})}{\int_{\mathbf{A}} p(\mathbf{A}, \mathbf{O}) d\mathbf{A}} \end{aligned} \tag{1}$$

The well-known articulatory ambiguity for given acoustics would emerge as a broad and/or multimodal posterior distribution for the articulators. A time-local model of acoustics and articulators could then be disambiguated by continuity considerations e.g. using dynamic programming to find the best complete path through a sequence of articulator posterior distributions.

This is the approach we take. $p(\mathbf{A}, \mathbf{O})$ should encompass the full range of articulatory and acoustic states anticipated in natural speech, in the appropriate proportions (i.e. with the greatest likelihood for the most common speech sounds and their most common articulatory counterparts). If we had an unlimited database of real speech, along with the true underlying articulator positions, we could simply sample from this database at random, until we had enough points to provide an adequate sampling density in the joint feature space, then perform some kind of local density estimation and smoothing to approximate $p(\mathbf{A}, \mathbf{O})$ – for instance by 'blurring' each distinct articulator-acoustics pair to account for a small range of local values in both domains. This amounts to Parzen estimation [26].

In the absence of a large, representative database of actual articulator and acoustic pairs, we used a forward articulatory speech synthesis model to construct a range of more or less natural and phonetically-balanced sentences (drawn from the Harvard IEEE set). To the extent that this training data includes all the main speech sounds (and articulator configurations), this database should be adequate to model the joint distribution of articulators and acoustics, at least for the specific 'vocal system' being modeled by the synthesizer. Thus, if we define a set of distance vectors between a given point in the joint articulatory-acoustic space $(\mathbf{A}, \mathbf{O})$ and each of our training examples,

$$\Delta_i = \begin{bmatrix} \mathbf{A} - \mathbf{A}_i \\ \mathbf{O} - \mathbf{O}_i \end{bmatrix} \tag{2}$$

where $\{\mathbf{A}_i, \mathbf{O}_i\}$ are our N training patterns, we can approximate our joint distribution,

$$p(\mathbf{A}, \mathbf{O}) = \frac{|\mathbf{W}|}{N} \sum_{i=1}^{N} K\left(\Delta_i^T \mathbf{W} \Delta_i\right) \tag{3}$$

where $\mathbf{W}$ is a positive-definite weighting matrix that defines the 'width' of the Parzen smoothing windows in the articulator and acoustic feature spaces. $K(\cdot)$ is the Parzen window itself, for instance a unit-variance normalized Gaussian, $K(\alpha) = 1/\sqrt{2\pi}\exp\{-\alpha^2/2\}$.

The choice of $\mathbf{W}$ depends on the sparsity of the sample density in each dimension as well as assumptions about the smoothness of the joint distribution along those dimensions. Taking $\mathbf{W}$ as diagonal, a small entry in a particular dimension corresponds to a wide window in that dimension, allowing for density to be interpolated between relatively broadly-spaced samples, but at the same time smoothing out any variation in more densely-sampled regions of the space that occurs at a finer scale. One adaptive approach to this is to vary the effective window width in proportion to the local density – for instance, by finding the k nearest neighbors to a given point, then setting the window width at that point as some fixed factor times the average distance to these neighbors, and performing this separately for each dimension.

In practice, then, we can calculate posterior distributions for articulatory parameters (either as a group, or as subsets in which case unused dimensions are ignored) by taking the acoustic observations $\mathbf{O}$, then retrieving all the training patterns $\{\mathbf{A_i}, \mathbf{O_i}\}$ within the radius of support of the Parzen window K over the acoustic dimensions. Then, for a each value in a grid defined over the possible values for the articulators $\mathbf{A}$, the joint probability $p(\mathbf{A}, \mathbf{O})$ of the actual observation and the hypothesized articulator value is calculated via eqn. 3. Normalizing by the sum over all articulator values gives the posterior probability (according to eqn. 1) — although since the subsequent dynamic programming search is obliged to choose exactly one articulator value for each time step, a common scaling of all likelihood scores at a particular time will not change the optimal choice, and thus the normalization is not required in practice.

In the case of independent estimation of a single articulatory parameter, the result of this is that a sequence of acoustic observation vectors results in a table of joint probabilities with each row corresponding to one of the quantized possible articulator values, and each column corresponding to one time step. We can regard the columns as sets of scaled posteriors for the articulatory parameters, and use dynamic programming to find the most likely sequence of articulatory values by simultaneously applying a continuity constraint as a transition cost that penalizes large jumps in articulator position. Specifically, we estimate the sequence of articulator values $\{\hat{\mathbf{A}}_t\}$ by using dynamic programming to find the sequence that maximizes

$$\prod_t p(\hat{\mathbf{A}}_t|\mathbf{O}_t) q(\hat{\mathbf{A}}_t|\hat{\mathbf{A}}_{t-1}) \tag{4}$$

where $q(\hat{\mathbf{A}}_1|\hat{\mathbf{A}}_0)$ is defined as 1, and

$$q(\hat{\mathbf{A}}_t|\hat{\mathbf{A}}_{t-1}) = \exp\left\{-\frac{1}{2}\left(\left|\hat{\mathbf{A}}_t - \hat{\mathbf{A}}_{t-1}\right|/\sigma\right)^2\right\} \tag{5}$$

for $t > 1$. σ is taken as the 99th percentile of articulator first-order differences seen in the training data, and is calculated separately for increasing and decreasing changes. Joint estimation of multiple articulator dimensions can be performed similarly, for as far as it is practical to track every value in the uniformly-quantized articulator space space. Practically, this has limited us to two dimensions in the current work.

A key aspect of our approach is that we model $p(\mathbf{A}, \mathbf{O})$ (the joint density of articulators and acoustics) instead of $p(\mathbf{O}|\mathbf{A})$ (the distribution of the acoustics given a particular articulator configuration, as used, for example, in [11]). The significance of this is that we are modeling not only the relation between the two spaces, but also the overall likelihood of each configuration. This is particularly important in disambiguating articulatory configurations that may result in similar acoustic observation (i.e. a common articulation $\mathbf{A_1}$ and a much rarer variant $\mathbf{A_2}$ for which $p(\mathbf{O}|\mathbf{A_1}) \approx p(\mathbf{O}|\mathbf{A_2})$), but which may differ greatly in their *a priori* likelihood of occurring in normal speech (e.g. $p(\mathbf{A_1}) \gg p(\mathbf{A_2})$, so $p(\mathbf{A_1}|\mathbf{O}) \gg p(\mathbf{A_2}|\mathbf{O})$). In our experiments, we include a comparison model which illustrates the impact of not including the prior likelihood of each articulation by dividing it out from each joint probability to obtain the conditional i.e.

$$p(\mathbf{O}|\mathbf{A}) = p(\mathbf{A}, \mathbf{O})/p(\mathbf{A}) \quad (6)$$

Finding the articulator state $\mathbf{A}$ that maximizes this objective, $p(\mathbf{O}|\mathbf{A})$, instead of our true objective $p(\mathbf{A}|\mathbf{O})$ from equation 1, amounts to increasing the likelihood of each $\mathbf{A}$ in inverse proportion to its prevalence in the training data i.e. boosting the chances of the rarest articulator configurations, which naturally leads to much less satisfactory results. This is the weakness that arises from the modeling of a uniformly-sampled articulator space as performed by [11]. Note, however, that where they actually evaluated every possible $\mathbf{A}$ within some space of values, for computational ease we approximate this using equation 6, where the prior probability of the articulatory configuration, $p(\mathbf{A})$, is obtained from a histogram of articulator values (over all dimensions, or some subset) over the entire training set.

The remaining variants presented in the experiments use our full joint distribution, and estimate articulator values either individually and independently, or in pairs. When more than one articulator is estimated at the same time, estimation can exploit joint dependence between articulator behavior, which ought to improve performance in the cases where articulators are directly linked, but may be ambiguous when viewed independently. We will discuss our results from this perspective.

3. Experiments

Our codebook is based on synthesized speech obtained from the Task Dynamic Application (TaDA) developed at Haskins Laboratories. TaDA is a MATLAB implementation of the Task Dynamic model of speech articulator coordination [27]. As such, it uses articulator positions as basis functions from which to synthesize speech. There are eight relevant articulator positions: tongue tip constriction degree (TTCD) and location (TTCL), tongue body constriction degree (TBCD) and location (TBCL), lip aperture (LA) and protrusion (PRO), velum (VEL) and glottis (GLO). TaDA simultaneously generates input parameters for the HLSyn speech synthesis software, which synthesizes more natural-sounding speech. The speech output of HLSyn is then transformed into 13 Mel Frequency Cepstral Coefficients (MFCCs) [28], using a 10ms window size and 5ms window advance rate. The articulator positions from TaDA, matched with the MFCCs constitute the final codebook.

The codebook used in our experiments was built from a training data set composed of 40 natural and phonetically-balanced sentences, drawn from the Harvard IEEE Corpus. To input the sentences into TaDA, we used the program's capability to receive orthographic input. From within TaDA, this orthography is then converted into phonemes via a dictionary lookup procedure.

We ran several experiments in order to compare three distinct articulatory inversion methods. The first was an implementation of a previous attempt by Richards [11] and the others were two variants of our own method. Specifically, the first attempted to estimate a single articulator (labeled 1A), and the second sought to estimate two articulators simultaneously (2A). All three methods were tested on two conditions, represented by the same two testing sentences. One sentence was designed to be close to the training sentences (labeled 'Easy'), in order to mimic a larger training set. This sentence contained only words that appeared in the training set sentences. The other sentence was more novel, with only 29% of words in the training set (labeled 'Hard').

The literature on articulatory inversion provides no consensus about how to compare estimated articulator paths with actual paths. Both correlation and geometric distance measures are well represented in the diversity of studies on this topic. We chose to use simple Euclidean distance as a way to compare the actual articulator paths (as determined by TaDA) with the ones estimated by the various inversion methods. For the Richards method, there were a total of 16 distance measurements calculated (8 articulators gathered for 2 testing sentences). The same number of measurements was also calculated for the 1A method. For the 2A method there were more measurements necessary. We continued to track each articulator for each testing word, but also for each accompanying articulator. This brought the total number of measurements to 112 (8 articulators gathered for 2 testing sentences with 7 possible accompanying articulators).

3.1. Results

Tables 1 and 2 summarize the results of the different models in terms of mean squared error in the normalized articulator estimates relative to the ground truth. When looking at the results of the Richards method versus our 1A method, a substantial improvement can be seen. This can be seen even across articulators. The mean Euclidean distance for the Easy testing word improved 23.8% when using the 1A method, as opposed to the Richards method. Surprisingly, an even larger disparity was seen for the Hard testing word, where there was an improvement of 30.0% for the 1A method. Moreover, the 1A results were about the same or better for every articulator, with the exception of lip protrusion, in which 1A lowered performance by 50% on the Easy test sentence only. Conversely, the most dramatic improvement was seen with tongue tip constriction degree which improved by 67.3% on the Easy sentence.

In comparing methods 1A and 2A, an improvement can also be seen. There is some variability in the improvement which depends on the articulator pairs chosen. However, the improvement is evident in the overall case. For the Hard testing sentence, the mean across all articulators (target and accompanying) is 2.8% lower than the 1A mean. This improvement is not evident for the Easy sentence, however, which is 18.4% worse, it would seem. This is mainly due to one accompanying articulator – tongue body constriction degree – which proved to be a major hindrance to the estimation of all target articula-

Table 1: Results of estimating articulator positions by different models: 'Easy' sentence. Best values in each column are in bold.

Alg.	with	GLO	LA	PRO	TBCD	TBCL	TTCD	TTCL	VEL	Mean
Rich	–	38.5	15.9	18.8	37.7	10.8	29.3	17.4	12.9	22.7
1A	–	18.8	7.3	28.2	37.6	8.2	**9.6**	15.9	13.1	17.3
2A	GLO	18.8	7.0	27.7	13.3	8.9	9.9	16.6	12.3	14.3
	LA	14.1	7.3	19.0	13.0	8.6	14.1	10.5	9.3	12.0
	PRO	18.8	8.0	28.2	**12.0**	9.0	12.7	14.3	12.3	14.4
	TBCD	18.8	36.3	40.9	37.6	61.6	38.4	56.6	50.8	42.6
	TBCL	12.1	6.7	21.4	13.4	8.2	14.4	16.0	10.5	12.8
	TTCD	14.4	7.2	**17.9**	14.5	9.3	**9.6**	12.2	8.8	**11.7**
	TTCL	**10.9**	6.7	19.9	14.1	9.5	14.3	15.9	**8.3**	12.4
	VEL	15.6	**6.6**	18.9	12.5	**8.1**	9.9	**9.3**	13.1	11.7

Table 2: Results of estimating articulator positions by different models: 'Hard' sentence. Best values in each column are in bold.

Alg.	with	GLO	LA	PRO	TBCD	TBCL	TTCD	TTCL	VEL	Mean
Rich	–	48.7	37.6	20.8	27.9	25.3	13.1	18.5	11.1	25.4
1A	–	31.1	28.2	19.7	12.2	16.6	11.9	13.4	9.5	17.8
2A	GLO	31.1	27.5	19.7	14.3	16.3	11.5	14.4	9.8	18.1
	LA	28.1	28.2	**19.6**	12.9	13.4	11.1	15.1	**8.5**	17.1
	PRO	31.1	26.8	19.7	16.5	18.2	11.9	12.7	8.8	18.2
	TBCD	27.5	**24.5**	19.7	12.2	14.9	10.2	**11.2**	**8.5**	**16.1**
	TBCL	28.9	26.9	19.7	14.8	16.6	**9.6**	13.4	9.3	17.4
	TTCD	26.8	27.9	19.7	**11.2**	12.6	11.9	13.6	8.7	16.6
	TTCL	**26.7**	28.1	19.7	15.5	**9.7**	12.0	13.4	**8.5**	16.7
	VEL	27.2	27.6	19.7	13.5	14.5	11.5	13.8	9.5	17.2

tors. If one removes these data points, then overall mean for the 2A method shows a 15.1% improvement over the 1A method. The largest improvement of articulator estimation was seen with the pairing of tongue body constriction degree and lip protrusion. This allowed the estimation of the former to be improved 68.1% over the 1A method for the Easy sentence. The largest mean improvement across target articulators was seen with the assistance of tongue tip construction degree as the accompanying articulator, which improved estimation of the articulators by 32.3% for the Easy sentence.

Thus, a complicated picture arises when considering method 2A. It was not a clear win over 1A and, even though the overall picture was positive, the specifics of the picture are mixed. Some articulators seem to be a great help as an accompaniment to other articulators, and some other articulators gain benefits from being accompanied without regard to which articulator. For instance, velum, glottis and lip protrusion seem to be aided by the majority of accompanying articulators. At the same time, tongue tip constriction degree and location and lip aperture appear to help in the estimation of most other articulators. The picture is mixed, though, as the estimation of some articulators was hindered by 2A. To tongue body constriction degree, estimation with almost every other articulator was detrimental, while glottis confused nearly every articulator it was paired with.

Figures 1 and 2 show results for the estimation of individual articulators (tongue body constriction location and lip aperture, respectively) for each of the three methods applied. For all plots, the thin red line represents the actual articulator path, while the thicker blue line represents the estimated articulator values. Behind each plot is shown the local-match scores used to determine the estimated path with the dynamic programming algorithm. These scores can also be thought of as probabilities, at a given instant in time, of the articulator taking on a particular value, given the observed acoustics at that time. The two articulators chosen were jointly estimated for the representation of the 2A case, as indicated. Thus, the 2A results are projections of the 3-dimensional estimation space which has the dimensions LA, TTCL and time. For both articulators, the 2A estimation produced superior results to either of the 1A estimations. Additionally, the 1A estimation was, for both individually, superior to the Richards estimation.

4. Discussion and Conclusions

We have presented an approach to estimating articulator configuration directly from acoustics based on a model of the joint distribution of the two, multidimensional spaces that incorporates both the link between articulators and acoustics, and the prior probabilities in both spaces, based on a paired corpus that is taken to reflect the balance of gestures in real speech. This approach can accommodate arbitrarily complex relationships, including ambiguities (multimodality) in either domain. When ambiguity occurs, valid articulation can still be inferred based both on continuity in articulator space (as enforced by transition constraints through time) and on differing priors among the alternative explanations.

Our transition modeling is somewhat deficient compared to the careful model of joint density: we use a single, global cost function to discourage large excursions in our dynamic programming best-cost paths, rather than, say, attempting to model the actual dynamics present in our training data. Indeed, it would be possible to extend the probabilistic model used at the frame level to obtain a posterior over *sequences* of articula-

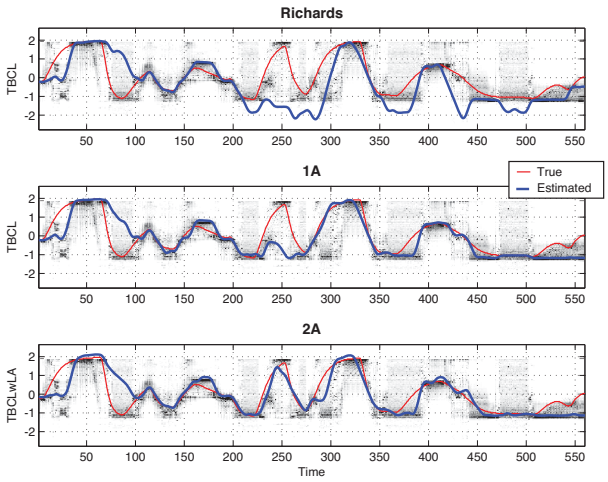


Figure 1: Example estimates for TBCL articulator from the three models. Top: Richards model (no priors). Middle: single-articulator current model. Bottom: TBCL from joint estimation of TBCL and LA. Ground truth (i.e. input to the synthesizer) is shown in each case, as is the underlying score surface input to the dynamic programming.

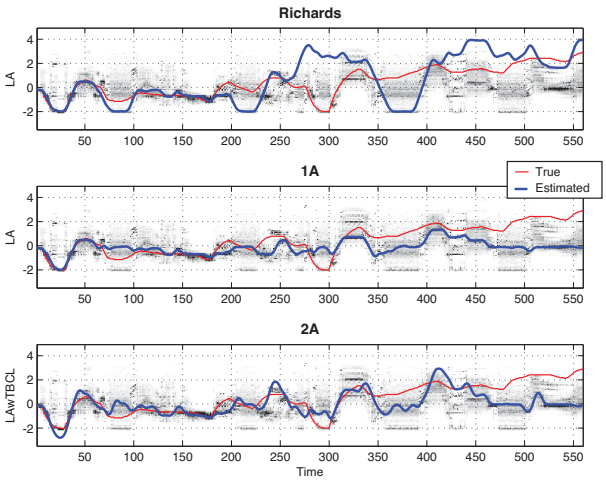


Figure 2: Example estimates for TBCL articulator from the three models as fig 1.

tors given sequences of observations, using the hidden Markov model (HMM). There is a problem here, however: In the conventional exposition of the HMM, local constraints are incorporated as the conditional distribution of observations given state, $p(\mathbf{O}|\mathbf{A})$, not the posterior probabilities of state given acoustics $p(\mathbf{A}|\mathbf{O})$ adopted as the goal here. The HMM then applies the prior of particular state sequences via the state transition costs, $p(\mathbf{A}_t|\mathbf{A}_{t-1})$ which incorporate both the likelihood of a particular transition and, implicitly, the overall likelihood of particular state configurations. One interpretation of our current approach is that we have taken state-specific variations in the transition probabilities and incorporated them in our local match scores, allowing a single, global, normalized transition cost. However, experiments to estimate and model transition behavior more accurately are an important direction for future work.

4.1. Future Work

The most significant challenges we face in the future revolve around moving toward speaker independence. Our results, no matter how much improved, are tied tightly to a specific – and in this case synthetic – speaker. The speaker that is implicitly represented by TaDA was our sole speaker for the experiments described herein. Moreover, the speech generated by TaDA is bound by a phonemic decomposition of the signal, which serves as the input to the synthesizer. Both of these factors mean that our training data, from which we build our codebook, is lacking in variation. Thus, we suspect that our codebook approach would struggle to predict the articulator motions of speech which is variable and substantially different from the TaDA speaker. That is to say, we suspect a challenge in applying our technique to additional speakers and to natural speech. There are several possibilities that could aid in overcoming this shortcoming. One idea is to average across a range of speakers. This could be accomplished by simply introducing variation into the training data in the form of new and different speakers. However, we tend to favor normalization of features into a speaker-independent space.

Of course, much rests on the assumption that we can obtain large amounts of training data in the future. From the standpoint of collecting synthetic speech, this may not pose too much a challenge. Different synthetic speakers can be created by varying the parameters of the synthesizer. Collecting speech-articulator pairing data for natural speech has been a challenge over the years. Several groups are currently working on compiling these data. Notably, the Speech Production and Articulation kNowledge Group (SPAN) is using Magnetic Resonance Imaging to capture data about the very same articulator used in TaDA [29]. This sort of endeavor is absolutely crucial to the future of solving articulator inversion.

Since to much of our methodology stands on the shoulders of the acoustic representation, it may be necessary to explore more sophisticated options for that representation in the future. Although MFCCs seem to perform well, especially in conjunction with geometric measures of similarity, they are by no means the ultimate choice. It is disappointing to note that very few representations of the speech acoustics have attempted to track the spectral changes caused by changes in the articulators, despite the fact that the articulatory functions' primary effect on the acoustics of the speech waveform consists of changing the source filter characteristics. In one of the next incarnations of our model, we will supplement or supplant our MFCC-based frame-by-frame acoustic data with information containing the magnitude and phase of changes in a finite number of frequency prominences, in the hope that such a modification will result in the acoustics being more strongly bound to the articulatory functions.

5. References

[1] P. Mermelstein and M. Schroeder, "Determination of Smoothed Cross-Sectional-Area Functions of the Vocal Tract from Formant Frequencies," *The Journal of the Acoustical Society of America*, vol. 37, p. 1186, 1965.

[2] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 5, pp. 417–427, 1973.

[3] E. Borowski and J. Borwein, *Dictionary of Mathematics*. Harper Collins, 1991.

[4] B. Lindblom, J. Lubker, and T. Gay, "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," *Journal of Phonetics*, vol. 7, pp. 146–161, 1979.

[5] S. Roweis, "Data Driven Production Models for Speech Processing," *Unpublished Ph. D. Thesis, California Institute of Technology, Pasadena, CA*, 1999.

[6] B. Atal, J. Chang, M. Mathews, and J. Tukey, "Inversion of articulatory-to-acoustic tranformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1555, 1978.

[7] J. Flanagan, *Speech Analysis Synthesis and Perception*. Springer, 1972.

[8] S. Dusan and L. Deng, "Estimation of articulatory parameters from speech acoustics by kalman filtering," 1998. [Online]. Available: citeseer.ist.psu.edu/dusan98estimation.html

[9] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *The Journal of the Acoustical Society of America*, vol. 100, pp. 1819–1834, 1996.

[10] T. Okadome, S. Suzuki, and M. Honda, "Recovery of articulatory movements from acoustics with phonemic information," in *Proc. 5th Seminar on Speech Production*, Kloster Seeon, Germany, May 2000, pp. 229–232.

[11] H. Richards, J. Mason, M. Hunt, and J. Bridle, "Deriving articulatory representations from speech with various excitation modes," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, pp. 1233–1236, 1996.

[12] M. Huckvale and I. Howard, "Teaching a Vocal Tract Simulation to Imitate Stop Consonants," in *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.

[13] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *The Journal of the Acoustical Society of America*, vol. 92, pp. 688–700, 1992.

[14] P. Perrier, L. Ma, and Y. Payan, "Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue," in *Proc. Interspeech*, Lisbon, 2005, pp. 1041–1044.

[15] M. Rahim, W. Keijn, J. Schroeter, and C. Goodyear, "Acoustic to articulatory parameter mapping using an assembly of neural networks," *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pp. 485–488, 1991.

[16] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, PhD thesis, Centre for Speech Technology Research, Edinburgh University, 2001.

[17] G. Westerman and E. Miranda, "Modelling the development of mirror neurons for auditory-motor integration," *Journal of New Music Research*, vol. 31, no. 4, pp. 367–375, 2002.

[18] K. Shirai and T. Kobayashi, "Considerations on articulatory dynamics for continuous speech recognition," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8, 1983.

[19] ——, "Estimating articulatory motion from speech wave," *Speech Communication*, vol. 5, no. 2, pp. 159–170, 1986.

[20] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," *Proc. 5th Seminar on Speech Production*, pp. 237–240, May 2000.

[21] S. King and A. Wrench, "Dynamical system modeling of articulator movement," in *Proc. International Congress of Phonetic Sciences*, San Francisco, CA, 1999.

[22] G. Ramsay and L. Deng, "Optimal filtering and smoothing for speech recognition using astochastic target model," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, pp. 1113–1116, 1996.

[23] G. Bailly, "Learning to speak. sensori-motor control of speech movements," *Speech Communication*, vol. 22, no. 2-3, pp. 251–267, 1998. [Online]. Available: citeseer.ist.psu.edu/bailly98learning.html

[24] F. Guenther, "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," *Psychological Review*, vol. 102, no. 3, pp. 594–621, 1995.

[25] K. Markey, "The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development," Ph.D. dissertation, University of Colorado, 1994.

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd Edition*. Wiley, New York, 2001.

[27] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *Acoustical Society of America Journal*, vol. 115, no. 5, pp. 2430–2430, 2001.

[28] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

[29] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, pp. 1771–1776, 2004.