A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model

Sankaran Panchapagesan[a)] and Abeer Alwan

Department of Electrical Engineering

University of California,

Los Angeles,

CA 90095

(Dated: October 5, 2010)

Abstract

In this paper, a quantitative study of acoustic-to-articulatory inversion for vowel speech sounds by analysis-by-synthesis using the Maeda articulatory model is performed. For chain matrix calculation of vocal tract (VT) acoustics, the chain matrix derivatives with respect to area function are calculated and used in a quasi-Newton method for optimizing articulatory trajectories. The cost function includes a distance measure between natural and synthesized first three formants, and parameter regularization and continuity terms. Calibration of the Maeda model to two speakers, one male and one female, from the University of Wisconsin X-ray microbeam (XRMB) database, using a cost function, is discussed. Model adaptation includes scaling the overall VT and the pharyngeal region, and modifying the outer VT outline using measured palate and pharyngeal traces. The inversion optimization is initialized by a fast search of an articulatory codebook, which was pruned using XRMB data to improve inversion results. Good agreement between estimated midsagittal VT outlines and measured XRMB tongue pellet positions was achieved for several vowels and diphthongs for the male speaker, with average pellet-VT outline distances around 0.15 cm, smooth articulatory trajectories, and less than 1% average error in the first three formants.

PACS numbers: 43.70.-h,

Keywords: speech inversion, acoustic-to-articulatory inversion, chain matrix, Maeda articulatory model

I. INTRODUCTION AND REVIEW OF PREVIOUS WORK

Acoustic-to-articulatory inversion or speech inversion is the problem of recovering the vocal tract (VT) shapes that produced a given speech signal. Potential benefits of successful inversion include the use of estimated articulatory parameters for efficient speech coding and improved speech recognition, computer-aided language learning using recovered VT outlines, and improved understanding of speech production, e.g. coarticulation.

Data-driven methods based on Artificial Neural Networks (such as Mixture Density Networks), Kalman filters, Hidden Markov Models and other techniques have become popular in recent years[1–8]. Here, we focus instead on analysis-by-synthesis methods where inversion is performed by adjusting the parameters of an articulatory synthesizer to match acoustic features computed from the input speech[9–12], as shown in the block diagram in Figure 1. Such methods would lead to a better understanding of the speech process and improved speech production models. Discussions of several techniques may be found in [12–14].

The main challenges faced in inversion by analysis-by-synthesis are:

(1) Complexity of speech production models; since the articulatory-to-acoustic or forward mapping in the loop of Figure 1 is computationally expensive, efficient techniques need to be developed for optimizing articulatory parameters.

(2) Inherent non-uniqueness of the inverse mapping, and local optima of the cost function; it is known from perturbation theory that for a lossless acoustic tube, both the poles and zeros of the input impedance are needed to determine the area function. Since only poles (formant frequencies) can be estimated from the speech signal of a vowel, for a theoretical, lossless VT, an infinite number of different VT area functions can result in a given set of formant frequencies[15,16]. Even for a lossy VT, it is a mathematical fact that an infinite number of different area functions exist that can produce the same first few formant frequencies and amplitudes, if the area function space is of

[a)]Electronic address: panchap@ee.ucla.edu

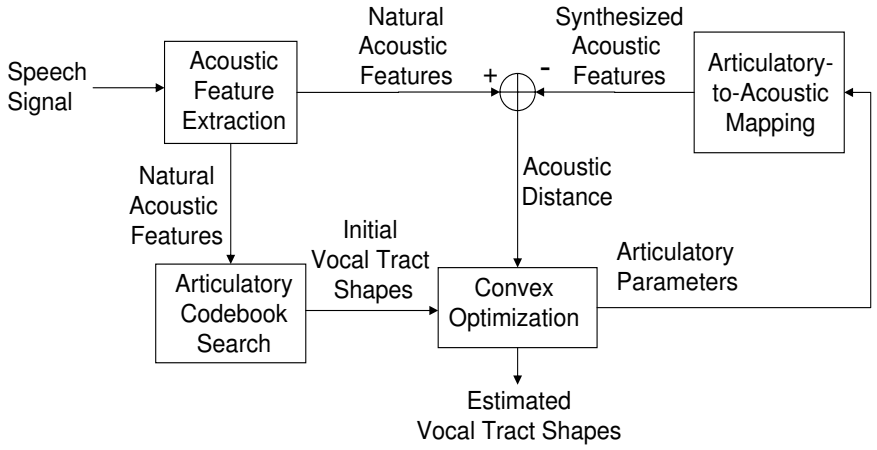


FIG. 1. Acoustic-to-articulatory inversion using analysis-by-synthesis.

higher dimension than the space of the first few formant frequencies and amplitudes[9]. In an empirical study using simultaneously measured acoustic and articulatory X-ray microbeam data (discussed below) of one speaker, it was found that the set of articulatory configurations producing similar acoustics was unimodal/unique for most speech sounds, but multimodal/non-unique for a few (/r/,/l/ and /w/)[17]. Use of articulatory models to constrain the VT area function, regularization and continuity terms in the cost function, and initialization using articulatory codebooks help resolve the non-uniqueness and local optima issues in analysis-by-synthesis[9,11–14,18].

(3) Incomplete knowledge about the shape and dynamics of the vocal tract for a given speaker, and

(4) Insufficient data to learn from or to evaluate inversion results.

The main issues are therefore: choice of acoustic features, the articulatory-to-acoustic mapping, the cost function to be optimized, construction and search of articulatory codebooks to initialize the optimization, the optimization techniques used, and evaluation of results.

The VT resonances are important for characterizing VT acoustics and for perception, and are closely related to the VT shape. For vowels, acoustic distance measures between

natural and synthesized formant frequencies are therefore often minimized[18,19]. Cepstral distance measures are also useful and very flexible[11,20,21] and will be discussed below in Section II.E.

Articulatory models decrease non-uniqueness by constraining the area function to be similar to those from human talkers. The Mermelstein[22] and Maeda[23] models describe the VT midsagittal outline and area function using a relatively small number of parameters (10 for the Mermelstein model and 7 for the Maeda model) which control the shapes and positions of articulators such as the jaw, tongue, lips and larynx.

The non-uniqueness of the inverse solution can also be resolved by including regularization and continuity terms in the optimization cost function[11,13,18,19]. The regularization term is designed to discourage VT configurations farther from the mean or neutral position, and usually takes the form of the sum of squares of articulatory parameters minus their nominal values[18,19]. The continuity term can be the 'geometric' distance from the articulatory parameters of the previous frame in a frame-wise optimization[13], or sum of squares of the first time-derivatives of articulatory parameters over several frames for global optimization over the speech segment[19]. The continuity terms also result in smoother estimated articulatory trajectories, which are desirable since human articulation is controlled by muscles of finite power and therefore human articulatory trajectories are necessarily smooth.

An articulatory codebook is used to initialize the optimization, because of the computationally intensive forward mapping, and local optima of the cost function[9,12–14]. The codebook consists of articulatory vectors and corresponding computed acoustic vectors, and is designed to cover both the articulatory and acoustic spaces with low redundancy. There is hence a trade-off between codebook size and resolutions in articulatory and acoustic spaces. The issues involved in the design and search of codebooks are discussed in greater detail in [13] and [14]. Codebooks specially constructed by dividing articulatory space into hypercubes within which the articulatory-acoustic mapping is approximately linear, have also been used to obtain inverse solutions[19]. Since the cost function includes continuity terms, dynamic programming is used to perform codebook search efficiently[13,19].

Techniques that have been used for more refined optimization of the cost function include direct search methods like the Hooke-Jeeves and coordinate descent methods which do not require the gradient of the cost function[10,13,18], gradient-based methods[11], and iterative solutions of variational equations[19]. A finite difference approximation may be used for the gradient of the formants with respect to articulatory parameters[19] and gradients may be precomputed at each codevector in the case of the hypercube codebook of [19]. Genetic algorithms that do not use a codebook have also been used[24].

Vocal tract outlines estimated by inversion for static vowels and fricatives have been compared against X-ray microbeam measurements of gold pellets placed on the tongue[18,25], and vocal tract outlines estimated for static vowels compared against real vocal tract shapes from X-ray images[26]. Simultaneously recorded articulatory and acoustic data that are publicly available, include the X-Ray Microbeam (XRMB) Speech Production Database from the University of Wisconsin, Madison[27], and the Edinburgh Multi-CHannel Articulatory (MOCHA) database[28]. In the XRMB database, articulatory data are available in the form of X-ray microbeam measurements of gold pellets placed on the tongue, teeth/jaw, and lips, along with simultaneously recorded acoustic data, for several speakers uttering a series of tasks. In the MOCHA database, similar articulatory data are available from Electromagnetic Articulography (EMA). In both databases, no information is available in the pharyngeal region since all XRMB pellets or EMA coils were placed either in the oral cavity or on the face. However, except for the larynx, some information is available on the positions of all the other important articulators (jaw, tongue body and tip, lips). A reasonable geometric error measure for inversion can therefore be obtained by comparing estimated VT outlines against measured positions of tongue and lip XRMB pellets. The available geometric information may also give clues as to the weights or constraints that need to be placed on the displacements of different articulators in order to more accurately recover the VT shape for a particular speaker and speech sound.

In this paper, we perform a systematic study of acoustic-to-articulatory inversion for non-nasalized vowel sounds by analysis-by-synthesis using the Maeda articulatory model,

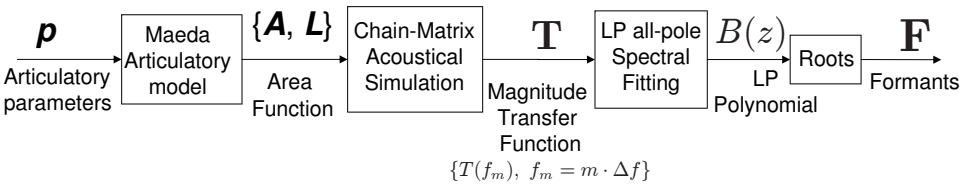


FIG. 2. Articulatory-to-Acoustic Mapping, Computation of Formants

and the XRMB database. We use the first three formants as acoustic features and develop efficient algorithms for codebook search and subsequent convex optimization. Calibration and adaptation of the Maeda model are discussed in detail for two speakers, one male and one female, from the XRMB database. Adaptation of the model includes scaling the overall VT and the pharyngeal region separately, modifying the model outer VT outline using measured palate and pharyngeal wall traces. XRMB dynamic articulatory data were also used to prune the codebook and improve inversion results. Inversion results are presented for the male speaker, and after quantifying both acoustic and geometric errors in inversion, error analysis is performed.

Sections II to V.B of this paper are organized around the block diagram of Figure 1, and the details of the different quantities and blocks in the figure are described. The articulatory-to-acoustic mapping, including the Maeda articulatory model, chain-matrix acoustic simulation, computation of cepstra and formants, and choice of acoustic features are described in Section II. The inversion cost function is given in Section III and its minimization using an articulatory codebook for initialization and subsequent convex optimization are described in Sections IV and V. Calibration and adaptation of the Maeda model are addressed in Section VI. Inversion results and error analysis are presented in Section VII and discussed in Section VIII.

II. THE ARTICULATORY-TO-ACOUSTIC MAPPING

Figure 2 shows the block diagram of the articulatory-to-acoustic mapping used in our work.

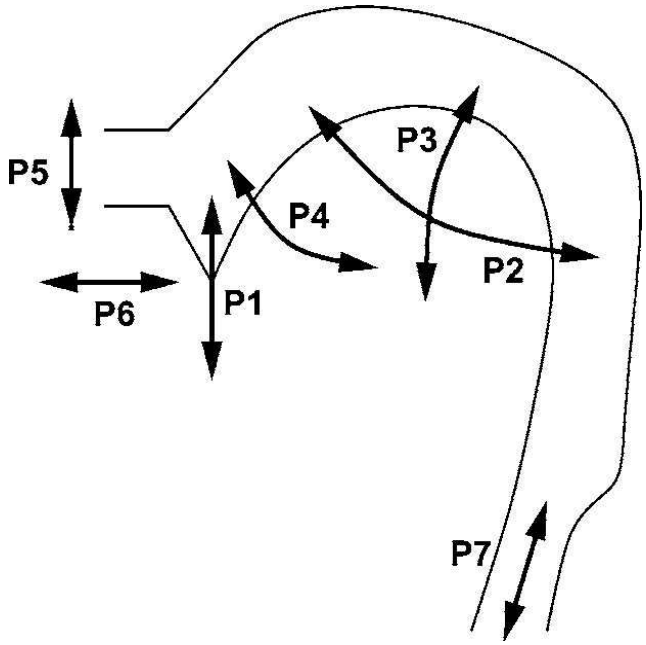


FIG. 3. Maeda articulatory model[23]: dependence of inner midsagittal VT outline on parameters (Reprinted with permission from [19], Copyright 2005, Acoustical Society of America). The parameters are: P1 - jaw (up/down), P2 - tongue body position (front/back), P3 - tongue body shape (arched/flat), P4 - tongue tip position (up/down), P5 - lip height (up/down), P6 - lip protrusion (front/back), and P7 - larynx height (up/down).

A. The Maeda Articulatory Model

In the Maeda articulatory model[23], the outer midsagittal VT outline consisting of the hard and soft palates (velum) and rear pharyngeal wall is fixed for a speaker (except for larynx height). The dependence of the inner midsagittal VT outline on parameters is shown in Figure 3. The inner VT outline is controlled by seven parameters: jaw position, tongue body position and shape, tongue tip position, lip height and protrusion, and larynx height. The VT outlines are described using a system of semi-polar grid lines, and the offsets, $\mathbf{v}$, of the inner VT outline along the grid lines are obtained as a linear combination of basis offset vectors:

$$\mathbf{v} = V\mathbf{p} + \mathbf{m}_v \tag{1}$$

where $\mathbf{p}$ is the articulatory parameter vector, V is the matrix containing the basis offset vectors, and $\mathbf{m}_v$ is the mean offset vector. V and $\mathbf{m}_v$ were obtained from a factor analysis of tongue shapes. The parameters $\mathbf{p}_i, 1 \leq i \leq 7$ are normalized by mean and standard deviation, and vary in the range [-3,3].

Midsagittal widths $d(x)$ along the length of the tract x, are converted to areas using the heuristic formula[23,29]:

$$A(x) = \alpha(x) d(x)^{\beta(x)} \tag{2}$$

where $\alpha(x)$ and $\beta(x)$ are *ad hoc* coefficients that vary along the tract. Using the semi-polar grid, the area function is obtained as a sequence of varying areas and lengths of 29 uniform tubes. The lengths of the tube sections in the area function are the distances between the midpoints of consecutive midsagittal grid line segments between the outer and inner VT outlines.

B. Chain Matrix Computation of VT Acoustic Response

The chain matrix method is one of the preferred approaches for computing the acoustic response of the vocal tract given its area function[13,30]. Here, the pressure, P, and volume velocity, U, at the input and output of an acoustic tube, for a linear wave, are related in the frequency domain by:

$$\begin{pmatrix} P_{out} \\ U_{out} \end{pmatrix} = \begin{pmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{pmatrix} \begin{pmatrix} P_{in} \\ U_{in} \end{pmatrix} \tag{3}$$

where the subscripts *in* and *out* denote the input and the output of the tube respectively. $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ are referred to as the chain parameters of the tube, and the matrix formed is called the chain matrix (CM).

If the vocal tract for a non-nasalized vowel sound is approximated as a series of N uniform tubes starting at the glottis and ending at the lips, the overall CM, K, is just the product of the individual CMs:

$$K = K_N \cdot K_{N-1} \cdot \cdots \cdot K_1 \tag{4}$$

where K_n is the CM of the nth tube. The transfer function of the vocal tract for a non-nasalized vowel sound may then be shown to be:

$$H = \frac{U_L}{U_G} = \frac{1}{(\mathcal{A} - \mathcal{C} Z_L)} \tag{5}$$

where U_G and U_L are the volume velocities at the glottis and lips, respectively, $\mathcal{A}$ and $\mathcal{C}$ are the elements of the CM of the overall vocal tract, and Z_L is the radiation impedance at the lips often approximated by that of a pulsating disk of air at the mouth opening[31]. The chain matrix method may also be extended to compute vocal tract transfer functions for other speech sounds such as nasals, nasalized vowels, fricatives, and laterals[13,30,32,33].

C. Chain Matrix for the Sondhi-Schroeter Model of the Vocal Tract

In our work, we follow the Sondhi-Schroeter model for wave propagation in a vocal tract used in [12, 30, 34] and [13], where frequency dependent losses due to air viscosity, heat conduction and yielding tract walls are taken into account. For this model the CM parameters of a uniform lossy cylindrical tube of area A (not to be confused with the CM parameter $\mathcal{A}$) and length L at angular frequency ω are given by ([30]):

$$\mathcal{A}_n = \cosh \frac{\sigma L_n}{c} \qquad \mathcal{B}_n = -\frac{\rho c \gamma}{A_n} \sinh \frac{\sigma L_n}{c} \tag{6}$$

$$\mathcal{C}_n = -\frac{A_n}{\rho c \gamma} \sinh \frac{\sigma L_n}{c} \qquad \mathcal{D}_n = \cosh \frac{\sigma L_n}{c} \tag{7}$$

where ρ is the density of air, and c is the speed of sound in air. Details on the values of the different parameters and the formulae for calculating γ and σ are given in [30]. The important thing to be noted is that γ and σ are only functions of frequency and do not depend on the area or the length of the tube.

The CM and the transfer function are typically computed for a set of equally spaced frequencies, and then used to compute quantities of interest like cepstra, the all-pole LPC spectral envelope and formant frequencies.

D. Computation of Formants

The steps involved in the computation of formants for given Maeda model parameters $\mathbf{p}$ are shown in Figure 2. First the VT area function is obtained as a series of N uniform tubes of varying areas and lengths: $\{\mathbf{A}, \mathbf{L}\}$, $\mathbf{A} = [A_1\ A_2\ \ldots\ A_N]$ $\mathbf{L} = [L_1\ L_2\ \ldots\ L_N]$. The chain matrix method is then used to compute the VT transfer function (H, Equation 5). The magnitude of the VT transfer function is:

$$T(f) = |H(f)| = \frac{1}{|\mathcal{A} - \mathcal{C}Z_L|} \tag{8}$$

where $\mathcal{A}$ and $\mathcal{C}$ are the elements of the overall CM of the vocal tract, and Z_L is the radiation impedance at the lips. $T(f)$ is computed at frequencies: $f_i = i \cdot \frac{F_{max}}{N_f},\ 0 \leq i \leq N_f$ where (N_f+1) is the number of frequency samples, and $F_{max} = f_s/2$ where f_s is the speech sampling frequency, for comparison with natural acoustic features.

The formant frequencies can then be computed from the roots of the denominator polynomial of an all-pole envelope fitted to $T(f_i),\ f_i = i \cdot \frac{F_{max}}{N_f},\ 0 \leq i \leq N_f$ using Spectral Linear Prediction[36].

The most computationally intensive step in Figure 2 is the calculation of the VT CM using Equations 4 to 7 since there may be up to $N = 30$ sections in the area function, and $T(f)$ may be desired at $N_f = 30$ or more frequency points depending on the sampling rate and frequency resolution.

E. Choice of Acoustic Features

The formants have a close relationship with the VT shape and the first three formants are therefore often used as acoustic features for inversion of vowels[18,19]. However, formant estimation can be difficult for high-pitched talkers, consonants, and semi-vowels.

As described in Section II.B, the acoustic quantity that is calculated first during articulatory synthesis is the VT transfer function. The calculation of formants involves finding the roots of an all-pole model fitted to samples of the transfer function at a set of uniformly

spaced frequencies. It would therefore be computationally simpler to match the computed VT transfer function with natural speech signal spectra, than matching computed and natural formants. Matching spectra would also effectively result in matching the formant spectral peaks, and explicit formant estimation would not be necessary.

However, it is difficult to directly compare computed spectral magnitude values with estimated natural values. The natural spectrum first needs to be smoothed, the voice source spectral tilt needs to be removed, and sensitivity to formant bandwidths needs to be decreased due to inaccuracies in the speech production model. The raised sine lifter introduced in [20] may be used to decrease the spectral tilt resulting from the voice source, and to emphasize the formant peaks. Mel frequency warping is also used to account for the fact that perturbations of the logarithm of the area function more linearly affect the logarithms of the formant frequencies (as a first order approximation)[15]. These operations are all performed more conveniently in the cepstral domain, and captured in a linear weighting matrix on cepstra[11,21,39].

However in this paper, we first explore formants as acoustic features for analysis-by-synthesis, and quantify and study the resulting inversion errors. A comparison of analysis-by-synthesis using cepstra and formants is a topic for future work.

III. THE OPTIMIZATION COST FUNCTION

As discussed in Section I, the objective function to be minimized (E) is the sum of acoustic (E_{acou}), regularization (E_{reg}) and 'geometric' continuity (E_{geo}) terms[12,18,19]:

$$E = E_{acou} + c_{reg}E_{reg} + c_{geo}E_{geo} \tag{9}$$

where c_{reg} and c_{geo} are weights, and

$$E_{reg} = \sum_{t=1}^{T} ||\mathbf{p}(t)||^2 \tag{10}$$

$$E_{geo} = \sum_{t=1}^{T-1} ||\mathbf{p}(t+1) - \mathbf{p}(t)||^2 \tag{11}$$

where $\{\mathbf{p}(t),\ 1 \leq t \leq T\}$ is the articulatory vector sequence, and the Euclidean norm is used. The acoustic term E_{acou} is discussed below. Note that the entire articulatory parameter sequence is simultaneously optimized. The way in which the weights c_{reg} and c_{geo} are chosen to achieve the occasionally competing goals of acoustic match, realistic VT shapes and smooth articulatory trajectories is discussed in Section VII.

The cost function of Equation 9 is computed in the "convex optimization" block of Figure 1.

A. Acoustic Cost with Formants

We first explored formants as acoustic features for analysis-by-synthesis, with the acoustic term in the cost function being:

$$E_{acou} = \sum_{t=1}^{T} \sum_{n=1}^{3} \left(\log F_n(t) - \log \bar{F}_n(t)\right)^2 \tag{12}$$

where $F_n(t)$ and $\bar{F}_n(t)$ are respectively the computed and natural n^{th} formants for the frame at time t. It is well known that:

$$|\log F - \log \bar{F}| \approx \frac{|F - \bar{F}|}{\bar{F}} \tag{13}$$

Therefore, Equation 12 approximately measures the sum of the squares of the relative errors in the formants, with the approximation becoming increasing accurate as the relative errors decrease. The error in the LHS is less than 2.5% when the RHS is 0.05, and less than 0.5% when the RHS is 0.01.

The limitation in the number of formants used is mainly due to the loss in accuracy of the speech production acoustic model at higher frequencies. At frequencies above around 3kHz (e.g. see [37]), the assumption of plane-wave propagation starts to break down, and the effect of transverse modes in the vocal tract becomes more significant. Other possible sources of error in computed acoustics include inaccurate modeling of losses, zeros due to side branches such as piriform sinuses, etc. With a more accurate acoustic model, a larger spectral frequency range including the fourth and higher formants could help reduce inversion

ambiguity, and needs to be investigated. But it is still of interest to see how successful inversion can be with limited acoustic information using a commonly used acoustic model.

IV. CONSTRUCTION AND EFFICIENT SEARCH OF THE ARTICULATORY CODEBOOK

As discussed in Section I, a codebook is needed to initialize the analysis-by-synthesis because of the computationally intensive forward mapping, local optima, and the non-uniqueness of the inverse mapping.

A. Codebook Construction and Pruning using XRMB Data

We followed the method of codebook construction using log formant bins described in [21]. Formant vectors are conveniently lower dimensional and also important for characterizing VT acoustics.

We first obtained 2×10^6 random articulatory configurations with a minimum area along the VT greater than 0.05 cm^2, and total VT length between 14 cm and 19 cm (the VT length for the nominal configuration of the Maeda model is around 16.3 cm). These ranges of area and length are wide for vowels, which usually have minimum areas greater than 0.15 cm^2, and areas smaller than 0.1 cm^2 typically result in frication[14]. Corresponding acoustic vectors were calculated for the random samples. With a log formant bin width corresponding to 15% relative error and an ∞-norm of 0.8 in articulatory space, the codebook size was around 230000 vectors. Constraining the minimum area to be greater than 0.15 cm^2, we obtained a pruned codebook of around 180000 vectors.

This large codebook still contains many unrealistic articulatory configurations, which may hinder the retrieval of realistic articulatory trajectories for an input acoustic vector sequence. While the Maeda model imposes some realistic constraints on VT shapes, combinations of extreme values of Maeda parameters often result in unrealistic configurations. Some of these could be eliminated with more information about VT geometry during speech.

We developed a novel method to further prune the codebook using the tongue and lip pellet positions measured in the XRMB database. First the Maeda model VT outlines were shifted (and scaled if necessary for a given speaker) so that the model and measured palate positions behind the teeth are aligned (as in Figures 11 to 16 in Section VII). The speech utterances of the speaker in the XRMB database were segmented using a simple energy-based endpoint-detector. Each XRMB measurement frame includes the positions of four pellets on the tongue and two lip pellets (except for errors such as pellet detachment). Lip pellets were shifted vertically by the approximate height between them during a token of /m/ for the speaker, for comparison with the lip height from the model. Cubic spline interpolation was used to obtain a partial tongue outline from the tongue pellets. From the intersections of the partial tongue outline with the grid lines used in the Maeda model, the offsets along some (typically around 12) of the grid lines may be obtained.

For one out of every 5 XRMB frames (i.e. approximately every 34.5 ms), 'measured' lip height and partial tongue grid offsets were determined, and the distances, d from corresponding tongue grid offsets and lip heights for all codebook vectors were determined. By eliminating all codevectors sufficiently distant (i.e., with d greater than a threshold) from any of the measured configurations, the codebook size was greatly reduced. Taking d to be the maximum magnitude difference, for a threshold of 0.15 cm, the codebook size was around 43000.

B. Codebook Search

The bin structure of the codebook in the formant domain can also be exploited for efficient search. First the bin containing an input formant vector is identified, and the search at time t then continues only in it and neighboring bins.

For dynamic speech segments, since the cost function includes the geometric distance, the search for the optimal codevector sequence involves dynamic programming (DP)[12]. For the DP search, we used two kinds of pruning. At each time t, from the identified formant

bins, only the best n_1 codevectors according to $E_{acou} + E_{reg}$ were considered for the DP iteration, and after the iteration, only n_2 codevectors were retained for the next iteration. Good search results were obtained even with $n_1 = 200$ and $n_2 = 20$, for a fraction of the original search time. The DP search may be further improved by using distance beams to prune paths instead of n_1-best and n_2-best sorting.

V. CONVEX OPTIMIZATION OF THE COST FUNCTION

A. BFGS Quasi-Newton Method and Derivatives of the Cost Function

Further optimization is needed after codebook initialization to obtain both a better acoustic match with the input speech, and smoother articulatory trajectories.

We developed an efficient way of calculating the derivative of the CM of the VT with respect to the area function, since the computation of the VT CM is the most expensive step in synthesis as noted at the end of Section II. This was then used in the Broyden-Fletcher-Goldfarb-Shanno (BFGS)[38] quasi-Newton method to optimize the cost function of Equation 9. The BFGS method has better (superlinear) asymptotic convergence than some other methods used in the past for optimization of area functions. The direct search methods of [12] and [18] and the iteration in the variational approach of [19] which appears to be a type of fixed point method, have linear convergence.

The BFGS method requires $\dfrac{\partial E}{\partial \mathbf{p}}$, the gradient of the cost function with respect to articulatory parameters. Although the articulatory parameter trajectory is simultaneously optimized, here we ignore time dependence for the sake of clarity. $\dfrac{\partial E_{reg}}{\partial \mathbf{p}}$ and $\dfrac{\partial E_{geo}}{\partial \mathbf{p}}$ can easily be calculated from Equations 10 and 11. Details may be found in [39]. The functional dependencies in computing E_{acou} are (see Figure 2):

$$\mathbf{p} \rightarrow \{\mathbf{A}, \mathbf{L}\} \rightarrow \mathbf{T} \rightarrow B(z) \rightarrow \mathbf{z} \rightarrow \mathbf{F} \rightarrow E_{acou} \tag{14}$$

where $\mathbf{z}$ are roots of $B(z)$ and $\mathbf{F}$ are formants. $\dfrac{\partial E_{acou}}{\partial \mathbf{p}}$ can be computed by applying the chain rule for derivatives. $\dfrac{\partial E_{acou}}{\partial \mathbf{F}}$ is relatively straightforward to calculate from Equations

12, as is $\frac{\partial \mathbf{F}}{\partial \mathbf{z}}$ from $z_i = e^{j2\pi F_i/f_s}$ (where the notation $\frac{\partial \mathbf{x}}{\partial \mathbf{y}}$ is used to denote the matrix of partial derivatives $\left[\frac{\partial \mathbf{x}(i)}{\partial \mathbf{y}(j)}\right]$ when $\mathbf{x}$ and $\mathbf{y}$ are both vectors). Calculation of $\frac{\partial \mathbf{z}}{\partial \mathbf{T}}$ involves calculating derivatives of the roots of a polynomial with respect to its coefficients, and derivatives of the auto-correlation sequence calculated from $\mathbf{T}$ with respect to $\mathbf{T}$ (involved in the LP Spectral Fit)[36]. $\frac{\partial \mathbf{A}}{\partial \mathbf{p}}$ and $\frac{\partial \mathbf{L}}{\partial \mathbf{p}}$ can be calculated from the equations of the Maeda articulatory model, which were discussed in Section II.A.

We focus on the step $\{\mathbf{A}, \mathbf{L}\} \rightarrow \mathbf{T}$, i.e., the chain matrix calculation of the VT transfer function, which is the most computationally intensive step.

B. Chain Matrix Derivatives with respect to the Area Function

By Equation 8, $\mathbf{T}$ depends on the CM parameters $\mathcal{A}$ and $\mathcal{C}$ of the VT and the radiation impedance Z_L. Therefore, to compute $\frac{\partial \mathbf{T}}{\partial \mathbf{A}}$ and $\frac{\partial \mathbf{T}}{\partial \mathbf{L}}$, we need to compute the derivatives of $\mathcal{A}$ and $\mathcal{C}$, which are given by Equations 4 to 7, with respect to $\{\mathbf{A}, \mathbf{L}\}$. Note that $\mathcal{A}$ and $\mathcal{C}$ are elements of the matrix K in Equation 4. The details of the calculation of $\frac{\partial \mathbf{T}}{\partial \mathbf{A}}$ and $\frac{\partial \mathbf{T}}{\partial \mathbf{L}}$ from $\frac{\partial K}{\partial \mathbf{A}}$ and $\frac{\partial K}{\partial \mathbf{L}}$ may be found in [39].

We first calculate $\frac{\partial K}{\partial A_n}$. Observe from Equations 6 and 7, that the CM of each section depends only on its own area and length, and not on those of other sections. This simplifies the derivative calculation from Equation 4:

$$\frac{\partial K}{\partial A_n} = [K_N \cdots K_{n+1}] \cdot \frac{\partial K_n}{\partial A_n} \cdot [K_{n-1} \cdots K_1] \tag{15}$$

If we define:

$$P_n = K_{n-1} K_{n-2} \cdots K_1, \quad 2 \leq n \leq N \tag{16}$$

$$Q_n = K_N K_{N-1} \cdots K_{n+1}, \quad 1 \leq n \leq N-1 \tag{17}$$

and let:

$$P_1 = Q_N = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{18}$$

then:

$$\frac{\partial K(\mathbf{A},\mathbf{L})}{\partial A_n} = Q_n \cdot \frac{\partial K_n}{\partial A_n} \cdot P_n, \quad 1 \leq n \leq N \tag{19}$$

From Equations 6 to 7, we can show:

$$\frac{\partial \mathcal{A}_n}{\partial A_n} = 0 \qquad \frac{\partial \mathcal{B}_n}{\partial A_n} = -\frac{1}{A_n} \cdot \mathcal{B}_n \tag{20}$$

$$\frac{\partial \mathcal{C}_n}{\partial A_n} = \frac{1}{A_n} \cdot \mathcal{C}_n \qquad \frac{\partial \mathcal{D}_n}{\partial A_n} = 0 \tag{21}$$

Therefore, $\frac{\partial K_n}{\partial A_n}$ is very easily obtained from A_n and the elements of K_n.

The partial derivatives with respect to the lengths of the area function can also similarly be calculated in an efficient way without much extra calculation.

VI. CALIBRATION OF THE MAEDA MODEL TO A SPEAKER

A. Calibration Cost Function and Method

For analysis-by-synthesis to be able to recover accurate vocal tract shapes for a given speaker, the Maeda articulatory model first needs to be calibrated to the speaker. That is, we need to verify that acoustic features (e.g. formants) computed from measured VT shapes for the speaker match simultaneously measured natural acoustic features. Measured XRMB pellet positions need therefore to be fitted by VT outlines, from which acoustic features can be computed. The task of calibration is made more difficult by the fact that the XRMB pellets do not give any information on the tongue shape in the pharyngeal region. Fitting VT outlines to XRMB pellets therefore involves both interpolation of the tongue shape between pellets in the oral region, and extrapolation of the tongue shape into the pharyngeal region of the vocal tract. While only four model grid offsets along the tongue are needed to uniquely recover the jaw and three tongue parameters (see Section II.A), this could result in parameters outside the nominal range, discontinuity across frames, and errors at other grid points especially in the pharyngeal region due to mismatch of the model with the speaker.

Toutios et al.[40] used constrained quadratic programming with variational regularization to obtain continuous Maeda tongue parameter trajectories within the nominal range of [-3,3], fitting four measured EMA sensors on the tongue. Cubic spline interpolation was used to obtain tongue offsets between sensor positions. After fitting VT shapes to measured sensor positions, they verified that measured natural formants lay within the range of variation of computed formants, when the larynx height parameter was varied within [-3,3].

For calibration of the Maeda model to a speaker, it is necessary to verify that there exist articulatory parameters within the nominal range of [-3,3], such that geometric perpendicular distances between VT shapes and measured pellet positions and acoustic distances between computed and measured acoustic features are both simultaneously small for a set of calibration frames. A general calibration method may therefore be developed by including an acoustic distance term in the cost function used to fit VT shapes to measured pellet positions in [40]. This is equivalent to adding an extra term to the cost function for inversion (Equation 9) that measures the distance between VT shapes and known pellet positions.

The cost function of interest for calibration is:

$$E_{cal}(\mathbf{p}, \Theta) = E_{acou} + c_{fit}E_{fit} + c_{reg}E_{reg} + c_{geo}E_{geo} \tag{22}$$

where E_{acou} is as in Equation 12, c_{reg}, E_{reg}, c_{geo}, E_{geo} are as in Equations 9, 10 and 11, c_{fit} is a weight, and E_{fit} measures the error in the fit between the tongue pellets and the VT outline:

$$E_{fit} = \sum_{t=1}^{T} ||V\mathbf{p}(t) + \mathbf{m}_v - \mathbf{v}(t)||^2 \tag{23}$$

where $\mathbf{v}(t)$ are the interpolated tongue offsets along model grid lines in the oral region, V is the matrix of corresponding basis offset vectors, and $\mathbf{m}_v$ are the corresponding mean offset values (see Equation 1).

Θ in Equation 22 consists of different various parameters and "constants" of the Maeda model that could vary with speaker, including[41]:

1. Overall geometric (length) scaling factor, or separate scaling factors for the oral and pharyngeal portions of the vocal tract

2. The outer VT outline

3. V and $\mathbf{m}_v$ (Equations 1 and 23)

4. The coefficients used to convert midsagittal widths to cross-sectional areas ($\alpha(x)$ and $\beta(x)$ in Equation 2)

For fixed Θ, codebook search and BFGS optimization can be used to optimize $E_{cal}(\mathbf{p}, \Theta)$ only as a function of $\mathbf{p}$ as in Sections IV.B and V for inversion. After optimization, low values for E_{fit} and E_{acou} would indicate that the model is calibrated for the chosen vowel sounds for the speaker. If it is not possible to make E_{fit} and E_{acou} simultaneously small, then the Maeda model (i.e., Θ) would have to be adapted to better fit the speaker.

The optimization approach we have developed in this paper has the advantage that it can be modified or extended without much difficulty to adapt all these different parameters in Θ. For example, when only a partial outer VT outline is available, as in the XRMB database, we can fix all other parameters in Θ, fix the Maeda parameters $\mathbf{p}$ to fit measured pellets for a set of calibration vowel frames (i.e., obtain a low value of E_{fit}), and then optimize $E_{cal}(\mathbf{p}, \Theta)$ only as a function of the outer VT outline to improve the acoustic match and the calibration.

We used the first three formants of the cardinal vowels /a/, /i/ and /u/ to perform calibration. Since these three cardinal vowels capture to some extent the range of variation of vocal tract shapes and formants for a speaker, a match for these would be a minimum requirement from a calibration method. In total, six frames from Task 14 were used, two for each cardinal vowel. The formant based cost function of Equation 12 was used for E_{acou}. Since we use static vowel frames, $c_{geo} = 0$ in Equation 22. We also take $c_{reg} = 0$ unless resulting parameters lie outside the nominal range.

The following steps are used for calibration:

1. We first obtained tongue shapes in the oral region by an average of cubic spline (used in [40]) and Hermite cubic polynomial interpolation between XRMB pellets. Cubic spline

interpolation sometimes results in overshoot of the interpolated tongue shape over the palate in some cases when pellets were very close to the palate as in /i/. Hermite polynomial interpolation maintains monotonicity of the interpolated shape between samples, and averaging the two polynomials gave a trade-off between smoothness and monotonicity. To obtain tongue grid offsets, Maeda model VT outlines were shifted so that the model and measured palate positions behind the teeth are aligned. Lip pellets were shifted vertically by the approximate height between them during a token of /m/ for the speaker, and horizontally averaged and shifted by an ad hoc speaker-specific distance.

2. Then, the Maeda model outline was scaled so that the rear pharyngeal wall outline of the model is approximately aligned with that of the speaker, by a visual match. This gives a VT length scaling factor. In our work, the overall VT scaling factor was applied with respect to the reference female speaker of the Maeda model.

3. The outer outline of the model is modified using the partial palate and pharyngeal traces available for the speaker from the XRMB database. There is an important point to be noted while modifying the pharyngeal portions of the model's outer VT outline to match the measured pharyngeal trace for the speaker. The pharyngeal trace provided in the XRMB database extends, for some speakers, from a point in the laryngo-pharynx or oro-pharynx to a point in the naso-pharynx as can be seen from the example Figure 5.9 in the XRMB database manual[27]. Since the naso-pharynx is not used for vowel production, the upper point of the pharyngeal trace provided for the speaker in the XRMB database cannot be used to adapt the Maeda model pharyngeal outline for these speakers. Only the lower portion of the provided pharyngeal trace may be reliable for the purpose of model adaptation. Also, as noted in [27], the XRMB pharyngeal traces are only coarse approximations derived from vocal tract images that are not very sharp.

   The unknown portions of the outer VT outline are initialized to the corresponding

portions of the nominal model outline.

4. A large number (e.g. 2×10^5) of random parameter combinations are obtained, uniformly distributed in the nominal range.

5. The random set of parameters above is pruned to eliminate those articulatory vectors with outlines that extend beyond the partial palate and pharyngeal traces available for the speaker from the XRMB database. For the pruned random codebook, acoustic vectors (first three formants) are computed.

6. For each calibration analysis frame, the adaptation codebook is searched using $E_{acou} + c_{fit}E_{fit}$, with the value of c_{fit} chosen large enough (we used 0.01) to emphasize the fit of the VT outline to the measured pellets more than the acoustic match. The acoustics computed from the parameters may not be very accurate due to model mismatch, sparseness of codebook sampling, or due to the unknown portions of the speaker's actual outer VT outline being very different from the model's nominal outer VT outline. However, the inclusion of acoustic distance in the codebook search serves to regularize the geometric fit, and the parameters obtained by codebook search will be in the nominal range, approximately fit the measured tongue pellets and shifted lip pellets, and also be such that the computed acoustic features match measured acoustic features approximately.

7. The unknown portion of the outer VT outline, and the larynx height parameter (P7 in Fig 3) for each frame are simultaneously adjusted using all calibration analysis frames to minimize E_{acou}. Since the outer VT outline is fixed in the Maeda model, adjusting it will affect the acoustics of all sounds. Therefore the outer VT outline needs to be adapted using all calibration frames together. We also combined the adaptation of the outer VT outline with the optimization of P7 for each frame as P7 is left free by the tongue pellet data and needs to be determined using the acoustics. A continuity/smoothness cost on the optimized outer VT outline is also included. The

parameters P1 to P6 obtained via the codebook search in the previous step are kept fixed, as they determine the fit of the inner VT outline to the measured tongue and shifted lip pellets.

8. One parameter to adapt at this point is the pharyngeal scaling factor. An indication that this needs to be adapted, will be given by out-of-range values of P7 in the optimization above, or by errors in the computed third formant of /u/ and the second formant of /i/ which depend upon the pharyngeal cavity. The range of pharyngeal sections to scale would be decided by the lowest point of the pharyngeal outline trace provided in the XRMB database. An optimal pharyngeal scaling may be chosen from a discrete set of values (for example from 0.7 to 1.3 with spacing of 0.01) so that the average error in the third formant of /u/ and the second formant of /i/ are minimized. The pharyngeal scaling factor was applied over the overall VT scaling factor chosen in Step 2 above.

9. Finally, the calibration cost function $E_{acou} + c_{fit}E_{fit}$ is optimized with respect to the articulatory parameters, keeping the outer VT outline fixed. If the resultant parameters lie outside the nominal range the regularization cost E_{reg} is also included. The weight c_{reg} can be varied to satisfy the parameter limits while reducing E_{acou} and E_{fit} as much as possible.

After this, both E_{acou} and E_{fit} should be sufficiently small, for example with less than 3% error in the first three formants, and less than 0.1 cm average error in offsets, and the optimized parameters should be within the nominal range, to indicate that calibration is verified.

The steps involving the codebook formation and search are there in order to take the acoustic cost into account. These can be skipped to simplify the process, if parameters obtained by optimizing just E_{fit} (if necessary with parameter range constraints) already satisfy the calibration requirements of small E_{acou} and E_{fit}. A simple way to force parameters to lie in the nominal range is using the regularization cost E_{reg} in addition to E_{fit}, with the

resulting optimization just a regularized least squares problem.

In the calibration steps above, changing either the articulatory parameters or the model speaker-dependent parameters such as the other VT outline and the pharyngeal scaling factor will affect the acoustic match and/or the fit of the model outlines to the measured pellets. Some iteration of steps 5 to 9 may therefore be needed to achieve calibration by this method.

If the calibration requirements of small E_{acou} and E_{fit} are not satisfied, then the model is not satisfactorily calibrated. The above calibration steps considered adaptation of the Maeda model using length scaling factors for oral and pharyngeal regions, and modification of the outer VT outline, optimizing its unknown portions. Failure to calibrate the model by adapting these parameters indicates that some other basic aspect of the model needs to be adapted, such as the range of allowable parameters, or the basis offset vectors, or the coefficients $\alpha(x)$ and $\beta(x)$ used to compute area functions from midsagittal widths. Rotation of the model to better fit the speaker's articulatory data should also be considered.

We evaluated the inversion method on two XRMB speakers, one female ("JW46") and one male ("JW11"). The two speakers were selected for this initial study based on the fact that their measured palatal outlines are similar to that of the Maeda nominal palatal outline (after the palate positions behind the teeth are aligned). However, the calibration steps would be the same for other speakers since the model's outer VT would be adapted if necessary. We next discuss the calibration of the Maeda model for the two speakers using the above calibration procedure.

B. Calibration for Speaker JW46

For JW46, the best scaling factor for the model shapes to fit the palate and pharyngeal traces is found to be approximately 0.94, by a visual match.

Following all the calibration steps listed above, including the construction and search of the calibration codebook, calibration was successful for /i/ and /u/, with parameters within

nominal ranges, less than 3% error in the first three formants, and less than 0.1 cm average distance between tongue pellets and model outline. A separate pharyngeal scaling factor was not found to be necessary or useful for Speaker JW46.

However, it was difficult to calibrate /a/ for speaker JW46, as can be seen from Figure 4. While either the acoustic or the geometric fit can be improved, it was not possible to simultaneously obtain small values for both E_{acou} and E_{fit}. From the failure to calibrate /a/ for JW46, it is not clear that the optimized outer VT outline is even necessary or appropriate for the speaker.

One aim of calibration is essentially to verify that the extrapolation of known tongue shapes in the oral region into the pharyngeal region performed by the model is appropriate for the speaker, taking into account the measured acoustic features for different vowels. For /a/, the acoustic match is typically obtained with pharyngeal areas being smaller with respect to oral areas. It seems clear that the model extrapolation into the pharyngeal region is not satisfactory, when the fit of the outline to tongue pellets is good. This could be due to the slight raising of the tongue tip for this speaker's /a/. A known issue with Maeda model is that changing the tongue tip parameter also changes areas in the laryngeal region[23].

The Maeda model basis vectors could be modified, for example by scaling the pharyngeal portions of the basis vectors, without modifying them in the oral region. Also, the dependence of laryngeal areas on the tongue tip parameter could be removed. This could provide good fit to tongue pellets as with the unadapted model, and also provide satisfactory extrapolation in the pharyngeal region to match measured acoustics.

A detailed systematic study of this is beyond the scope of this paper.

Since the model could not be satisfactorily calibrated for /a/ of speaker JW46, no inversion results are presented for this speaker. In earlier work using a speaker independent codebook, we obtained generally realistic estimated VT shapes that approximately fit measured pellet positions for some diphthongs and vowel sequences[39].

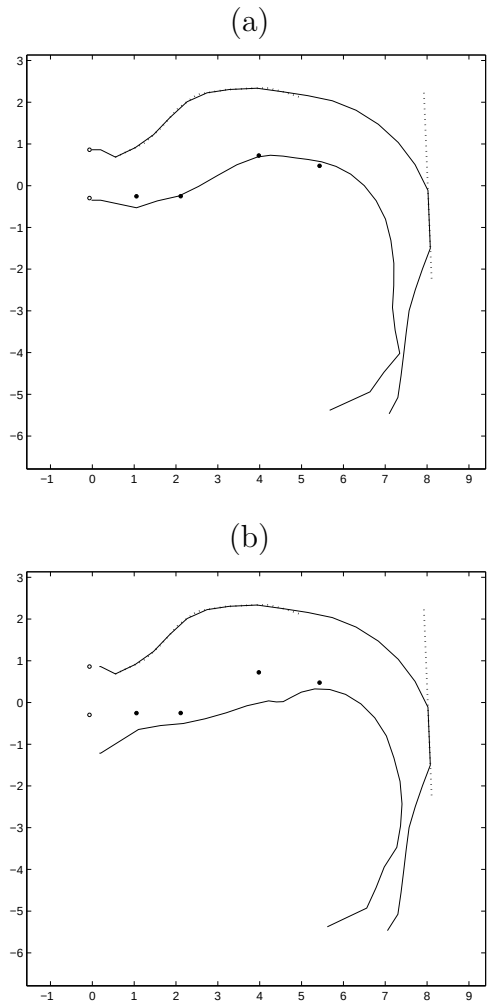


FIG. 4. Model VT shapes for /a/ of JW46, after optimization of the outer VT outline and optimization of Equation 22 with respect to articulatory parameters. (a) Average pellet to VT outline distance is less than 0.1 cm, but the average formant error is 14.7% (b) Average formant error is 5.7% , but average tongue pellet to VT outline distance is close to 0.3 cm.

C. Calibration for Speaker JW11

For speaker JW11, with an overall scaling factor of 1.19, a pharyngeal scaling factor of 0.83, and a modified and optimized outer VT outline, calibration for the three cardinal vowels was successful with parameters within nominal ranges, less than 3% error in the first three formants, and around 0.1 cm average distances between tongue pellets and model outlines. As mentioned earlier, the overall VT scaling factor given above is with respect

to the reference female speaker in the Maeda model and the pharyngeal scaling factor is applied over the overall VT scaling.

For /i/ tongue outlines are slightly further away from the palate than measured pellets. We suspect that these errors may be reduced by adapting the coefficients ($\alpha(x)$ and $\beta(x)$ in Equation 2) used to convert midsagittal widths to cross-sectional areas in this region of the palate.

For /u/, it was observed that for model tongue outlines to fit the measured pellets, the tongue body shape parameter (P3 in Fig. 3) had to be slightly outside the nominal range. There was still some acoustic mismatch when the model tongue outlines did fit the measured pellets, which could again perhaps be reduced by adapting $\alpha(x)$ and $\beta(x)$.

Investigation of these issues will be the focus of future work.

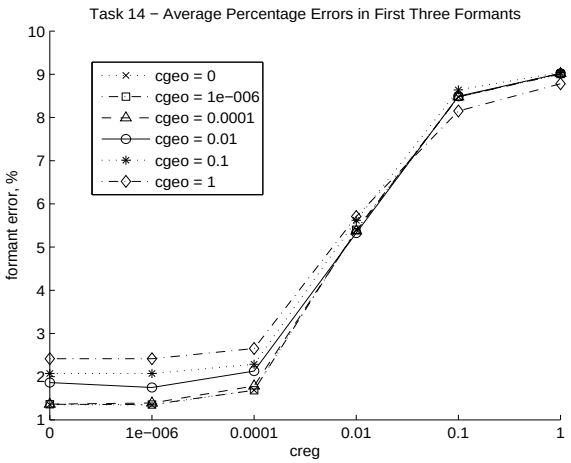
A speaker-specific codebook was constructed for JW11.

VII. RESULTS OF INVERSION EXPERIMENTS

The inversion method was evaluated for speaker JW11 on vowels, diphthongs and vowel sequences from utterance tasks 13, 14 and 15 of the XRMB database[27]. From Task 13, which consists of words of the form /sVd/ where V is a vowel/diphthong, we use the diphthongs /aɪ/, /ɔɪ/, /aʊ/ and /eɪ/ from the words/nonwords "side, soid, sowd and sayed" respectively. From Task 14 we use separately articulated vowels, a list of which may be found in Table I. From Task 15, we use the vowel sequences /iu/, /iɑ/, /uɑ/, /ɑu/, /ɑi/ and /ui/.

We downsampled speech signals to 8kHz, and manually extracted formants from the LPC analysis of the speech signals for Tasks 13, 14 and 15. Frames were centered around times at which XRMB pellet positions were measured, with a frame rate of around 146Hz. A lower frame rate would probably suffice and will be explored in the future.

(a)



(b)

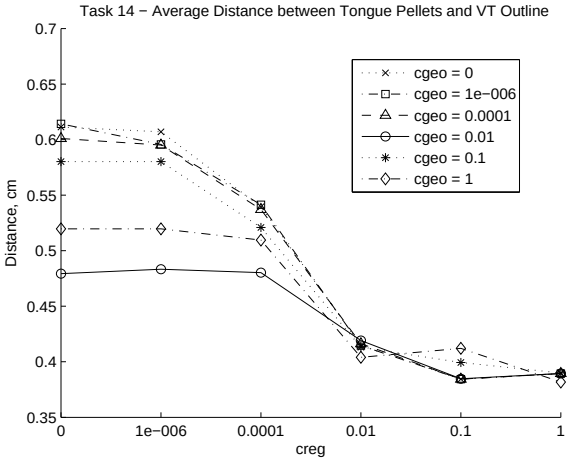
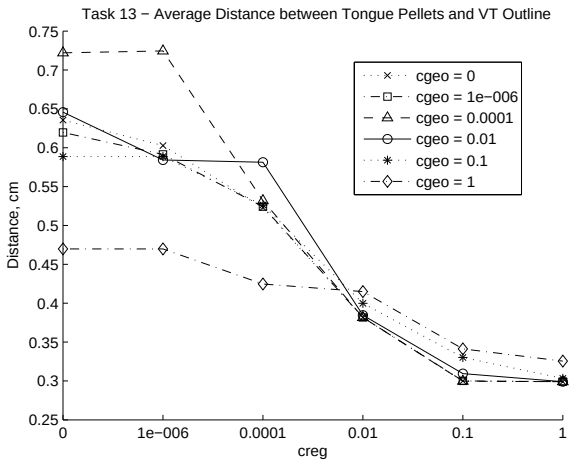


FIG. 5. Task 14, Codebook search results using unpruned codebook with 184819 vectors, varying c_{reg} and c_{geo}. (a) Average errors in first three formants. (b) Average distance between tongue pellets and estimated VT outlines.

A. Codebook Search Results

The goals of inversion using analysis-by-synthesis are to obtain a good match between input and synthetic acoustic features (i.e., low E_{acou}), realistic estimated VT shape sequences (related to E_{reg}) and smooth articulatory trajectories (low E_{geo}). The values of c_{reg} and c_{geo} in the cost function may need to be carefully chosen, as discussed in Section III, to achieve a balance between these three simultaneous goals. The acoustic and geometric error

(a)



(b)

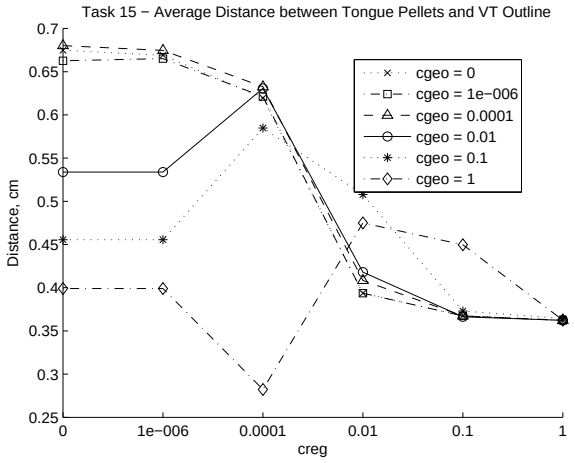
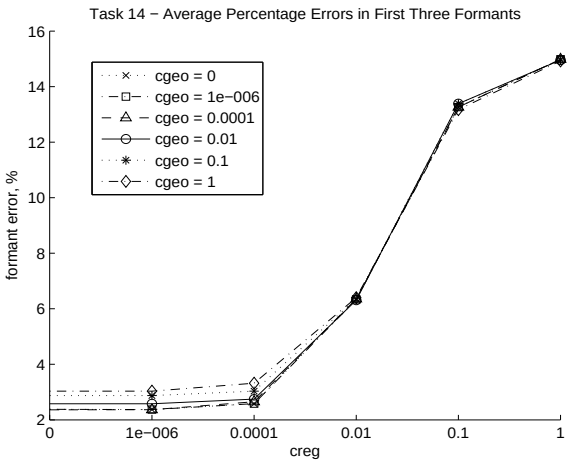


FIG. 6. Codebook search results using unpruned codebook with 184819 vectors, varying c_{reg} and c_{geo}. Average distance between tongue pellets and estimated VT outlines for (a) Task 13 and (b) Task 15.

measures used to evaluate inversion results are, respectively, the average percentage error in the first three formants, and the average perpendicular distances from measured tongue pellet positions to the estimated VT outlines (i.e., the corresponding nearest line segments of the VT outline). Since the lip pellets need to be translated by ad hoc distances before comparison with the model lip outline, only a visual match is used here.

For the vowels, diphthongs and vowel sequences from Tasks 13, 14 and 15, we investigated whether it was possible to get low acoustic and geometric errors for any combination of c_{reg}

(a)



(b)

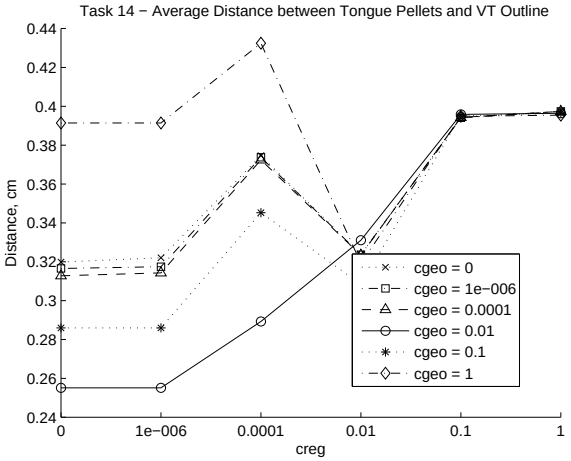


FIG. 7. Codebook search results for Task 14 using XRMB data-pruned codebook with 43086 vectors, varying c_{reg} and c_{geo}. (a) Average errors in first three formants. (b) Average distance between tongue pellets and estimated VT outlines.

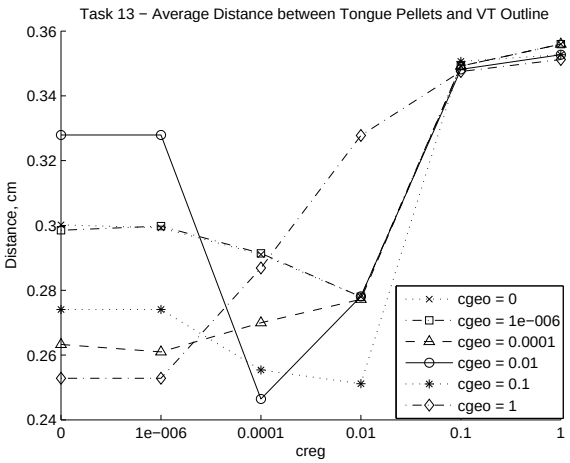
.

and c_{geo}, for both the unpruned codebook with 184819 vectors, and the XRMB data-pruned codebook with 43806 vectors, that were discussed in Section IV.A.

The results of formant-based codebook search with varying c_{reg} and c_{geo} are shown in Figures 5 and 6 for the large unpruned codebook.

It is seen from Figure 5(a) that, as expected, the acoustic error (average percentage error in the first three formants) generally increases as c_{reg} and c_{geo} are increased. The acoustic

(a)



(b)

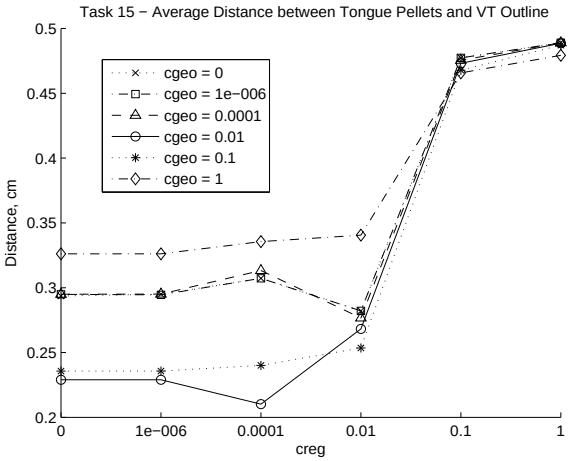


FIG. 8. Codebook search results using XRMB data-pruned codebook with 43086 vectors, varying c_{reg} and c_{geo}. Average distance between tongue pellets and estimated VT outlines for (a) Task 13 and (b) Task 15.

.

error variation for Tasks 13 and 15 is similar to that in Figure 5(a) for Task 14. From Figure 5(b), it is also seen that for Task 14, the geometric errors (average distances between tongue pellet and estimated VT outlines) decrease from higher values to around 0.4 cm and formant errors increase to about 9% as c_{reg} is increased. The geometric errors from codebook search results for Tasks 13 and 15 are shown in Figure 6 (a) and (b) respectively.

In inversion by analysis-by-synthesis, the aim is to improve the geometric fit by reducing

the acoustic error, starting from an initial sequence of articulatory parameters for which the acoustic and geometric errors are both relatively small. If c_{reg} is large, it would not be possible to reduce the acoustic error much further with convex optimization. With a well calibrated model, this implies that the geometric error would also not decrease much. It appears that with the unpruned codebook, for the selected representative vowels, it is not possible to obtain values for c_{reg} and c_{geo} that would give good initial sequences of vocal tract shapes with both acoustic and geometric errors relatively low, for the three tasks.

The results of formant-based codebook search for varying c_{reg} and c_{geo}, for the XRMB-pruned codebook of 43086 vectors, are shown in Figures 7 and 8.

For the pruned codebook, for Task 14, while the acoustic error increases to around 15% as c_{reg} is increased over the same range, the geometric error varies over a smaller range of between 0.25-0.44 cm for the three tasks, for the range of variation of c_{reg} and c_{geo}. Also, for $c_{reg} = 0.0001$ and $c_{geo} = 0.01$, the average geometric error is around 0.20-0.30 cm and the average formant error is around 3%. Also note that the average geometric error is the lowest for both Tasks 13 and 15 for this combination of parameters for the values considered. This implies that either of these tasks could have been used to estimate the values of c_{reg} and c_{geo}. The set of discrete values we have considered here for c_{reg} and c_{geo} is very sparse, and their values could possibly be more finely tuned.

B. Results of Convex Optimization

We fixed $c_{reg} = 0.0001$ and $c_{geo} = 0.01$, and performed convex optimization of the formant-based cost function after codebook search. The inversion acoustic and geometric errors after codebook search and convex optimization are given in Tables I, II and III for Tasks 14, 13 and 15 respectively.

Figure 9 shows an example of articulatory parameters before (dotted lines) and after (solid lines) optimization for the vowel sequence /ɑi/ from Task 15 of JW11. It can be seen that the parameters vary more smoothly after optimization. Figure 10 shows computed

formants after codebook search and convex optimization compared with natural formants for the same test case.

It is observed from Tables I, II and III, that the formant errors (related to the acoustic term in the cost function) always decrease after convex optimization, usually by a significant amount. Also, articulatory trajectories become smoother after optimization (related to the continuity term in the cost function). However, the geometric error between measured XRMB pellets and estimated VT outlines does not always decrease after optimization, and in fact the average geometric error over phonemes increases for both Tasks 14 and 13, as seen from Tables I and III respectively. We discuss this further in Section VIII below.

Measured XRMB gold pellet positions are plotted against the estimated VT outlines and shown for four evenly spaced frames each from /aɪ/, /ɔɪ/ and /aʊ/ in Figure 11, and for eight evenly spaced frames from /eɪ/ in Figure 12, all from Task 13. For the second half of /aʊ/, the mouth rounding is not recovered and the estimated VT shapes are unrealistic, with a wide mouth opening. This is not reflected in the geometric error which does not include the error in the estimated positions of measured lip pellets. Perhaps tighter pruning of the codebook with XRMB data would eliminate these unrealistic shapes for /aʊ/. For /aɪ/, while the estimated VT shapes are realistic and acoustic error is low, the error between VT shapes and pellets is close to 0.5 cm. The inversion results for /aʊ/ and /aɪ/, together with the low acoustic errors for both, indicates non-uniqueness in the acoustic-to-articulatory mapping for these cases. Since the inversion method does not currently handle /s/ and /d/, it does not capture the context of the diphthongs of Task 14 which were taken from words/nonwords of the form /sVd/. Perhaps the results could be improved with dynamic information if the inversion method were extended to fricatives and stops. We discuss this further below in Section VIII. For /eɪ/, while the error in recovered tongue shape is small (0.12 cm average tongue pellet-VT outline distance), there is some error in the recovered lip pellets. It must be recalled that the lip pellets are shifted by ad hoc distances and plotted, which inherently has some error.

For /oʊ/ of Task 14, the average distance between tongue pellets and estimated VT

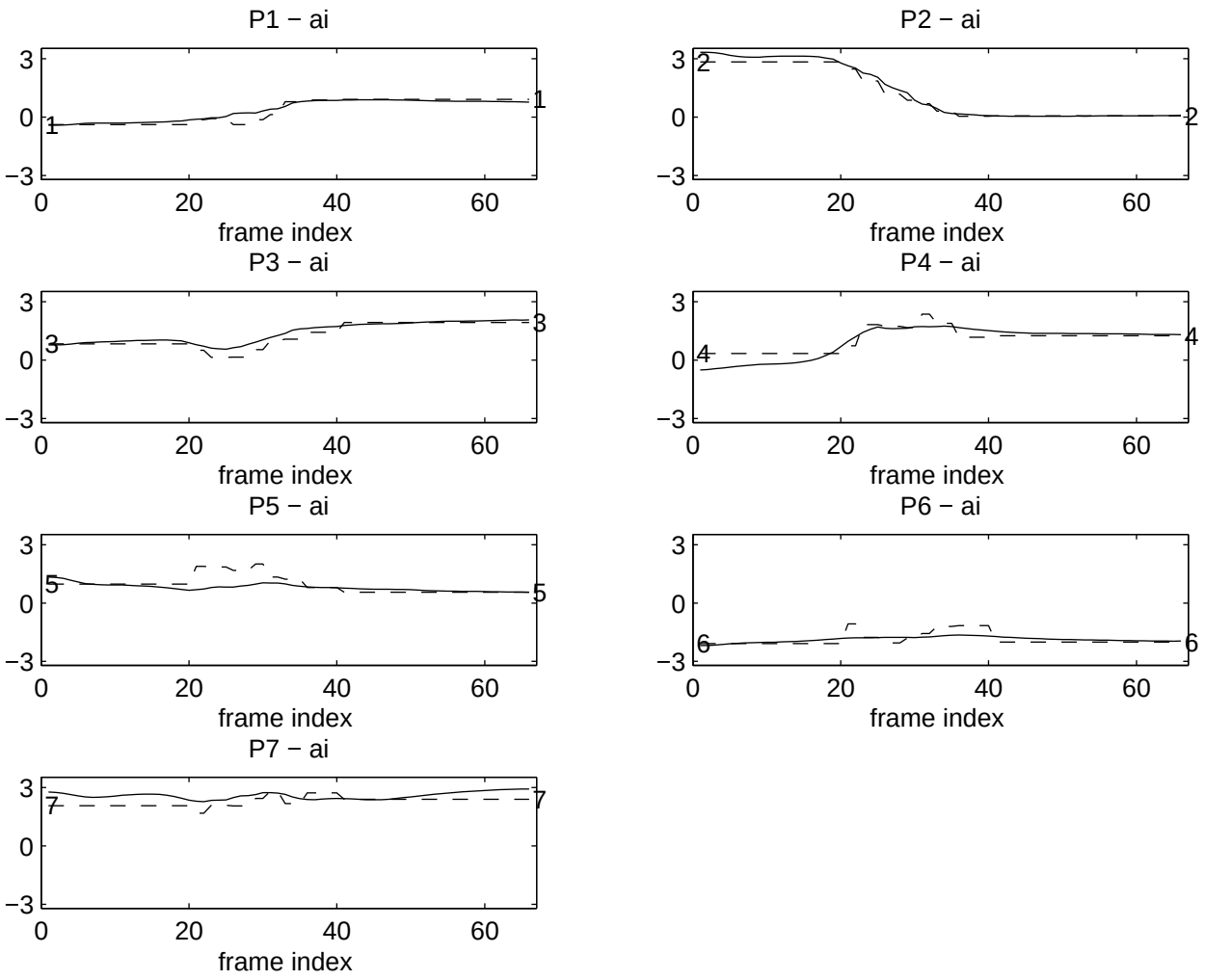


FIG. 9. Example of articulatory parameters before (dashed lines) and after (solid lines) optimization. /ɑi/ from Task 15 of Speaker JW11 (see corresponding formants in Figure 10 and VT shapes in Figure 14). In each subfigure the value of the corresponding articulatory parameter is plotted along the y-axis which is limited approximately to the range [-3,3], the nominal range of the Maeda model parameters.

outline is 0.12 cm, with the average formant error being 2.12%. Since the constriction for /oʊ/ is in the soft palate region where the outer VT outline is not really available, there is some acoustic mismatch after inversion. Inclusion of data from /oʊ/ in calibration might improve the acoustic error after inversion for /oʊ/.

The low inversion errors for /eɪ/ and /oʊ/ suggest that the inversion method is capable

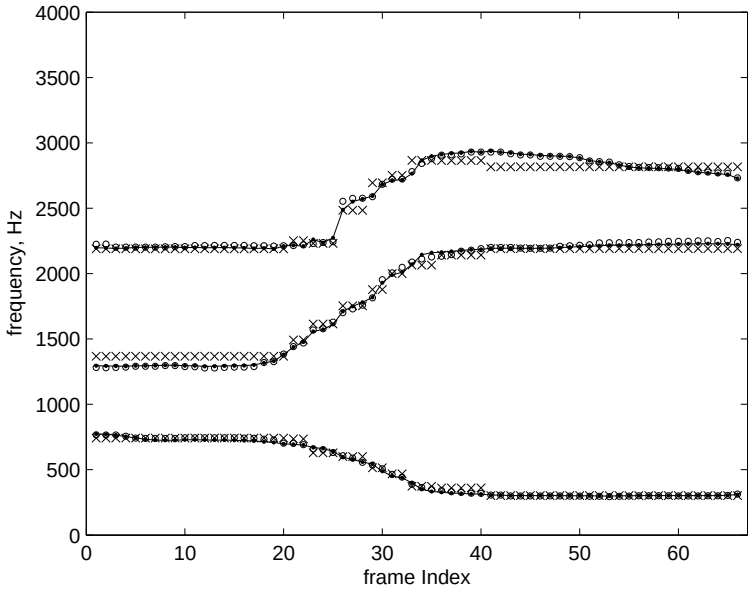


FIG. 10. Natural (circles), Codebook (crosses) and Optimized (lines) formants for /ɑi/ from Task 15 of Speaker JW11 (see corresponding parameters in Figure 9 and VT shapes in Figure 14).

of recovering finer articulatory contrasts in some cases.

Estimated VT shapes and measured pellets are plotted for one frame each from nine relatively static vowels from Task 14 in Figure 13. Figures 14, 15 and 16 show the results of inversion of vowel sequences /ɑi/, /ɑu/ and /ui/ of speaker JW11. These are in increasing order of geometric errors (see Table III). While the estimated VT shapes for vowel sequences /ɑi/ and /iɑ/ in Task 15 had low geometric error of around 0.10 cm (see Figure 14 and Table III), the estimated VT shapes for vowels /ɑ/ and /i/ in static context in Task 14 have larger errors of around 0.30 cm (see Table I and Figure 13). The acoustic errors were low in both the static and dynamic cases. This seems to imply that trajectory information is useful for recovering the VT shape also for cardinal vowels such as /ɑ/ and /i/. Inversion results are very poor for /ə/ and /ɔ/ in Task 14, again probably due to non-uniqueness in VT shapes for their formant values, which is discussed in the following section.

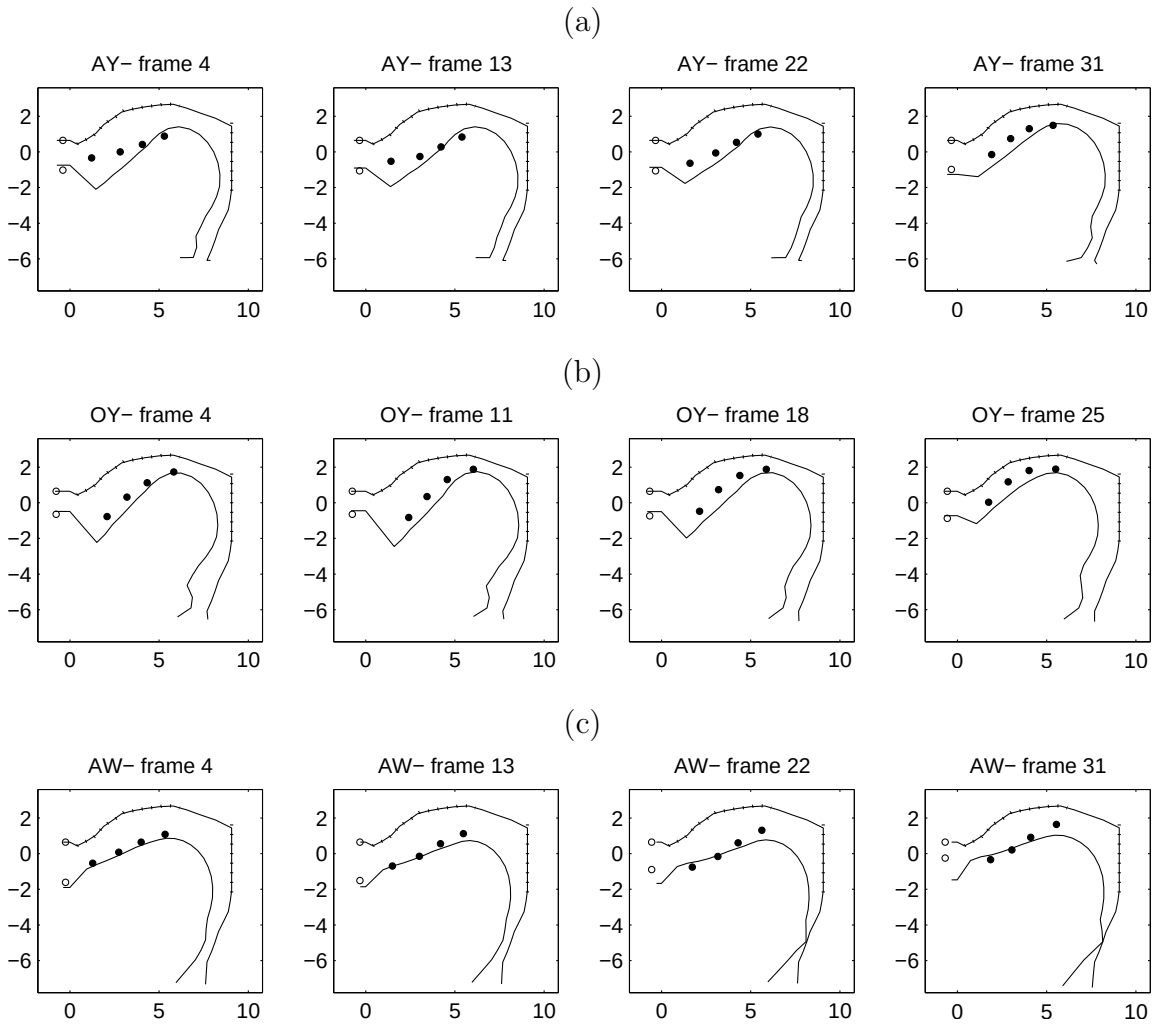


FIG. 11. Speaker JW11, Task 13 (a) /aɪ/ from 'side' (b) /ɔɪ/ from 'soid' (c) /aʊ/ from 'sowd' - Measured XRMB tongue (solid circles) and shifted lip (empty circles) pellet positions plotted against estimated VT outlines (solid lines). Vowel labels above figures are given in DARPA format. For the three diphthongs, average formant errors are 1.21%, 0.42% and 0.37% respectively, and average distances between tongue pellets and estimated VT outlines are 0.49 cm, 0.33 cm and 0.26 cm respectively.

VIII. DISCUSSION

As noted above, it is seen from Tables I to III that the acoustic error always decreases after the convex optimization. This is expected since the acoustic error is the main component of the optimization cost function in analysis-by-synthesis. The hope in performing

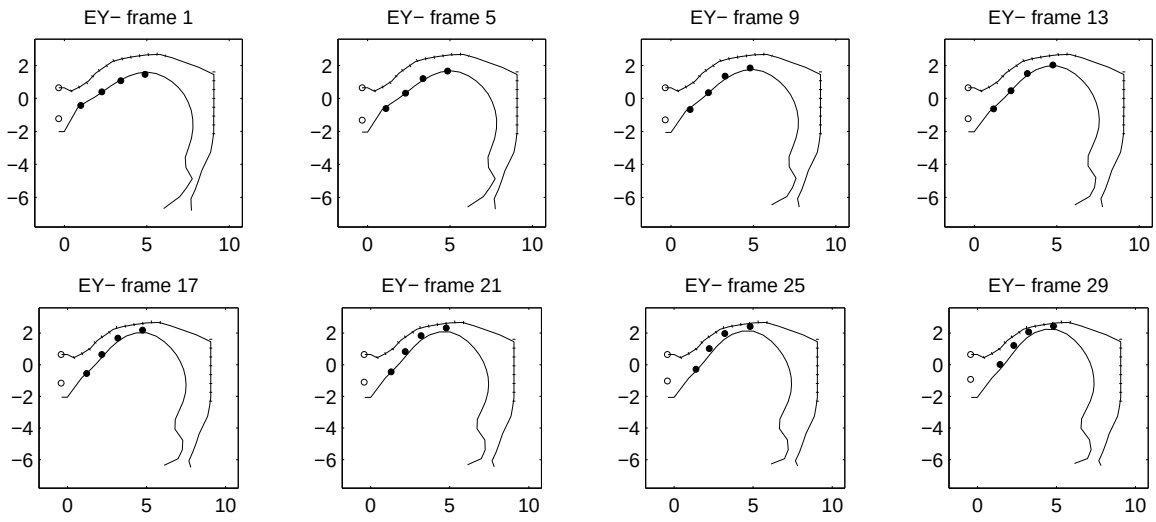


FIG. 12. Speaker JW11, Task 13, /eɪ/ - Measured XRMB tongue (solid circles) and shifted lip (empty circles) pellet positions plotted against estimated VT outlines (solid lines). Vowel labels above figures are given in DARPA format. The average formant error is 0.34% and the average distance between tongue pellets and estimated VT outline is 0.12 cm.

analysis-by-synthesis is that the geometric error would also decrease as the acoustic error decreases. However, it is seen that the geometric inversion error does not always decrease after convex optimization, and in fact increases for many phonemes. Also, the geometric error is high for many phonemes.

By optimizing the calibration cost function (Equation 22) using codebook search and convex optimization, we verified that the model was well calibrated for all the test phonemes in Tables I to III . That is, for each test phoneme, it was verified that there exist articulatory parameter sequences with low acoustic error between calculated and measured formants ($< 3\%$) and low geometric errors between calculated VT outlines and measured formant XRMB pellet positions (< 0.10 cm). Therefore poor calibration was not to blame for high geometric inversion errors.

As discussed in Section I, it is well known that for many phonemes, due to the non-uniqueness of the acoustic-to-articulatory inverse mapping, the analysis-by-synthesis cost function has multiple local optima. If the initial codebook sequence of VT shapes is not

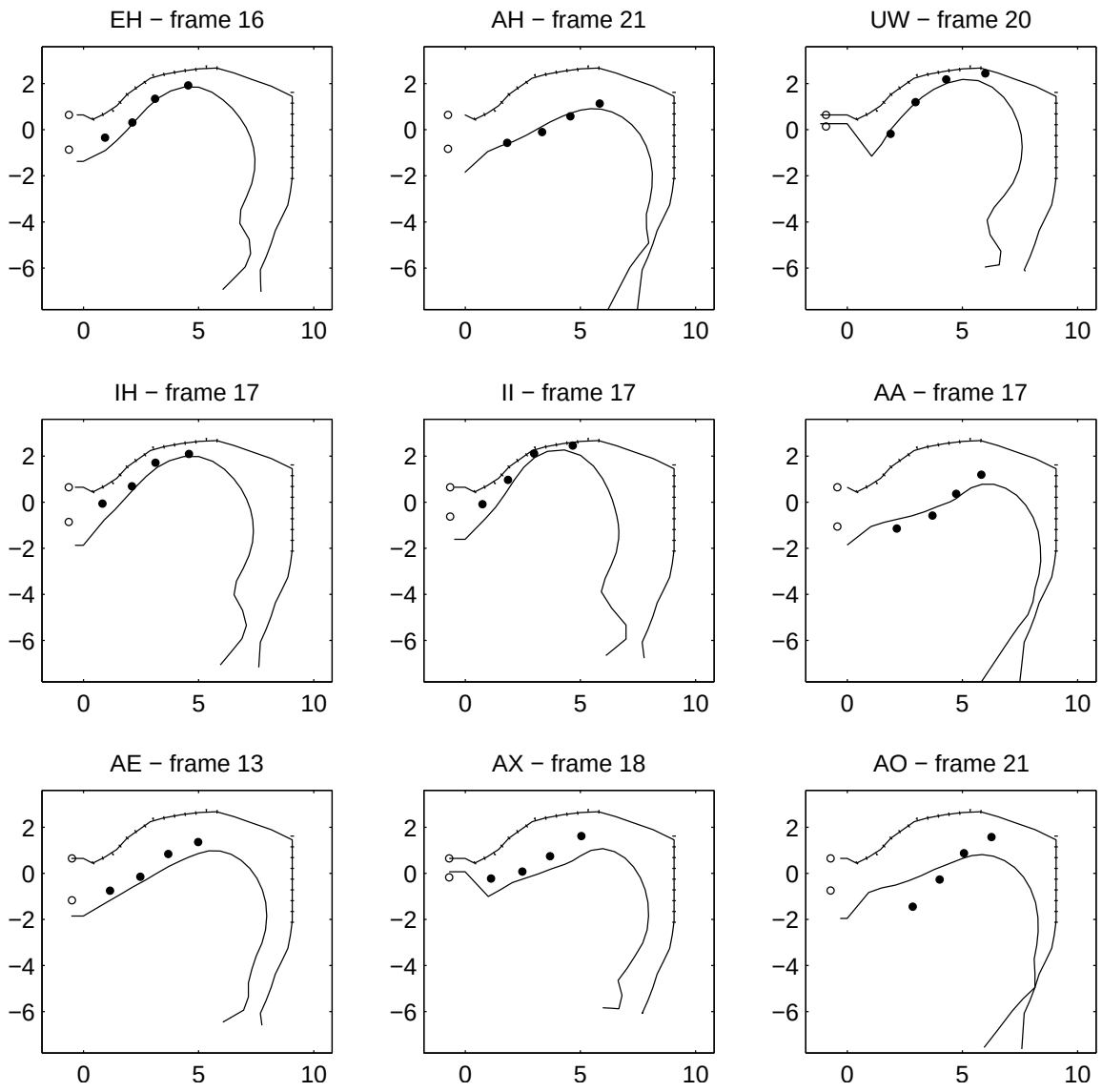


FIG. 13. Speaker JW11, Task 14, Representative frames from relatively static vowels - measured XRMB tongue (solid circles) and shifted lip (empty circles) pellet positions plotted against estimated VT outlines (solid lines). Vowel labels above figures are given in DARPA format. See Table I for the equivalent IPA labels, average formant errors and the average distance between tongue pellets and estimated VT outlines.

near the actual sequence of VT shapes but rather near one of the other non-unique inverse solutions, then optimizing the cost function will converge to the corresponding local optimum, with improved acoustic fit but possibly larger pellet to VT outline distances. This is

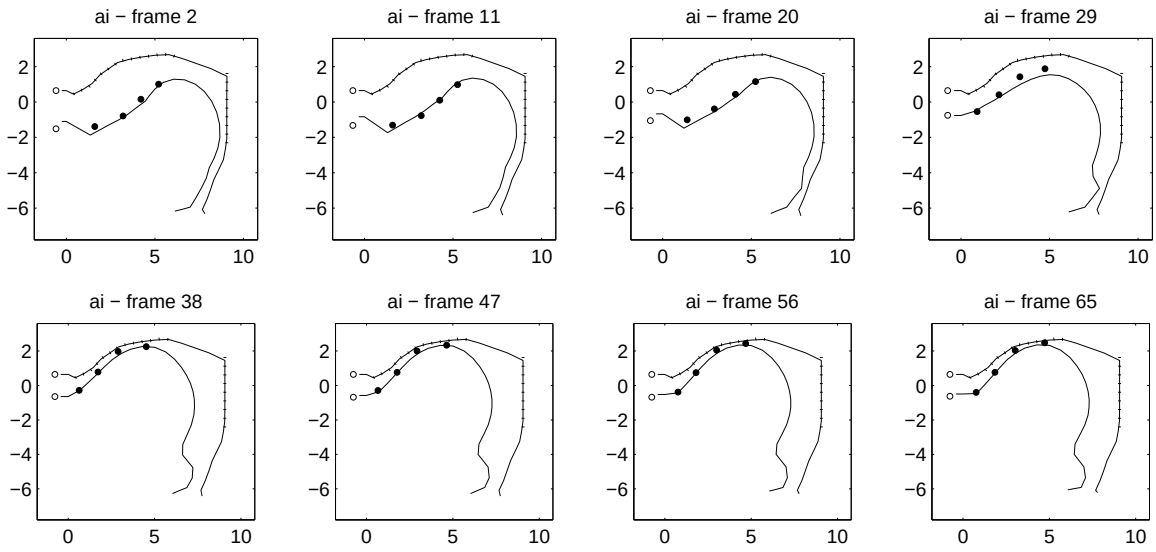


FIG. 14. Speaker JW11, /ɑi/ - Measured XRMB tongue (solid circles) and shifted lip (empty circles) pellet positions plotted against estimated VT outlines (solid lines). The average formant error is 0.33% and the average distance between tongue pellets and estimated VT outline is 0.10 cm.

particularly observed for phonemes /ɪ/ through /ɔ/in Table I, and phonemes /aɪ/ through /aʊ/ in Table II.

The effect of the non-uniqueness of the acoustic-to-articulatory mapping may also be observed in the results with the unpruned codebook (Figure 5) compared to that with the pruned codebook (Figure 7). For $c_{reg} = c_{geo} = 0$, while the acoustic error is lower (around 1.5% average formant error) for the unpruned codebook than for the pruned codebook (around 2.5%) the geometric error is much higher (0.6 cm compared to 0.32 cm). This is to be expected since the criterion optimized is the acoustic distance which is indeed lower with the unpruned codebook at the price of using unrealistic articulatory shapes.

From the above discussion, it is clear that the initial articulatory sequence obtained from codebook search is crucial for the success of inversion by analysis-by-synthesis. This was one of the reasons why the codebook was pruned using XRMB data for the speaker, to eliminate unrealistic VT shapes that result from a naive sampling of the articulatory space within the nominal range. Even with the pruned codebook, the non-uniqueness is a serious problem

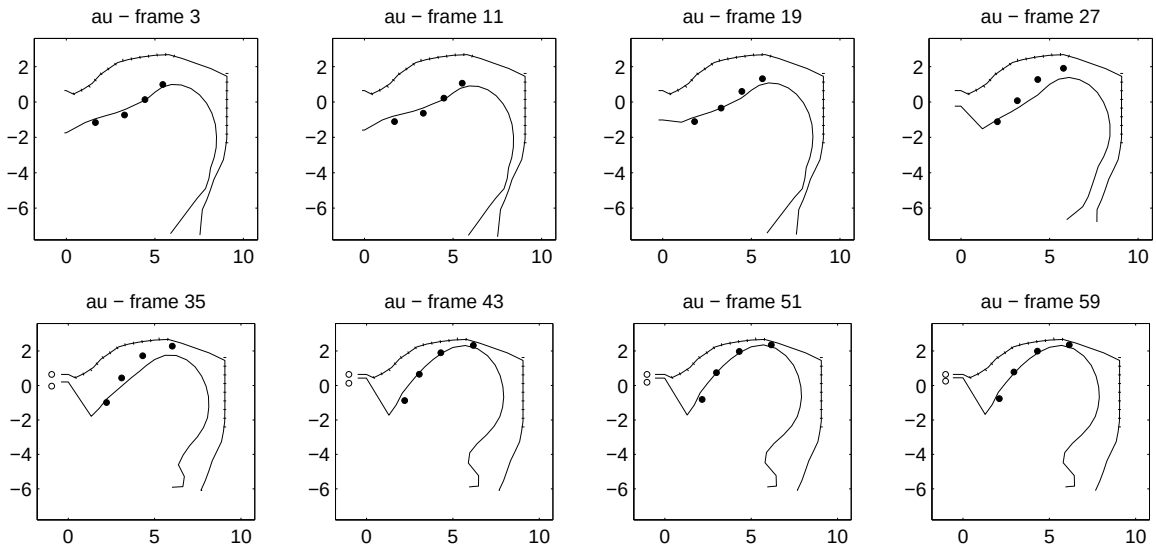


FIG. 15. Speaker JW11, /ɑu/ - Measured XRMB tongue (solid circles) and shifted lip (empty circles) pellet positions plotted against estimated VT outlines (solid lines). The average formant error is 2.43% and the average distance between tongue pellets and estimated VT outline is 0.20 cm.

for several phonemes such as /æ/, /ə/ and /ɔ/ of Task 14 and /aɪ/ of Task 13, where the geometric inversion errors are high. The current codebook search and inversion cost function do not always yield a good initial parameter sequence for optimization. Alternate cost functions and codebook search strategies may perhaps work better and need to be investigated.

It is generally thought that the dynamic information in speech helps to reduce the effect of the non-uniqueness problem in inversion. Since the regularization and continuity terms in the optimization cost function resolve the non-uniqueness in analysis-by-synthesis, it is plausible that the non-uniqueness for a given phoneme or part of a dynamic phoneme could be correctly resolved by estimation of correct VT shapes for the left and right phonetic contexts.

Comparing the results of inversion for Tasks 14 and 13 in Tables I and II respectively, we see that several static vowels produced in isolated contexts in Task 14 have lower geometric errors than the diphthongs in Task 13, which seem to have more dynamic information. Also,

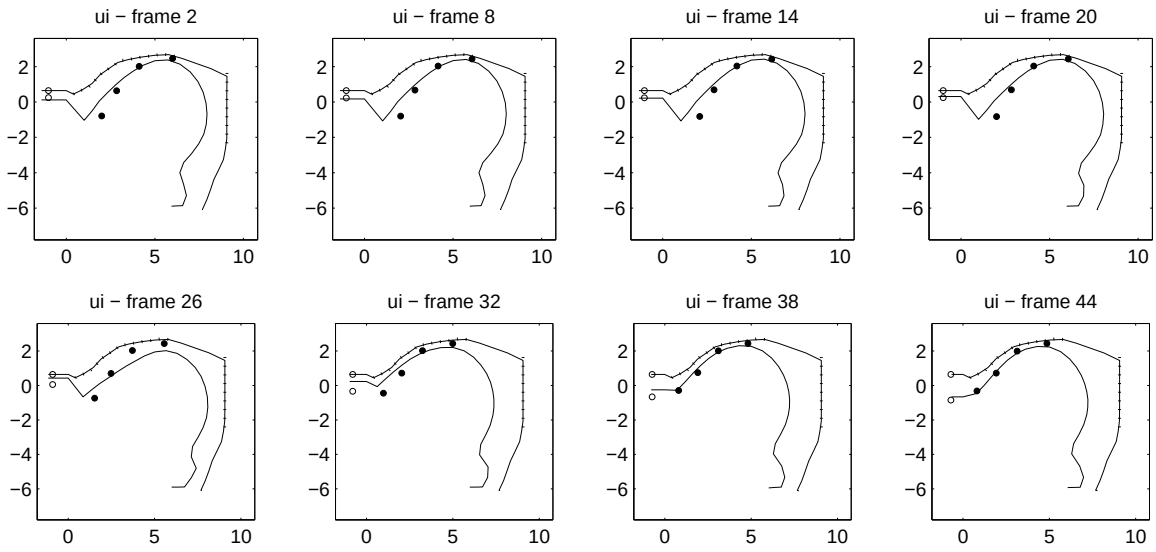


FIG. 16. Speaker JW11, /ui/ - Measured XRMB tongue (solid circles) and shifted lip (empty circles) pellet positions plotted against estimated VT outlines (solid lines). The average formant error is 2.20% and the average distance between tongue pellets and estimated VT outline is 0.26 cm.

a greater proportion (5 out of 11) of the static vowels studied seem to show improvement in geometric error after optimization compared to the diphthongs (1 out of 4).

However, it should be noted that the diphthongs of Task 13 were from contexts of the form /sVd/ where V is the diphthong. Since the inversion method does not currently handle fricatives and stops like /s/ and /d/, the left and right contexts of the diphthongs were not taken into account in the inversion, and there is actually missing dynamic information for the diphthongs compared to the static vowels of Task 14. Combined with the non-uniqueness of the inverse mapping (since calibration was verified for all test phonemes), the larger geometric errors for the diphthongs studied are therefore better explained and support the hypothesis that dynamic/contextual information are needed for accurate recovery of vocal tract shapes. It seems likely that the results for the diphthongs in Task 13 could be improved if the inversion method could handle /s/ and /d/, and inversion were performed on the entire utterance with all phonemes simultaneously.

The methods of analysis-by-synthesis that we have used are mostly "standard"; use of

a codebook, the general form of the cost function, use of optimization algorithms, are all generally found in the literature. One of the contributions of our paper is a better evaluation of the standard techniques of analysis-by-synthesis for inversion, by comparing estimated VT shapes against measured XRMB pellets. Our paper is, as far as we know, essentially the first one to do so for dynamic vowel sounds; [18, 25] and [26] only studied static vowels. We have also quantified the geometric error using the average perpendicular distance from pellets to estimated VT outline. Previous papers in the past have generally used acoustic criteria alone to judge results (the acoustic match is expected to be good in analysis-by-synthesis), or have used phonetic/linguistic human judgement to evaluate how realistic the results are[19]. We have also outlined a systematic procedure for adaptation/calibration of the Maeda model to a new speaker, which makes inversion by analysis-by-synthesis possible.

The inversion method used in this paper is, however, highly speaker dependent: the Maeda model is adapted to the speaker using a measured (partial) outer VT outline for the speaker and some calibration data; a speaker-specific codebook needs to be constructed, and some dynamic articulatory data from the speaker are needed to prune the codebook and improve inversion results. While some data are also needed to estimate appropriate values for coefficients c_{reg} and c_{geo} in the cost function, these estimates are more likely to work well for different speakers, if the codebooks are well-pruned. Our error analysis of results for Tasks 13, 14 and 15 with the XRMB data-pruned codebook showed that either Tasks 13 or 15 (containing dynamic diphthongs and vowel sequences) could be used to estimate reasonable values of c_{reg} and c_{geo}. Note that the "test set" utterances of Tasks 13, 14 and 15 were not used in the codebook pruning. However, due to paucity of appropriate analyzed data, we currently use six analysis frames (two each of /a/, /i/ and /u/) from the test utterances to calibrate the Maeda model to the speaker. Frames from other utterances should serve equally well for this purpose.

The need for extensive articulatory data to perform the codebook pruning for each speaker is a drawback. There are many articulatory configurations, with articulatory parameters inside the nominal range of $[-3, 3]$, that are probably never assumed for any speaker

and any sound in a given language (e.g. high jaw and tongue, but wide open lips), which are however present in a codebook naively constructed by sampling the articulatory space. Pruning using XRMB data utilizes parameter correlations to eliminate these unlikely configurations from the codebook, resulting in decreased inversion errors and also reducing the size of the codebook and improving the efficiency of codebook search. It remains to be determined whether articulatory data from a set of training speakers can be used to prune a codebook for a new test speaker.

Also, while some articulatory data from a given speaker appear to be unavoidably necessary for *accurate* recovery of VT shapes from speech sounds, generally reasonable shapes could still possibly be recovered using a nominal/standard articulatory model. Useful information that could be inferred may include constriction location along the tongue, and simultaneous articulatory gestures needed to produce certain sounds. Earlier results with a speaker independent codebook for speaker JW46 indicate that this is indeed possible, at least for some vowels and diphthongs[39]. However, we need to first study whether accurate recovery is possible at all with current methods, even with sufficient articulatory data available. This is where our paper's contributions lie.

For applications like language learning, one could also use articulatory data from non-native speakers to improve the codebook. Patterns of speech production in speakers can be studied off-line and fine phonetic contrasts that may be common could be learned and used to help speakers of one language learn another.

In Section I, we listed the main challenges faced in inversion: (1) complexity of speech production models, (2) inherent non-uniqueness of the inverse mapping, and local optima of the cost function, (3) incomplete knowledge about the shape and dynamics of the vocal tract for a given speaker, and (4) insufficient data to learn from or to evaluate the inversion results.

It is clear that all four factors remain big challenges in inversion. We developed efficient codebook search and optimization techniques to deal with the complexity of the articulatory-to-acoustic mapping. As explained earlier the primary reason for the poor re-

sults for some vowels or diphthongs is the non-uniqueness of the acoustic-to-articulatory mapping. For these sounds, there exist several competing VT shape sequences in the codebook that produce the same formants, even after codebook pruning using XRMB data. The current codebook search using the inversion cost function does not always pick a good initial parameter sequence from the candidates. Alternate codebook search strategies and cost functions need to be investigated. For dynamic information to be more useful in resolving the non-uniqueness, the inversion method should be extended to other speech sounds such as fricatives and stops.

Other factors that could contribute to the the inversion error are the limitation of the articulatory model and data to the midsagittal plane, the possible speaker-dependence of the coefficients $\alpha(x)$ and $\beta(x)$ used to convert midsagittal widths to cross-sectional areas (Equation 2) and variation of $\alpha(x)$ and $\beta(x)$ with the midsagittal width itself.

Restricting the articulatory model (and the articulatory data) to the midsagittal plane has both advantages and limitations. One advantage that the number of parameters that control the vocal tract shape and thereby the area function are reduced. An articulatory space of smaller dimension would likely give fewer possibilities to achieve a given set of acoustic features and possibly reduce the non-uniqueness problem of the acoustic-to-articulatory mapping. However, since the tongue can in reality move in three dimensions, many different 3-D VT shapes could map to the same 2-D midsagittal outline as the VT moves between different phonemes for dynamic speech sounds. The error in the mapping from the midsagittal outline to the area function could be large in such cases. Also, the same midsagittal VT shape could also map to different acoustic features. These factors would cause inversion errors.

The purely midsagittal description of the VT would also be a serious limitation of the model for some sounds such as the lateral /l/. During the production of /l/ there could be a lateral occlusion in the midsagittal plane which would result in zero areas in the area function computed using $\alpha(x)$ and $\beta(x)$. But the area function is actually not zero due to the presence of lateral channels along the side of the occlusion. Asymmetric lateral channels would also

lead to zeros in the speech signal, which is not captured by a midsagittal model. The inversion of laterals using a midsagittal model would therefore be very unreliable. Inversion errors for vowels and diphthongs could also be higher in lateral contexts.

In this paper, we have considered limited types of adaptation of the Maeda model to the speaker - only vocal tract length scaling and modification of the outer VT outline. Results could be improved with more information about the VT geometry for the given speaker, mainly the entire outer VT outline consisting of the hard and soft palates and rear pharyngeal wall extending down to the laryngeal region. The XRMB database does not include information on the soft palate (velum) and on the laryngeal region, which are limiting factors in our experiments since the velum outline had to be interpolated, and the length of the pharyngeal region was adapted in an ad hoc manner based on the acoustic match during calibration. By optimizing the calibration cost function (Equation 22) using codebook search and convex optimization, we verified that the adapted model was well calibrated for all the test phonemes of speaker JW11.

However, our unsuccessful attempt at calibrating the Maeda model for speaker JW46 indicated that superficial adaptation of the Maeda model was insufficient for this speaker. The coefficients $\alpha(x)$ and $\beta(x)$ and the VT outline basis vectors (deformation modes) of the Maeda model vary with speaker and can cause large inversion errors for some speakers if they are not adapted. The parameters used in calculating the chain matrix of a tube section may also be adapted. The approach we have developed in this paper has the advantage that it can be extended without much difficulty to optimize all these different parameters. This is a topic of future work.

The mapping from midsagittal widths to areas using the coefficients $\alpha(x)$ and $\beta(x)$ and Equation 2 is ad hoc, and inaccuracies are possible at different ranges of midsagittal widths, for example at very small and very large midsagittal widths. The shift of the measured XRMB lip pellets for comparison with estimated VT is also ad hoc. While it is only used for a visual comparison and not for quantitative evaluation in the results section of the paper, the error in the lip height and protrusion is included in the calibration cost function. Since

the lip opening affects the lip aperture and radiation impedance and therefore the computed formants, this may also be a source of error in the inversion results. Investigation of these issues will also be a topic of future work.

A mapping from XRMB pellet positions to Maeda articulatory parameters would be very useful in learning correlations between articulatory parameters and better articulatory constraints, perhaps also across speakers. With such a mapping, estimated articulatory parameter trajectories could also be compared with actual ones. The optimization-based method in Toutios et al.[40] could give such a mapping, provided the model is well calibrated to the speaker, as discussed in Section VI.

We have not used any a priori model of articulatory dynamics, and used only the constraints provided by the articulatory model and the regularization and continuity terms in the cost function. The inversion could be improved by using a model of articulatory dynamics such as the task dynamic model from gestural phonology, where the fundamental units of speech production are taken to be gestures, which are the coordinated action of articulators[42].

In summary, the methods that we have proposed and discussed in the paper provide an improved understanding of speech production, of the limitations of articulatory and acoustic models, and of inversion by analysis-by-synthesis. We proposed a systematic framework for calibration and adaptation of the Maeda model to new speakers with XRMB data, by optimizing a novel cost function. The optimizations for model adaptation and inversion by analysis-by-synthesis used an elegant and efficient calculation of the derivatives of the chain matrix of a tube with respect to its area. A quantitative study of inversion of vowels and diphthongs was performed, and results were significantly improved by codebook pruning. Good match between estimated midsagittal VT outlines and measured XRMB tongue pellet positions was achieved for several vowels and diphthongs, with average pellet-VT outline distances around 0.15 cm.

Acknowledgments

This work was supported in part by the NSF. We thank John Westbury of the University of Wisconsin for sharing the X-Ray Microbeam Database, which was supported in part by R01 DC 00820 from the NIDCD. We used Matlab programs for both the Maeda model and chain matrices written by Edward Riegelsberger[14]. We thank the editor and the three anonymous reviewers for helping to improve the quality of this paper.

References

[1] B. S. Atal and O. Rioul, "Neural networks for estimating articulatory positions from speech", J. Acoust. Soc. Am. Suppl.1 **86**, S67 (1989).

[2] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi, "On the use of neural networks in articulatory speech synthesis", J. Acoust. Soc. Am. **93**, 1101–1121 (1993).

[3] 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", J. Acoust. Soc Am **92**, 688–700 (1992).

[4] S. Dusan, "Statistical estimation of articulatory trajectories from the speech signal using dynamic and phonological constraints", Ph.D. thesis, University of Waterloo (2000).

[5] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal", Ph.D. thesis, U. Edinburgh (2001).

[6] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model", Speech and Audio Processing, IEEE Transactions on **12**, 175–185 (2004).

[7] J. Hogden, P. Rubin, E. McDermott, S. Katagiri and L. Goldstein, "Inverting mappings from smooth paths through Rn to paths through Rm: A technique applied to recovering articulation from acoustics", Speech Communication **49**, 361-383 (May 2007).

[8] A. Lammert, D. Ellis and P. Divenyi, "Data-driven articulatory in-

version incorporating articulator priors", Proc. SAPA-08, 29-34 (2008) http://www.sapa2008.org/papers/127.pdf (date last viewed: 10/05/10).

[9] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", JASA **63**, 1535–1555 (1978).

[10] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Signal models for low bit-rate coding of speech", J. Acoust. Soc. Am. **68**, 780–791 (1980).

[11] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave", Speech Communication **5**, 159–170 (1986).

[12] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal", IEEE Trans. SAP **2**, 133–150 (1994).

[13] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production", in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sondhi, 231–267 (Marcel Dekker, New York) (1992).

[14] E. L. Riegelsberger, "The acoustic-to-articulatory mapping of voiced and fricated speech", Ph.D. dissertation, The Ohio State University (1997).

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

[16] P. Mermelstein, "Determination of vocal tract shapes from measured formant frequencies", J. Acoust. Soc. Am. **41**, 1283–1294 (1967).

[17] C. Qin and M. A. Carreira-Perpinan, "An empirical investigation of the non-uniqueness in the acoustic-to-articulatory mapping", in *Proc. Interspeech 2007*, 74-77 (2007) http://faculty.ucmerced.edu/mcarreira-perpinan/papers/interspeech07a.pdf (date last viewed: 10/05/10).

[18] V. N. Sorokin, "Determination of vocal tract shape for vowels", Speech Communication **11**, 71–85 (1992).

[19] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion", J. Acoust. Soc. Am. **118**, 444–460 (2005).

[20] B. Juang, L. Rabiner, and J. Wilpon, "On the use of bandpass liftering in speech recognition", IEEE Trans. Acoustics, Speech, and Signal Processing **35**, 947–954 (1987).

[21] J. Schroeter, P. Meyer, and S. Parthasarathy, "Evaluation of improved articulatory codebooks and codebook access distance measures", in *Proc. IEEE ICASSP*, 393–396 (1990).

[22] P. Mermelstein, "Articulatory model for the study of speech production", J. Acoust. Soc. Am. **53**, 1070–1082 (1973).

[23] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model", in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal, 131–149 (Kluwer Academic, Dordrecht) (1990).

[24] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests", Speech Communication **14**, 19–48 (1994).

[25] V. N. Sorokin and A. V. Trushkin, "Articulatory-to-acoustic mapping for inverse problem", Speech Communication **19**, 105–118 (1996).

[26] B. Potard, Y. Laprie and S. Ouni, "Incorporation of phonetic constraints in acoustic-to-articulatory inversion", J. Acoust. Soc. Am. **123**, 2310-2323 (2008).

[27] J. R. Westbury, *X-ray Microbeam Speech Production Database User's Handbook, Version 1.0*, 1-135, (Waisman Center, University of Wisconsin, Madison) (June 1994).

[28] `http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html` (date last viewed 10/05/10).

[29] J. M. Heinz and K. N. Stevens, "On the derivation of area functions and acoustic spectra from cineradiographic films of speech", J. Acoust. Soc. Am. **36**, 1037 (1964).

[30] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer", IEEE Trans. ASSP **35**, 955–967 (1987).

[31] J. Flanagan, *Analysis, Synthesis, and Perception of Speech*, 2nd edition, 36-38 (Springer-Verlag, Berlin) (1972).

[32] Z. Zhang, C. Espy-Wilson and M. Tiede, "Acoustic Modeling of American English Lateral

Approximants", Proceedings of Eurospeech 2003, Switzerland, (2003).

[33] S. Panchapagesan, "Modeling the Production of /l/ Based on MRI data", M.S. thesis, University of California, Los Angeles (2003).

[34] M. M. Sondhi, "Model for wave propagation in a lossy vocal tract", J. Acoust. Soc. Am. **55** (1974).

[35] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC", Computer Speech and Language **23**, 42–64 (2009).

[36] J. Makhoul, "Spectral Linear Prediction: Properties and Applications", IEEE Trans. ASSP **23**, 283–296 (1975).

[37] T. Vampola, J. Horacek and J. G. Svec "FE Modeling of Human Vocal Tract Acoustics. Part I: Production of Czech Vowels", Acta Acustica united with Acustica **94(3)**, 433–447 (2008).

[38] M. S. Gockenbach, "Online lectures on numerical optimization", Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan (Spring 2005), `http://www.math.mtu.edu/~msgocken/ma5630spring2005/lectures.html` (date last viewed 10/05/10).

[39] S. Panchapagesan, "Frequency warping by linear transformation, and vocal tract inversion for speaker normalization in automatic speech recognition", Ph.D. dissertation, University of California, Los Angeles (2008), `http://www.ee.ucla.edu/~spapl/paper/panchi_dissertation.pdf` (date last viewed: 11/28/09).

[40] A. Toutios, S. Ouni and Y. Laprie, "Protocol for a Model-based Evaluation of a Dynamic Acoustic-to-Articulatory Inversion Method using Electromagnetic Articulography", in *ISSP 2008*, 317-320 (2008), http://issp2008.loria.fr/Proceedings/PDF/issp2008-73.pdf (date last viewed: 11/28/09).

[41] B. Mathieu and Y. Laprie, "Adaptation of Maeda's model for acoustic to articulatory inversion", in *Proceedings of Eurospeech*, 2015–2018 (1997).

[42] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production", Ecological Psychology **1**, 333–382 (Lawrence Erlbaum Associates) (1989).

Vowel		Avg. Formant Error		Avg. Tongue Pellet-VT outline distance	
DARPA	IPA	Codebook	Optimized	Codebook	Optimized
OW	oʊ	3.72%	2.10%	0.14 cm	0.12 cm
EH	ɛ	2.20%	0.10%	0.21 cm	0.17 cm
AH	ʌ	2.23%	0.12%	0.15 cm	0.17 cm
EY	eɪ	1.73%	0.22%	0.26 cm	0.21 cm
UX	u	4.08%	0.21%	0.27 cm	0.25 cm
IH	ɪ	2.14%	0.36%	0.27 cm	0.30 cm
IY	i	1.64%	0.07%	0.28 cm	0.30 cm
AA	ɑ	3.47%	0.43%	0.22 cm	0.31 cm
AE	æ	0.84%	0.22%	0.38 cm	0.42 cm
AX	ə	3.83%	1.32%	0.35 cm	0.58 cm
AO	ɔ	4.31%	0.19%	0.66 cm	0.65 cm
Avg.		2.75%	0.49%	0.29 cm	0.32 cm

TABLE I. Task 14 of Speaker JW11, inversion errors after codebook search and convex optimization. The vowels are arranged in increasing order of average Tongue Pellet-VT outline distance after optimization. Vowel labels are given in both DARPA and IPA formats.

Vowel		Avg. Formant Error		Avg. Tongue Pellet-VT outline distance	
DARPA	IPA	Codebook	Optimized	Codebook	Optimized
AY	aɪ	4.08%	1.21%	0.41 cm	0.49 cm
OY	oɪ	2.64%	0.42%	0.27 cm	0.33 cm
AW	aʊ	2.87%	0.37%	0.16 cm	0.26 cm
EY	eɪ	2.38%	0.34%	0.15 cm	0.12 cm
Avg.		2.99%	0.58%	0.25 cm	0.30 cm

TABLE II. Task 13 of Speaker JW11, inversion errors after codebook search and convex optimization. Phoneme labels are given in both DARPA and IPA formats. The diphthongs were from the words/nonwords "side, soid, sowd and sayed" respectively.

Vowel Sequence	Avg. Formant Error		Avg. Tongue Pellet-VT outline distance	
	Codebook	Optimized	Codebook	Optimized
iu	3.64%	0.16%	0.17 cm	0.19 cm
ui	6.20%	2.20%	0.22 cm	0.26 cm
iɑ	3.04%	0.32%	0.13 cm	0.11 cm
ɑi	2.28%	0.33%	0.14 cm	0.10 cm
uɑ	5.80%	1.57%	0.28 cm	0.24 cm
ɑu	5.91%	2.43%	0.32 cm	0.20 cm
Avg.	4.48%	1.17%	0.21 cm	0.18 cm

TABLE III. Task 15, vowel sequences of Speaker JW11, inversion errors after codebook search and convex optimization.

List of Figures