RECOVERING VOCAL TRACT SHAPES FROM MFCC PARAMETERS

Sorin Dusan and Li Deng

Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

Recovering vocal tract shapes from the speech signal is a well known inversion problem of transformation from the articulatory system to speech acoustics. Most of the studies on this problem in the past have been focused on vowels. There have not been general methods effective for recovering the vocal tract shapes from the speech signal for all classes of speech sounds. In this paper we describe our attempt towards speech inverse mapping by using the mel-frequency cepstrum coefficients to represent the acoustic parameters of the speech signal. An inversion method is developed based on Kalman filtering and a dynamic-system model describing the articulatory motion. This method uses an articulatory-acoustic codebook derived from Maeda's articulatory model.

1. INTRODUCTION

Estimation of articulatory positions and movements from speech acoustics is commonly called *inverse mapping* in speech research, as it is an inverse of the natural transformation from articulators to speech acoustics. The main difficulties of the acoustic-to-articulatory mapping are due to the nonlinear and one-to-many characteristics of this inverse transformation. Some of the approaches to speech inversion have used an analytical nonlinear function for modeling the articulatory-to-acoustic transformation, whereas others used articulatory-acoustic codebooks derived from articulatory models or measurements of articulatory and acoustic parameters from human beings. There were many attempts of estimating the vocal tract shapes from the formant frequencies of the speech signal, but these parameters are not representative for all classes of speech sounds. The non-unique solution of the acoustic-to-articulatory mapping have motivated researchers to find optimal articulatory trajectories and vocal tract shapes by imposing dynamic constraints.

The inverse mapping in speech is a difficult and unsolved problem. Satisfactory solutions to this problem will have both theoretical and practical significance. It would help the motor theory of speech production, the articulatory phonology and have applications in speech and speaker recognition, speech synthesis, speech coding and teaching deaf people to speak.

Among the first researchers who approached this problem were Mermelstein and Schroeder who proposed methods of estimating the area function from formant frequencies [5]. Sondhi and Gopinath proposed a method of determination of vocal tract shape from impulse response at the lips [12]. A new method using the inverse filtering of the acoustic speech waveforms has been suggested by Wakita [14]. Shirai and Honda studied the estimation of articulatory motion using an articulatory dynamical model and nonlinear filtering [11]. They have used a nonlinear observation function relating the formant frequencies to articulatory parameters.

A theoretical study of speech inversion has been done by Atal et al. [1], using a computer sorting technique. They studied the acoustic-to-articulatory relationship by sampling the whole space of an articulatory model and creating the articulatory sets of vectors called fibers which map into the same acoustic vector. A study of estimation of articulatory trajectories using Kalman filtering has been done by Wilhelms et al. [15]. They used as acoustic features the short time spectra and experimented the method on vowels and some limited consonants. Schroeter et al., [9] proposed a method of estimating the articulatory parameters using a vocal tract/cord model and an articulatory-acoustic codebook. In that study they have used the LPC parameters as acoustic vectors and sampled the articulatory space between pairs of root shapes. This work has been extended later to a multi-frame approach. More recently, Schroeter and Sondhi [10] presented a method based on dynamic programming to search the articulatory codebooks. They have used the LPC derived cepstral coefficients as acoustic feature and introduced a lifter in computation of the acoustic distance and a dynamic cost in making a transition from a vocal tract shape to another one. Papcun et al. [6] further studied the inversion problem with a neural network trained on X-ray microbeam data. An optimization method based on conditional minimum efforts has been used by Sorokin [13] for determination of vocal tract shape for vowels from formant frequencies.

Ramsay and Deng [7] proposed a stochastic target model for estimating the articulatory parameters. They used the EM algorithm for estimating model parameters and Kalman smoothing to estimate the articulatory states. Another work using the dynamic programming search has been presented by Richards et al. [8]. They attempted to estimate the articulatory representation of speech using the cepstral coefficients and a codebook derived from the Distinctive Regions Model. A method of recovering articulator positions from acoustics based on human articulatory-acoustic data has been published by Hogden et al. [4]. They have used a vector-quantization method to build different articulatory-acoustic codebooks.

Improving our earlier method for estimating articulatory parameters from formant frequencies using the Iterated Extended Kalman filtering technique [3], we in this paper describe our new experiments on recovering vocal tract shape and its dynamics for vowels using the mel-frequency cepstrum coefficients (MFCC) as the acoustic measurement. The main contribution of this paper is selection and creation of articulatory-acoustic data and the implementation of filtering and smoothing techniques.

2. ARTICULATORY MODEL

In order to use the Extended Kalman Filtering we linearized the articulatory-to-acoustic function on small regions using an articulatory-acoustic codebook. To create this codebook we have used the Maeda's static articulatory model built by statistical analysis of X-ray films of a French female speaker. This articulatory model constructs the vocal tract shape from eight linear components representing the jaw, tongue body, tongue dorsum, tongue tip, lips and hight of pharynx. From these parameters the vocal tract area function is computed and a lossy vocal tract model transforms the area function into the vocal tract transfer function. We used an all-poles model of the vocal tract transfer function.

3. ARTICULATORY AND ACOUSTIC DATA

Our main idea for the acoustic-to-articulatory mapping is to use low dimensional parameters whose components are orthogonal to represent both articulatory and acoustic vectors. In this work we have chosen the Maeda's articulatory parameters (which are orthogonal to each other and explain most of the vocal tract data variance) and the MFCC parameters (constructed from orthonormal functions). The articulatory-acoustic nonlinear function $\mathbf{h}$ relating the articulatory vectors $\mathbf{x}$ to the acoustic vectors $\mathbf{y}$ is defined by the equation: $\mathbf{y} = \mathbf{h}(\mathbf{x})$. This analytical function is of many-to-one type and practically has been proved to be so by many articulatory compensation experiments. For this reason we did not create, as in other studies, an acoustic-articulatory codebook to search for each acoustic frame the closest acoustic entry in the codebook and get the corresponding articulatory parameters describing the recovered vocal tract shape. Instead, we created an articulatory-acoustic codebook in which many possible entries of articulatory parameters map into the same acoustic vector. In this way we allow each acoustic frame to be produced by different vocal tract shapes, as in natural speech this occurs due to compensatory articulation. The selection of the optimal vocal tract shape from all candidate shapes has been done introducing dynamic constraints by the dynamical model.

The articulatory parameters used to construct the $(\mathbf{x}, \mathbf{y})$ pairs of the codebook were the eight articulatory model parameters, whereas for the acoustic parameters we used the MFCCs. It is well known that the MFCCs are among the best acoustic features used in automatic speech recognition. The MFCCs are robust, contain much information about the vocal tract configuration regardless the source of excitation, and can be used to represent all classes of speech sounds. We devised a method of computing the MFCC parameters using a filterbank from both speech signal and vocal tract shape. This is because in our analysis-by-synthesis procedure we have to minimize the acoustic distance between the measured speech spectra and the model speech spectra. In both cases, from the all-poles LP models we computed the log energy spectrum and then applied it to a filterbank composed of critical band filters. The outputs of these filters were used to compute the MFCCs after multiplication with some orthonormal functions. We used 10 low-order MFCCs, not including the zero-th order that represents the log energy.

In our previous work, [3], we have created an articulatory-acoustic codebook only from middle vowels, and the transitions to and from vowels were not accurately modeled. In the current work, we constructed a separate articulatory-acoustic codebook by randomly sampling the articulatory space. The initial data points in the sampling represent 392 open vocal tract shapes selected from a total of 519 shapes from which the Maeda's articulatory model has been built. Subsequently, we created for each of the 392 original vocal tract shapes many vocal tract shapes which map approximately into the same acoustic vector as the original shape does. These simulated shapes can be very different and their corresponding vocal tract transfer functions are not exactly the same. Hence a fine covering of the acoustic space of this codebook has been accomplished. The entire articulatory-acoustic codebook we have created contains a total of 235,000 pairs of articulatory and acoustic vectors. The histograms of the 235,000 articulatory vectors (8 dimensions) from the codebook are shown in Fig. 1. The corresponding histograms of the MFCC vectors (10 dimensions) are shown in Fig. 2. This codebook is used to characterize the nonlinear function $\mathbf{h}$, which is linearized on many small regions using a clustering algorithm and a vector quantization (VQ) technique. The result of the VQ-clustering gives a total of 10,000 piecewise-linear regional models, which jointly approximate $\mathbf{h}$. For training the model parameters and recovering the vocal tract shapes we have used vowel tokens from TIMIT database and articulatory-acoustic data recorded with an electromagnetic midsagittal articulograph (EMMA).

4. METHODS

For estimating the dynamical model parameters we implemented the same method as in [2, 7]. We have used the Expectation-Maximization (EM) algorithm for ML estimation of model parameters. To model the dynamics of the articulators we used second-order critically damped linear models. These can be augmented to the first order state equation:

$$\mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k + \mathbf{w}_k \tag{1}$$

where $\mathbf{F}$ is the transition matrix and $\mathbf{w}$ is a white Gaussian noise with covariance matrix $\mathbf{Q}$. By expanding the nonlinear function $\mathbf{h}(\mathbf{x})$ in a Taylor series about a reference $\bar{\mathbf{x}}_k$,

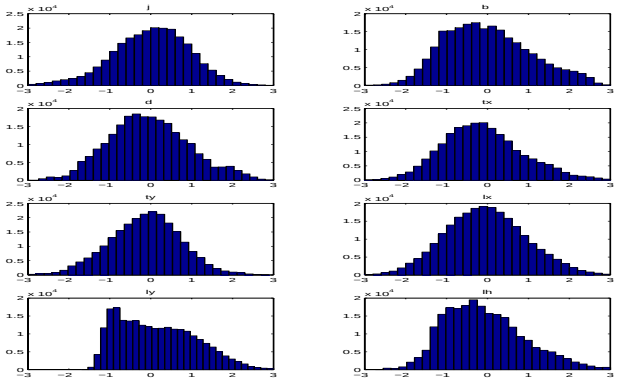

Figure 1: Distribution of articulatory vectors

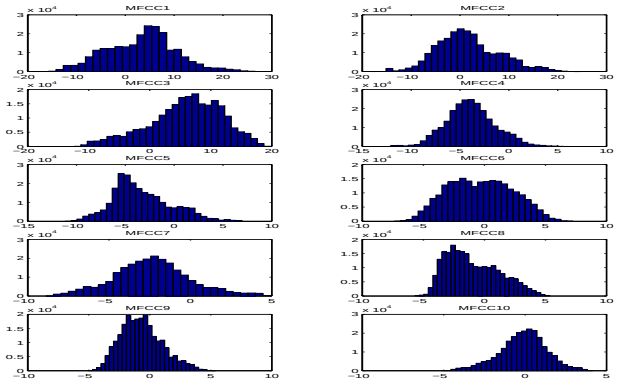

Figure 2: Distribution of acoustic vectors

as in [11], we obtained the linearized output equation:

$$\mathbf{y}_k = \mathbf{h}(\bar{\mathbf{x}}_k) + \mathbf{H}(\bar{\mathbf{x}}_k)(\mathbf{x}_k - \bar{\mathbf{x}}_k) + \mathbf{v}_k \qquad (2)$$

where $\mathbf{H}$ is the Jacobian matrix of $\mathbf{h}$ and $\mathbf{v}$ is a white Gaussian noise with covariance matrix $\mathbf{R}$. The two processes $\mathbf{w}$ and $\mathbf{v}$ are supposed to be uncorrelated. For each linear region a matrix $\mathbf{H}$ and a mean acoustic vector has been computed. Because of the nonlinearity in the function $\mathbf{h}$, the conditional probability of articulatory states $\mathbf{x}$ given the observations $\mathbf{Y}_N$ is not Gaussian and the EM algorithm will only converge to an approximate ML estimate of model parameters.

The EM algorithm iteratively estimates the parameters θ (including matrices $\mathbf{F}$, $\mathbf{Q}$ and $\mathbf{R}$) by maximizing the log-likelihood objective function:

$$\begin{aligned} J(\mathbf{X},\mathbf{Y},\theta) &= log\{L(\mathbf{X},\mathbf{Y},\theta)\} = \\ &-\frac{1}{2}\sum_{m=1}^{M}\sum_{k=0}^{N_m-1}\{(\mathbf{x}_{k+1}-\mathbf{F}\mathbf{x}_k)^T\mathbf{Q}^{-1}(\mathbf{x}_{k+1}-\mathbf{F}\mathbf{x}_k)\} \\ &-\frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{N_m}\{[\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)]^T\mathbf{R}^{-1}[\mathbf{y}_k-\mathbf{h}(\mathbf{x}_k)]\} \\ &-\frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{N_m}\{log|\mathbf{Q}|+log|\mathbf{R}|\}+constant \end{aligned} \qquad (3)$$

The training we have used is based on multiple observation sequences. The number of sequences M is a function of occurrences for each vowel in the training utterances, whereas the number of frames N_m varies according to the observation sequence m. In maximizing the objective function, we have used different expectations of the states given the observation sequences. These expectations have been computed using the Kalman filtering and smoothing.

The recovery of vocal tract shapes is based on estimating the articulatory states for each frame of the test data. After training the model parameters we have used the Iterated Extended Kalman filtering and smoothing techniques for estimating the articulatory states.

5. EXPERIMENTAL RESULTS

We performed some preliminary experiments of estimating the model parameters and the vocal tract shapes for 10 English vowels (/aa/, /ae/, /ah/, /ao/, /eh/, /ey/, /ih/, /iy/, /uh/ and /uw/) from the utterances of a female speaker from TIMIT database. The vowel tokens have been divided into the training and test sets. The selection of the speaker was based on data fitting with the Maeda's model female speaker in the two-dimensional space formed by the frequencies of the F1 and F2 formants. The EM algorithm for estimation of model parameters has been used for 10 iterations, with the algorithm convergence consistently being observed. The MFCCs have been computed for frames of 32 ms, with 10 ms frame shift, after preemphasis and Hamming windowing. We trained the models from these short vowel tokens without taking into account the preceding sounds for each of them. Because of that, we added before each observation sequence a simulated starting sequence. These starting sequences were built by linear interpolation between the corresponding MFCCs of the mean articulatory vector of the codebook and those of the first frame for each of the observation sequences. The mean articulatory vector of the codebook was close to zero, hence the transition of the articulatory parameters from this initial state to the first state of each observation sequence was smooth.

In Fig. 3 we show an example of recovering the vocal tract shapes from MFCCs of a TIMIT vowel /aa/. From the 16 MFCC frames of the /aa/ token we estimated the trajectories of the 8 articulatory parameters. From these parameters we recovered the vocal tract area functions and transfer functions as plotted in this figure. An example of recovering vocal tract shapes for an /ey/ token from EMMA is presented in Fig. 4. The Maeda's estimated articulatory parameters cannot be compared directly with the EMMA measurements. Instead we compared the vocal tract shapes derived from these two methods. In the experiments using the model parameters estimated from TIMIT data and the articulatory-acoustic data measured with an EMMA from a female speaker, we have found that the estimated vocal tract shapes are consistent with the ones derived from the actual EMMA measurements. In our experiments, we found that the trajectories of the estimated articulatory parameters from MFCC parameters are as smooth as those obtained using the formant frequencies,

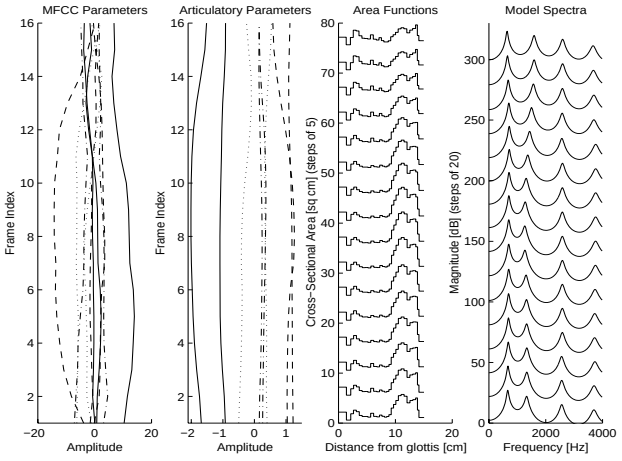

Figure 3: Recovered vocal tract shapes for /aa/ (TIMIT)

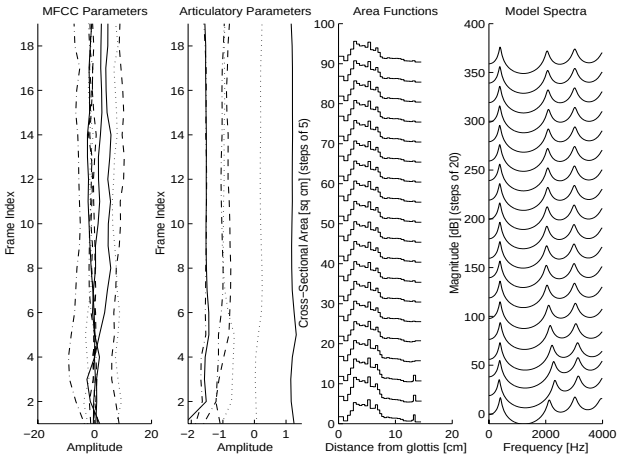

Figure 4: Recovered vocal tract shapes for /ey/ (EMMA)

even though the MFCC trajectories are not as smooth as those of formant frequencies.

6. SUMMARY

We present in this paper preliminary results of our work towards improving the recovery of vocal tract shapes from the speech signal, using the MFCC parameters. We have used the EM algorithm, with the E-step accomplished by the Iterated Extended Kalman filtering and smoothing, to estimate the model parameters. The method has been shown to be successful for vowel tokens in TIMIT data. The use of the method for all other classes of speech sounds is currently underway.

7. ACKNOWLEDGMENTS

We would like to thank A. Galvan, J. Ma, and G. Ramsay for implementation of Maeda's articulatory model, implementation of the VQ algorithm, and for discussions. This work is supported in part by NSERC, Canada, the Ontario Government, and by ICR of Univ. of Waterloo.

8. REFERENCES

1. 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(5):1535–1555, 1978.
2. V. Digalakis, J.R. Rohlicek, and M. Ostendorf. ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition. *IEEE Trans. SAP*, 1(4):431–442, 1993.
3. S. Dusan and L. Deng. Estimation of Articulatory Parameters from Speech Acoustics by Kalman Filtering. In *Proc. of CITO Researcher Retreat-Hamilton Canada*, 1998.
4. J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. Accurate Recovery of Articulator Positions from Acoustics - New Conclusions Based on Human Data. *JASA*, 100(3):1819–1834, 1996.
5. P. Mermelstein and M. R. Scroeder. Determination of Smoothed Cross-Sectional Area Functions of the Vocal Tract from Formant Frequencies. In D.E. Commins, editor, *Proceedings of the Fifth International Congress on Acoustics*, volume 1a., 1965.
6. 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. *JASA*, 92(2):688–700, 1992.
7. G. Ramsay and L. Deng. Maximum-Likelihood Estimation for Articulatory Speech Recognition Using a Stochastic Target Model. In *Proc. EUROSPEECH'95*, pages 1401–1404, 1995.
8. H. Richards, J. Mason, M. Hunt, and J. Bridle. Deriving Articulatory Representations of Speech. In *Proc. EUROSPEECH'95*, pages 761–764, 1995.
9. J. Schroeter, J.N. Larar, and M.M. Sondhi. Speech Parameter Estimation Using a Vocal Tract /Cord Model. In *ICASSP*,pages 308–311, 1987.
10. J. Schroeter and M.M. Sondhi. Dynamic Programming Search of Articulatory Codebooks. In *ICASSP*, pages 588–591, 1989.
11. K. Shirai and M. Honda. Estimation of Articulatory Motion. In *Dynamic Aspects of Speech Production*, pages 279–302. Tokyo University Press, 1976.
12. M.M. Sondhi and B. Gopinath. Determination of the Vocal-Tract Shape from Impulse Response at the Lips. *JASA*, 49(6):1867–1873, 1971.
13. V. Sorokin. Determination of Vocal Tract Shape for Vowels. *Speech Communication*, 11(1):71–85, 1992.
14. H. Wakita. Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms. *IEEE Trans. Audio Electroacoust.*, AU-21:417–427, 1973.
15. R. Wilhelms, P. Meyer, and H. W. Strube. Estimation of Articulatory Trajectory by Kalman Filter. In I.T. Young et al., editor, *Signal Processing III: Theories and Applications*, pages 477–480, 1986.