ARTICULATORY INVERSION AND SYNTHESIS: TOWARDS ARTICULATORY-BASED MODIFICATION OF SPEECH

Sandesh Aryal and Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University
{sandesh,rgutier}@cse.tamu.edu

ABSTRACT

Certain speech modifications, such as changes in foreign/regional accents or articulatory styles, are performed more effectively in the articulatory domain than in the acoustic domain. Though measuring articulators is cumbersome, articulatory parameters may be estimated from acoustics through inversion. In this paper, we study the impact on synthesis quality when articulators predicted from acoustics are used in articulatory synthesis. For this purpose, we trained a GMM articulatory synthesizer and drove it with articulators predicted with an RBF-based inversion model. Using inverted instead of measured articulators degraded synthesis quality, as measured through Mel cepstral distortion and subjective tests. However, retraining the synthesizer with predicted articulators not only reversed the effect of errors introduced during inversion but also improved synthesis quality relative to using measured articulators. These results suggest that inverted articulators do not compromise synthesis quality, and open up the possibility of performing speech modification in the articulatory domain through inversion.

Index terms– articulatory synthesis, articulatory inversion, speech modification, Maeda parameters

1. INTRODUCTION

In order to modify certain characteristics of speech such as duration, pitch, speaker identity and articulation styles, we must first decouple them from other factors that make up the speech signal. Some of these characteristics, such as duration and pitch, are easily extracted in the acoustic domain. Others, such as regional/foreign accents and articulation styles, are more challenging since speaker-dependent and linguistic information interact in complex ways when analyzing the formant structure of the utterance. These two sources of information, however, may be easily decoupled in articulatory space [1]. For this reason, researchers have incorporated articulatory parameters in a variety of speech modification problems such as voice transformation [2], foreign accent conversion [3], and flexible text-to-speech synthesis [4].

However, current technologies that collect articulatory parameters are impractical outside laboratory settings. These technologies, such as X-Ray Microbeam, ultrasound, electropalatography, and, electromagnetic articulography (EMA) are invasive, and in the case of X-ray microbeam also dangerous. In order to avoid the cumbersome process of measuring articulatory parameters, researchers have proposed several methods to invert articulatory parameters from the acoustic signal [5-8]. Inverted articulatory features have been found useful for speech recognition [9-11], but their effectiveness in speech modification is not well studied.

As a first step toward using articulatory inversion in speech modification, this article investigates the impact on synthesis quality of replacing measured articulators with predictions from articulatory inversion. Namely, we predict Maeda articulatory features [12, 13] from speech acoustics (MFCCs) using an RBF-based inversion method [5]. Then, we use a GMM-based articulatory synthesizer [6] to synthesize speech from either measured or predicted articulators. Finally, we compare these two types of synthesis using objective measures (Mel cepstral distortion) and subjective evaluation (listening tests).

Relation to prior work. Our work is most related to previous studies that incorporated articulatory parameters in speech synthesis and speech modification [3, 4, 6]. These previous studies, however, used directly measured articulatory parameters – see section 2 for a detailed discussion. In contrast, our study uses articulatory parameters predicted from acoustics through inversion. Also related to our work are models of infant motor learning based on articulatory inversion/synthesis [14, 15]. Because these studies focus on the process of motor learning, they generally use synthetic speech or restricted natural utterances (e.g., vowel/consonant patterns, babbling). In contrast, our work uses natural speech containing complete sentences.

The paper is organized as follows. In section 2 we review related work on speech modification in the articulatory domain. Section 3 describes the articulatory inversion model and the data-driven articulatory synthesizer we used in this work. In section 4 we compare the quality of the resulting speech synthesis when using actual articulators or predicted articulators.

2. RELATED WORK

A few studies have shown how to incorporate articulatory control for modifying speech characteristics [4, 6, 16]. Toda et al. [6] proposed a data-driven language-independent method for flexible articulatory speech synthesis. The authors used a GMM-based forward mapping to estimate acoustic parameters (Mel cepstral coefficients) from articulatory parameters (seven EMA positions, pitch and loudness). Then, they manipulated the EMA positions to simulate the effect of speaking with the mouth wide open. As a result of this manipulation, the authors observed a loss of high frequency components in fricatives. Though the articulatory manipulation was effective in modifying speech characteristics, it also reduced the synthesis quality compared to driving the GMM-based forward mapping with unmodified articulators. Ling et al. [4] showed that incorporating articulatory parameters in a HMM-based

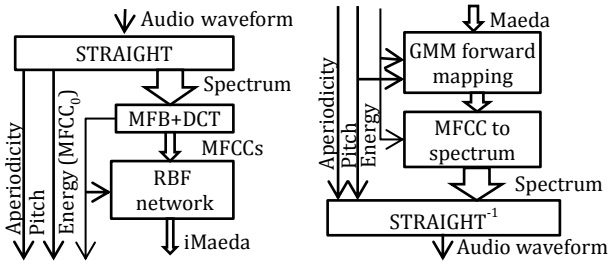


Figure 1: Block diagrams of the articulatory (a) inversion and (b) synthesis methods

synthesizer [17] improved synthesis quality, as opposed to using only text input. The authors used a five-state left-to-right HMM structure with no skip to train context-dependent phoneme models on a combination of articulatory (six EMA positions) and acoustic features (40th-order frequency-warped LSFs). The output distribution of acoustic parameters was modeled as a Gaussian distribution with the mean value given by a linear function of the articulatory parameters and the state-specific parameters. The authors showed the feasibility of modifying vowels by manipulating articulatory parameters alone. As an example, increasing the tongue-height parameters led to a clear shift in vowel perception from /ɛ/ to /ɪ/ in synthesis. Similarly, decreasing the tongue-height parameters led to a shift from /ɛ/ to /æ/. Improvements in synthesis quality relative to Toda et al. [6] come from the use of phonetic information and the ability of HMMs to model the temporal properties of speech better than a GMMs. In a recent study [3], we used articulatory parameters to convert utterances from a non-native speaker so they sounded more native-like. Our approach consisted of identifying mispronounced or accented diphones in the non-native utterances, and replacing them with units from the non-native speaker such that the substitute units closely matched the articulatory trajectories of a native speaker. Our method was able to reduce the perceived accentedness of the non-native utterance, though the reduction was limited by the availability of target units in the non-native corpus.

These previous studies illustrate the feasibility of performing speech modifications in the articulatory domain, assuming articulatory measurements are available. Though this is rarely the case, articulatory-based modification of speech may still be possible if acoustic features can be mapped accurately into the articulatory domain (i.e., through inversion). As a step towards this objective, the present study seeks to understand the effect of replacing measured articulators with inverted articulators. For this purpose, we use the GMM-based articulatory synthesizer of Toda et al. [6] since it does not require access to the phonetic transcription; this allows us to focus on issues in synthesis quality that are due exclusively to articulatory information.

3. ARTICULATORY INVERSION AND SYNTHESIS METHODS

To evaluate the effectiveness of inverted articulatory features we used the articulatory inversion and articulatory synthesis strategy outlined in Figure 1. Our articulatory inversion method predicts Maeda articulatory features from the audio signal through the following four steps. First, we extract *pitch* (f_0), *aperiodicity* and *spectral envelope* using STRAIGHT [18]. In a second step, we compute Mel Frequency Cepstral Coefficients (MFCCs) by warping the STRAIGHT spectral envelope according to the Mel-frequency scale and then applying a type-II discrete cosine transformation (DCT). Then, we map MFCCs into Maeda parameters with an RBF network; see section 3.1 for details. Finally, we smooth the trajectory of inverted Maeda parameters with a low-pass filter to match the natural smoothness of measured Maeda trajectories. In what follows, we refer to the filtered inverted Maeda parameters as *iMaeda* to differentiate them from the actual Maeda parameters.

Our articulatory synthesis method involves three steps, as illustrated in Figure 1(b). First, we use a GMM-based forward mapping to estimate spectral features ($MFCC_{1\text{-}24}$) from articulatory features (Maeda and delta-Maeda), log pitch ($\ln(f_0)$), and the energy parameter ($MFCC_0$). In a second step, we reconstruct the STRAIGHT *spectral envelope* from the estimated spectral coefficients ($MFCC_{1\text{-}24}$) and the energy parameter ($MFCC_0$) in the original speech. Specifically, given a vector of predicted MFCCs, the least-squares estimate of the spectral envelope is $\hat{\boldsymbol{s}} = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{e}$, where $\boldsymbol{F}$ is the Mel Frequency Filter Bank (MFB) matrix used to extract MFCCs from the STRAIGHT spectrum, and $\boldsymbol{e}$ is the exponential of the inverse DCT of MFCCs. In a final step, we use the STRAIGHT synthesis engine to generate the waveform using the estimated *spectral envelope*, *aperiodicity* and *pitch* (f_0).

In the following subsections, we provide details of the RBF-based articulatory inversion model and the GMM-based forward mapping model.

3.1 Inversion model

Following Chao and Carreira-Perpiñán [5], we use an RBF-based inversion model. Given a static acoustic feature vector $\boldsymbol{y}_t$ at frame t, the inversion model estimates the corresponding static Maeda parameters $\hat{\boldsymbol{x}}_t = \boldsymbol{w}_0 + \sum_{i=1}^{N} \boldsymbol{w}_i\, \phi(\|\boldsymbol{y}_t - \boldsymbol{c}_i\|)$, where $\boldsymbol{w}_i, i = 0,1,2..N$ are weight vectors, N is the number of hidden nodes in the RBF network, and $\boldsymbol{c}_i$, $i = 1,2 \ldots N$ are the centroids of the Gaussian basis functions. The basis function is defined as $\phi(r) = e^{-(r/\sigma)^2}$, where $r = \|\boldsymbol{y}_t - \boldsymbol{c}_i\|$ and σ is the spread parameter. The centroids $\boldsymbol{c}_i$ are obtained through k-means clustering of the acoustic feature vector $\boldsymbol{y}_t$ in a training set, whereas the weights $\boldsymbol{w}_i$ are learned using the pseudo-inverse method [19] with a regularization parameter, l. Parameters σ and l are selected through cross-validation while training the RBF network [5].

3.2 Forward mapping model

Following Toda et al. [6], we use a GMM-based forward mapping model coupled with global variance [20] to estimate the trajectory of spectral features from the trajectory of articulatory parameters. Assume $\boldsymbol{x}_t$ is an articulatory feature vector consisting of (*a*) static and dynamic (delta) Maeda parameters, (*b*) energy, and (*c*) log pitch at frame t. Let $\boldsymbol{y}_t$ be the target spectral feature vector ($MFCC_{1\text{-}24}$) of dimension D. Then, the distribution of the joint vector $\boldsymbol{z}_t = [\boldsymbol{x}_t^{\top}, \boldsymbol{y}_t^{\top}]^{\top}$ is modeled as

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} \alpha_m\, \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \quad (1)$$

where α_m is the weight of the m^{th} mixture component and $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ is its normal distribution with mean $\boldsymbol{\mu}_m^{(z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$. We will use the symbol $\boldsymbol{\lambda}^{(z)}$ to denote the parameter set for the GMM model, which consists of weights, mean, and covariance matrices of all individual mixture components. All model parameters are learned from the training set of joint vectors $\boldsymbol{z}_t$ using expectation-maximization (EM).

Given a GMM model, we calculate the maximum likelihood estimate of spectral features considering the dynamics and the global variance (GV) as follows. Let column vector $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \boldsymbol{Y}_3^\top, \ldots . \boldsymbol{Y}_L^\top]^\top$ denote the sequence of static and dynamic spectral features from all L frames in a sentence, where $\boldsymbol{Y}_t = [\boldsymbol{y}_t{}^\top, \Delta \boldsymbol{y}_t{}^\top]^\top$ is the target spectral feature column vector composed of static spectral features $\boldsymbol{y}_t$ and the corresponding dynamic features $\Delta \boldsymbol{y}_t$ at frame t. Similarly, let column vector $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \boldsymbol{X}_3^\top, \ldots . \boldsymbol{X}_L^\top]^\top$ denote the sequence of articulatory feature vectors of the same L frames where $\boldsymbol{X}_t = [\boldsymbol{x}_t{}^\top, \Delta \boldsymbol{x}_t{}^\top]^\top$. Also, consider the within-sentence variance of the d^{th} component of spectral features given by $v(d) = \frac{1}{L}\sum_{t=1}^{L}(y_t(d) - \bar{y}(d))^2$, where, $\bar{y}(d) = \frac{1}{L}\sum_{t=1}^{L} y_t(d)$ and $y_t(d)$ is the d^{th} component of static spectral feature vector at time frame t. Thus, the GV of the static spectral feature is written as $\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), v(3), \ldots v(d), \ldots v(D)]$ where D is the dimension of static spectral feature vector $\boldsymbol{y}_t{}^\top$, and $\boldsymbol{y}$ is the sequence of static spectral features $[\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \boldsymbol{y}_3^\top, \ldots . \boldsymbol{y}_L^\top]^\top$. Now, the time sequence of estimated spectral feature vectors (static only) is given by the following equation:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\ P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}) \tag{2}$$

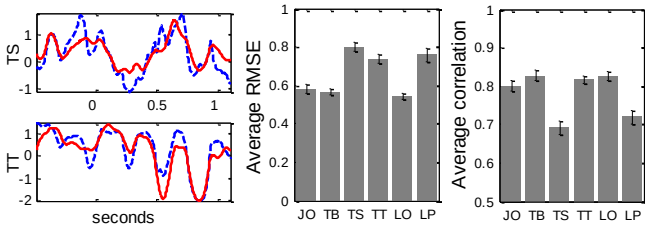
where $\boldsymbol{\lambda}^{(v)} = \{\boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\}$, $\boldsymbol{\mu}^{(v)}$ is the vector of average variance for all spectral features and $\boldsymbol{\Sigma}^{(vv)}$ is the corresponding covariance matrix, learned from the distribution of $\boldsymbol{v}(\boldsymbol{y})$ in the training set. The likelihood $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)})$ is computed as

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)}) = P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(v)})^{w} \cdot P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}). \tag{3}$$

The distribution of GV, $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$, is modeled by a single Gaussian $\mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)})$. The power term $w\ (= 1/2L)$ in equation (3) controls the balance between the two likelihoods. Following [20], we use EM to solve for $\hat{\boldsymbol{y}}$ in equation (2).

4. EXPERIMENTAL RESULTS

We evaluated our inversion/synthesis methods on an articulatory/acoustic dataset of 640 sentences uttered by a single speaker (*rgo)* described in [3]. Following Bawab et al. [21], we estimated six Maeda parameters (jaw opening, tongue back position, tongue shape, tongue tip height, lip opening and lip protrusion) from drift-corrected EMA (Electromagnetic Articulography) positions. The seventh Maeda parameter (larynx height) cannot be calculated from EMA. We then normalized the Maeda parameters to zero mean and unit variance. We computed 25 MFCCs ($MFCC_{0\text{-}24}$), from the STRAIGHT spectrum, and used $MFCC_0$ as the energy parameter and $MFCC_{1\text{-}24}$ as spectral features. *Pitch* and *aperiodicity* were also extracted from STRAIGHT analysis, and later used in waveform generation. Maeda and MFCCs were obtained for synchronous time steps sampled at 200Hz. Out of the 640 sentences, we randomly selected 100 sentences as a test set, and used the remaining 540 sentences to train the inversion and forward mappings.



(a) (b) (c)

Figure 2: (a) Trajectories of actual (dotted blue) and inverted (solid red) Maeda parameters of a typical sentence. (b) Accuracy of inverted Maeda features (JO=jaw opening, TB=tongue body position, TS=tongue shape, TT=tongue tip, LO=lip opening, LP=Lip protrusion). (c) Correlation coefficient between measured and inverted Maeda parameters.

4.1 Accuracy of the inversion model

Following Chao and Carreira-Perpiñán [5], we trained an RBF network with 25 input nodes (25 MFCCs), 1024 hidden nodes, and 6 output nodes using all non-silent frames from the training sentences. We then predicted Maeda parameters for all the non-silence frames in the test sentences. Figure 2(a) shows the trajectories of inverted and actual Maeda parameters (tongue shape and tongue tip) on a sample utterance, whereas Figure 2(b-c) shows the average RMSE (root mean squared error) and average CC (correlation coefficient) and between inverted and measured parameters. Predictions of tongue shape had the least accuracy (RMSE=0.80, CC=0.69), followed by lip protrusion (RMSE=0.76, CC=0.71) and tongue tip (RMSE=0.73, CC=0.81). Lip opening (RMSE =0.54, CC=0.83) had the highest accuracy among the six Maeda parameters.

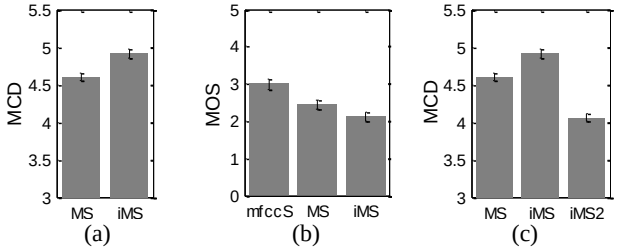
4.2 Synthesis quality with measured and predicted articulators

After establishing the accuracy of the inversion model, we designed an experiment to compare the quality of synthesis driven by either measured or predicted Maeda parameters. For this purpose, we trained a GMM-based forward mapping model with 256 mixture components using measured Maeda parameters. Then, we generated two sets of syntheses. The first set (*MaedaSynth*) was obtained by driving the GMM with measured Maeda features; the second set (*iMaedaSynth*) was obtained by driving the GMM with predicted Maeda parameters. We then calculated the average MCD (Mel cepstral distortion) between non-silent frames in the original and synthesized utterances as

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(MFCC_d^{(o)} - MFCC_d^{(s)}\right)^2}, \tag{4}$$

where $MFCC_d^{(o)}$ and $MFCC_d^{(s)}$ are the d^{th} MFCC coefficient of the original and synthesized sentences, respectively. As shown in Figure 3(a), the quality of *iMaedaSynth* (MCD=4.92dB) was 7% worse than that of *MaedaSynth* (MCD =4.60dB).

We also conducted a listening test to evaluate the subjective quality of *MaedaSynth* and *iMaedaSynth*. Given that both synthesis methods involved lossy compression of STRAIGHT spectra into MFCCs, we also evaluated a third set of synthesis (*mfccSynth*),



(a) (b) (c)

Figure 3: (a) Objective and (b) subjective evaluation of articulatory synthesis driven by measured and predicted Maeda parameters; MaedaSynth (MS): driven by measured Maeda parameters; iMaedaSynth (iMS): driven by iMaeda; mfccSynth (mfccS): synthesis following MFCC compression. (c) Objective comparison of synthesis quality of iMaedaSynth2 (iMS2) with MaedaSynth (MS) and iMaedaSynth (iMS).

which consisted of mapping STRAIGHT spectra into MFCCs and back to STRAIGHT spectra. For the perceptual tests, listeners[a] (n=30) were asked to rate 30 sentences (10 sentences for each of the three synthesis methods, presented randomly without repetition) using Mean Opinion Score (MOS: 1=Bad; 2=Poor; 3=Fair; 4=Good; 5=Excellent). Before the test began, participants were calibrated by listening to sample sounds with accepted MOS values. Results are shown in Figure 3(b). The baseline method (*mfccSynth*) was rated highest with an average MOS of 3.0. *MaedaSynth* and *iMaedaSynth* received average MOS of 2.4 and 2.1 respectively, a result that is consistent with the objective evaluation in Figure 3(a). In conclusion, both objective and subjective evaluations indicate that the articulatory-inversion model degraded the quality of synthesized speech. Does this result mean that articulatory inversion cannot be used to enable speech modification in the articulatory domain? Not quite, as we show next.

4.3 Retraining the forward mapping model with predicted articulators

Could the loss of quality of *iMaedaSynth* have been caused by training the articulatory inversion and forward mapping separately? To answer this question, we decided to retrain the GMM forward mapping using iMaeda parameters (from training sentences) instead of the measured Maeda parameters. Then, we generated a new set of syntheses (*iMaedaSynth2*) for the same test sentences in *iMaedaSynth*. To evaluate the synthesis quality, we computed MCD of *iMaedaSynth2* and compared against the MCD for *MaedaSynth* and *iMaedaSynth*. Results are shown in Figure 3(c): the *iMaedaSynth2* model (mean MCD: 4.06) not only outperforms the *iMaedaSynth* model (mean MCD: 4.92) but also the *MaedaSynth* model (mean MCD: 4.60).

To confirm these results, we conducted pairwise listening tests to compare the subjective quality of *iMaedaSynth* and *iMaeadaSynth2*. Participants (n=10) were asked to listen to parallel syntheses of the same sentence (one from *iMaedaSynth*, the other from *iMaedaSynth2*) and then select the one they perceived to be of better quality. Each participant listened to 60 such pairs (30 pairs of sentences presented twice in reversed order to avoid ordering effects). On average, the *iMaedaSynth2* was preferred 68% of the time over *iMaedaSynth* (95% confidence interval $\pm$5.38%). In a final perceptual study we compared *iMaedaSynth2* against *MaedaSynth*. On average, *iMaedaSynth2* was preferred 57% of the time over *MaedaSynth* (95% confidence interval $\pm$3.85%). Thus, these results indicate that training the articulatory synthesizer with predicted articulators not only reverses any errors introduced by the articulatory inversion model but also provides higher synthesis quality than what could be achieved if ground-truth articulators were available.

[a] Participants were recruited through Amazon Mechanical Turk. Only residents in the US were allowed to participate in the study.

5. DISCUSSION

The objective of this study was to evaluate the use of inverted articulatory parameters in data-driven articulatory speech synthesis. Our initial results show that replacing measured articulators with predicted articulators reduces the quality of a GMM-based synthesizer [6], as measured by Mel cepstral distortion and Mean Opinion Scores. However, the apparent loss of synthesis quality can be avoided by retraining the GMM on predicted Maeda parameters rather than on measured Maeda parameters. More importantly, driving the retrained synthesizer with inverted articulators (*iMaedaSynth2*) generates speech of higher quality than the original synthesizer driven by ground-truth articulators (*MaedaSynth*), as measured by Mel cepstral distortion and Mean Opinion Scores. Thus, it appears that the inversion step facilitates the synthesis process by eliminating variance in the articulators that is not predictive of (predicted by) acoustic information. These results suggest that inverted articulatory features can be used in speech synthesis without compromising synthesis quality, and open up the possibility of speech modification in the articulatory domain through articulatory inversion.

Our inversion results indicate that predictions of tongue-related parameters and lip protrusion are the least accurate. These results are consistent with previous studies [5, 6] and can be attributed to the higher degree of freedom of the tongue compared to jaw and lips. The poor subjective quality of *MaedaSynth* (MOS: 2.4) also deserves further discussion. At first, this result may suggest that there are issues with our articulatory synthesizer. However, the baseline synthesis method (*mfccSynth*), which sounds very similar to the original recordings and comparable to those rated as 4.1 MOS in [20], also received a low rating (MOS: 3.0). In previous work [22], we reported that recordings from a non-native speaker in the CMU-ARCTIC corpus received significantly lower MOS than those from a native speaker in that same corpus. Thus, the low ratings in our study can be attributed to the characteristics of the original recordings (i.e., utterances from a non-native speaker, EMA interfering with speech production).

The results presented here were obtained using a speaker-dependent articulatory inversion model. Further work is needed to test whether similar results can be achieved with subject-independent inversion models. Though the task appears challenging, Ghosh and Narayanan [11] have recently shown that articulatory features predicted from speaker-independent models can boost automatic speech recognition.

6. ACKNOWLEDGEMENTS

This work is supported by NSF award 0713205. We are grateful to Prof. Steve Renals and the Scottish Informatics and Computer Science Alliance (SICSA) for their support during RGO's sabbatical stay at CSTR (University of Edinburgh).

7. REFERENCES

[1] H. Hermansky and D. Broad, "The effective second formant F2'and the vocal tract front-cavity," in *ICASSP*, 1989, pp. 480-483.

[2] A. Toth and A. Black, "Using articulatory position data in voice transformation," *ISCA SSW6*, pp. 182-187, 2007.

[3] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign accent conversion through concatenative synthesis in the articulatory domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2301-2312, 2012.

[4] Z. H. Ling, K. Richmond, J. Yamagishi, and R. H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1171-1185, 2009.

[5] Q. Chao and M. A. Carreira-Perpiñán, "The geometry of the articulatory region that produces a speech sound," in *Asilomar Conference on Signals, Systems & Computers*, 2009, pp. 1742-1746.

[6] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, pp. 215-227, 2008.

[7] I. Y. Ozbek, M. Hasegawa-Johnson, and M. Demirekler, "Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) with Audio-Visual Information Fusion and Dynamic Kalman Smoothing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1180-1195, 2011.

[8] K. Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Proc. Interspeech*, 2009, pp. 2835-2838.

[9] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303-319, 2002.

[10] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "Robust word recognition using articulatory trajectories and Gestures," in *Interspeech*, 2010, pp. 2038-2041.

[11] P. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, pp. EL251-EL257, 2011.

[12] S. Maeda, "An articulatory model of the tongue based on a statistical analysis," *The Journal of the Acoustical Society of America*, vol. 65, p. S22, 1979.

[13] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, pp. 199-229, 1982.

[14] I. S. Howard and M. A. Huckvale, "Training a vocal tract synthesizer to imitate speech using distal supervised learning," in *Proc. SPECOM*, 2005, pp. 159-162.

[15] G. Bailly, "Learning to speak. Sensori-motor control of speech movements," *Speech Communication*, vol. 22, pp. 251-267, 1997.

[16] A. W. Black, H. T. Bunnell, Y. Dou, P. Kumar Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn, "Articulatory features for expressive speech synthesis," in *ICASSP*, 2012, pp. 4005-4008.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294-299.

[18] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *ICASSP*, 1997, pp. 1303-1306.

[19] S. Haykin and L. Li, "Nonlinear adaptive prediction of nonstationary signals," *IEEE Transactions on Signal Processing*, vol. 43, pp. 526-535, 1995.

[20] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222-2235, 2007.

[21] Z. Al Bawab, R. Bhiksha, and R. M. Stern, "Analysis-by-synthesis features for speech recognition," in *ICASSP*, 2008, pp. 4185-4188.

[22] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, pp. 920-932, 2009.