Palate-referenced Articulatory Features for Acoustic-to-Articulator Inversion

An Ji [1], *Michael T. Johnson* [1], *Jeffrey Berry* [2]

[1] Electrical and Computer Engineering, Marquette University, Milwaukee, WI, USA
[2] Speech Pathology and Audiology, Marquette University, Milwaukee, WI, USA

{an.ji, mike.johnson, jeffrey.berry}@marquette.edu

Abstract

The selection of effective articulatory features is an important component of tasks such as acoustic-to-articulator inversion and articulatory synthesis. Although it is common to use direct articulatory sensor measurements as feature variables, this approach fails to incorporate important physiological information such as palate height and shape and thus is not as representative of vocal tract cross section as desired. We introduce a set of articulator feature variables that are palate referenced and normalized with respect to the articulatory working space in order to improve the quality of the vocal tract representation. These features include normalized horizontal positions plus the normalized palatal height of two midsagittal and one lateral tongue sensor, as well as normalized lip separation and lip protrusion. The quality of the feature representation is evaluated subjectively by comparing the variances and vowel separation in the working space and quantitatively through measurement of acoustic-to-articulator inversion error. Results indicate that the palate-referenced features have reduced variance and increased separation between vowels spaces and substantially lower inversion error than direct sensor measures.

Index Terms: articulatory features, Electromagnetic Articulography (EMA), acoustic-to-articulatory inversion

1. Introduction

The use of articulatory information can be beneficial for a wide variety of speech tasks. Such information can improve the performance of automatic speech recognition (ASR) systems by accounting for speech production knowledge [1-3], or increase the quality of synthesis algorithms [4-6]. For tasks such as pronunciation assessment accurate articulatory information can help provide more detailed user feedback [7]. The goal of articulatory inversion is to recover articulatory trajectories from the acoustic signal for such purposes. One challenge for this task is the selection of appropriate articulatory feature variables that represent articulator movements and vocal tract structure in a meaningful way across differing speakers and physiologies Often kinematic sensor positions are used directly for inversion methods, but this is not always the best representation. In this paper we investigate articulatory feature representations that use palate-referenced measures of vocal tract area rather than direct sensor positions and evaluate their performance.

The kinematic data used in this work is Electromagnetic Articulography (EMA) sensor measurements. The EMA modality is currently a popular approach to tracking articulatory motion due to its relative low cost and balance between spatial and temporal resolution. The measured trajectory consists of a set of position coordinates for each sensor during speech. Toda [8] has shown that speech spectra can be produced from EMA measurements by learning statistical dependencies between position trajectory and the corresponding speech signal, indicating that raw EMA measures relate to the output of the speech generation process. While most work in this area has used sensor position data directly [9-12], this has significant weaknesses in terms of usefulness for acoustic-to-articulatory inversion and applications. One key weakness is that differences in subject physiology and sensor placement cause absolute sensor positions to vary significantly across speakers. Even more significant, however, is the fact that raw sensor data does not include all relevant vocal tract information, most notably information about the location of the palate relative to the sensor. This information is typically available, since palate data is typically collected along with other baseline calibration measures as part of the EMA data collection process.

This paper proposes the use of palate-referenced articulatory measures as reliable and phonetically meaningful features to characterize vocal tract shapes from EMA measurements. These features are evaluated through comparisons of the feature space across different vowels any by comparing acoustic-to-articulator inversion accuracy.

2. Method

2.1. Maeda's vocal tract model

A model-based approach is used to estimate the vocal tract configuration and identify appropriate features. Several theoretical models that describe speech production process have been proposed [13-15]. In this work the Maeda model mid-sagittal vocal tract representation is used, as shown in Figure 1. The Maeda model represents the articulatory working space with seven key parameters that relate to the cross-sectional areas of the vocal tract, determined from a factor analysis of x-ray vocal tract contour data [16].

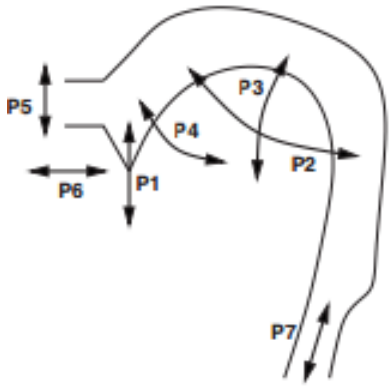

Figure 1: *Maeda's articulatory model: P1 jaw height, P2 tongue dorsum length, P3 tongue dorsum shape, P4 tongue apex position, P5 lip separation, P6 lip protrusion, P7 larynx height.*

2.2. Dataset

The Marquette EMA-MAE corpus [17] includes synchronous acoustic and three-dimensional kinematic data collected at 400 Hz. Acoustic records were obtained using a cardioid pattern directional condenser microphone positioned approximately 1 meter from participants. The corpus includes approximately 45 minutes of synchronized acoustic and kinematic data for each speaker, including word, sentence, and paragraph level speech samples.

As shown in Figure 2, articulatory sensors included the jaw (MI) (interior lower front incisor), lower lip (LL), upper lip (UL), tongue dorsum (TD), and tongue tip (TT), all placed in the midsaggital plane. In addition, there were two lateral sensors, one (LC) at the right corner of the mouth to help indicate lip rounding and one (LT) in the right central midpoint of the tongue body to help indicate lateral tongue curvature.

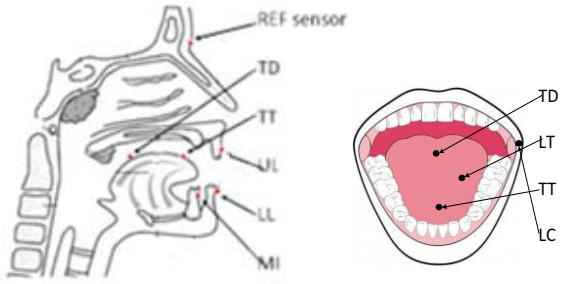

Figure 2: *EMA-MAE sensor placement*

2.3. Proposed articulatory feature

We have developed a geometric transformation from the EMA kinematic measurements to vocal tract (VT) parameters based on an extension of the Maeda model. These parameters include the following articulatory feature variables:

Table 1. *Articulatory features*

	Description
VT1	Tongue dorsum normalized horizontal position
VT2	Tongue dorsum vertical height to hard palate
VT3	Tongue body normalized horizontal position
VT4	Tongue body vertical height to hard palate
VT5	Tongue apex normalized horizontal position
VT6	Tongue apex vertical height to hard palate
VT7	Normalized horizontal lip protrusion
VT8	Normalized vertical lip separation

We measured the distance between the center incisors and the middle point of the back molar from each speaker's bite plate record. This distance is used as a normalization scalar when calculate horizontal position of the tongue, to give a better information of tongue's position relative to the whole vocal tract regardless of the difference among individuals. The horizontal (x-axis) variables VT1, 3, 5, and 7, are all calculated directly from sensor position divided by this normalization constant. This will lead to improvement in cross-subject variability but not variability or inversion accuracy within a single subject. The vertical (y-axis) variables VT2, 4, and 6; however, are computed as the vertical distance between the sensor position and the palate, representing vocal tract height at the sensor positions, including two midsagittal positions and one lateral position. It is hypothesized that these vertical articulatory variables will be significantly more representative of vocal tract height and therefore of acoustic spectral characteristics both within and across subjects. Lip protrusion VT7 is taken directly from the sensor x position without any normalization, and vertical lip separation VT8 is calculated as

$$VT8 = \frac{(UL_y - LL_y) - (UL_y - LL_y)_{closed\ position}}{(UL_y - LL_y)_{max}} \quad (1)$$

representing lip separation rescaled to a [0,1] working space.

2.3.1. Bite-plate correction

All sensor measures are referenced to an origin at the upper front incisor, with the data orientation referenced to the subject's head orientation. In addition to this baseline physical normalization, we use a bite-plate with each subject to determine the orientation of the maxillary occlusal plane where the upper and lower mandible meet, and apply a bite-plate correction to all sensor data so that the planes of the working space are the midsagittal plane (x-y plane) and the maxillary occlusal plane (x-z plane) [18].

2.3.2. Palate mesh estimation

The EMA-MAE palatal reference data includes a trace of the mid-sagittal palate line and a series of transverse traces across the palate. We use the thin plate spine (TPS) method [19] for estimation of the palate mesh from the collected palatal reference data. This approach mathematically warps a flat thin metal plate to approximate a set of control points. The TPS method allows control of the smoothness of the resulting mesh via a smoothness parameter λ , which in this work we have set equal to 0.05 [20].

2.4. Acoustic-to-articulator inversion

In this experiment, we implemented an HMM based inversion to estimate the articulatory parameters from the acoustic signal. The core inversion approach is similar to that in [21], with two parallel streams trained separately in acoustic and articulatory feature space. Three state left-to-right mono-phone HMMs with one Gaussian per state are used for training and testing. Twelve Mel-Frequency Cepstral Coefficients (MFCCs) plus energy, along with their first and second derivatives are used for the acoustic features. The 198 sentences from one subject are divided into training (178) and testing (20) sets. Two sets of articulatory feature vectors are implemented, the first being the direct x and y position values of the designated EMA sensors, and the second being the proposed articulatory features of Table 1, along with their first and second derivatives.

2.5. Evaluation

Several evaluation metrics are used to compare the baseline and proposed features sets. The first is simply the variance of the features, overall and within specific vowel configurations, with an emphasis on the variance in the vertical direction where the palate referencing has significant impact on the feature information. The second is the normalized RMS error or the acoustic-to-articulator inversion output, computed as

$$E_{rms} = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2}}{std(y)} \quad (2)$$

where m is the number of examples in the test set, y_i is the true articulatory variable value, $f(x_i)$ is the inversion output, and $std(y)$ is the standard deviation of the articulatory variable across the full test set.

The third metric is the correlation between the actual articulator motion y_i and the estimated motion $f(x_i)$.

3. Experiments and results

3.1. Working space analysis

Figure 3 compares the working spaces for the vowel [iy] for a female native English speaking subject. Focusing on the y-dimension, it can be seen that the overall working space is smaller and more compressed in the proposed palate-referenced feature space.

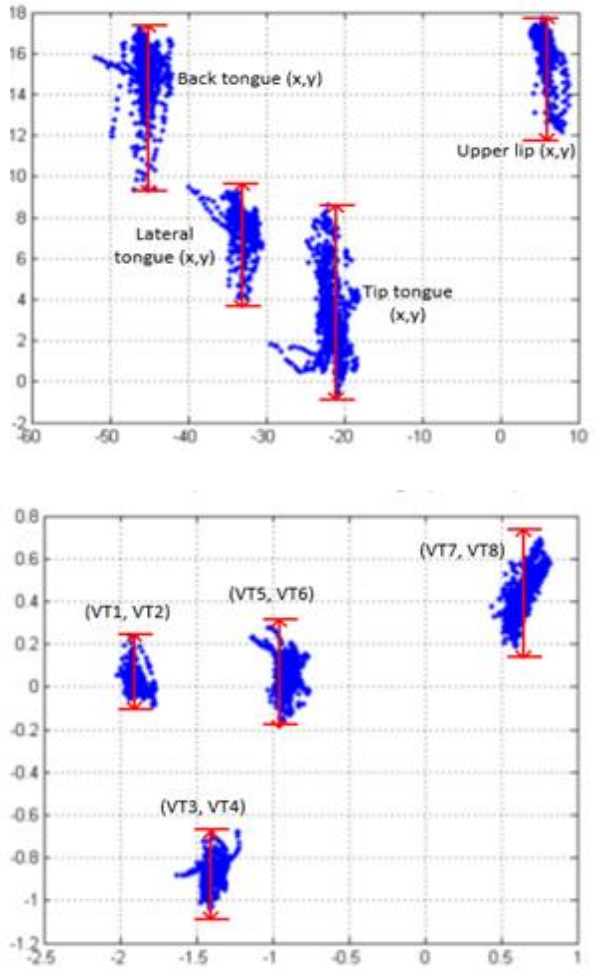

Figure 3: *Working space for vowel [iy] for direct sensor measures (upper) and proposed articulatory features (lower).*

Table 2. *Average variance in vertical sensor dimension within baseline and proposed feature spaces.*

Vowel	Average variance, position features	Average variance, articulatory features
[ey]	±8.5%	±4%
[uw]	±8%	±5.5%
[aa]	±12%	±6.5%

Figure 4 compares the working spaces for three different vowels. It can be seen that the overlap between the vowel spaces is significantly reduced using the proposed articulatory variables.

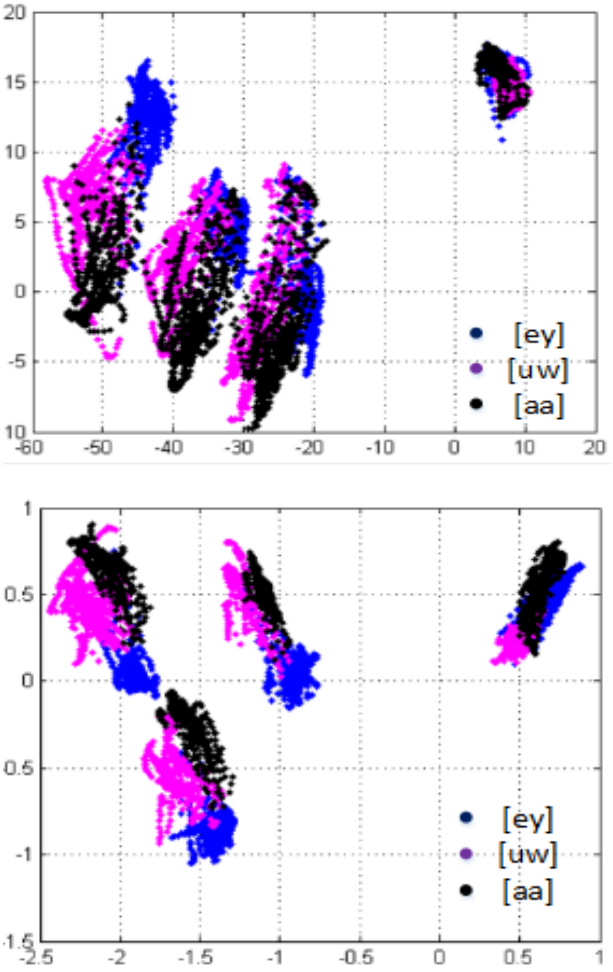

Figure 4: *Working space comparing variance and overlap between [ey], [uw], and [aa] for direct sensor measures (upper) and proposed articulatory features (lower)*

3.2. Acoustic-articulator inversion accuracy

Figure 5 illustrates the measured and reconstructed time trajectories of raw sensor coordinates and articulatory feature in vertical dimension for a test utterance.

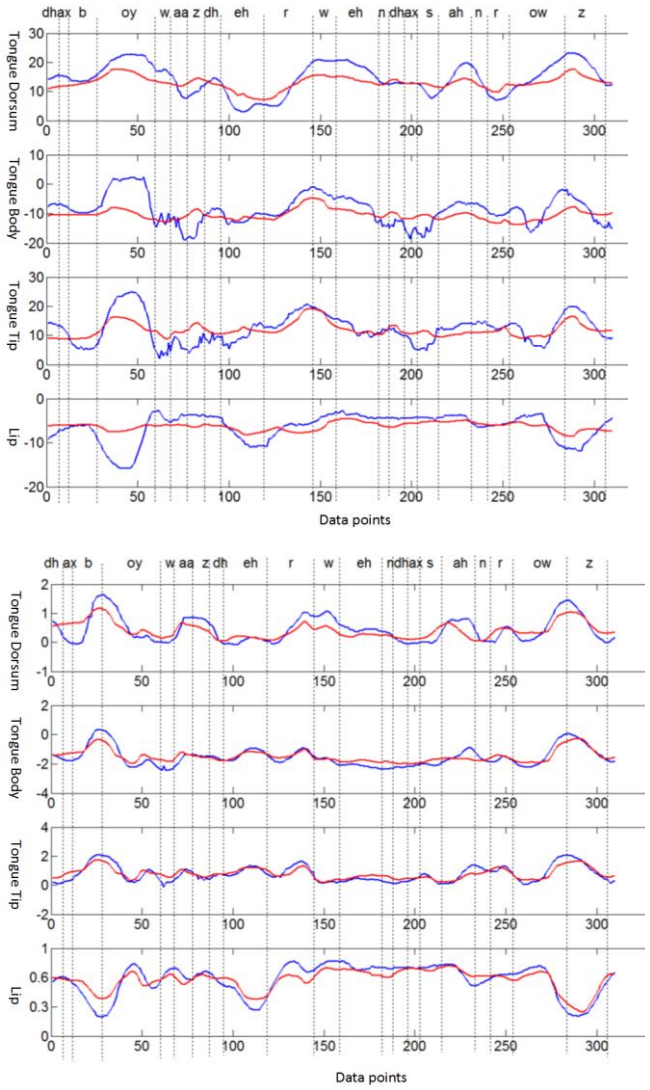

Figure 5: *Measured (blue lines) and reconstructed (red lines) trajectories of the direct measures (upper) and articulatory features (lower), in the test sentence "The boy was there when the sun rose". Phone boundaries are shown by vertical bars.*

Table 3. *RMS error and correlation coefficients between acoustic-to-articulator inversion estimates and actual trajectories.*

	E_{rms}		r	
	Sensor space	AF space	Sensor space	AF space
Dorsum	0.975	0.644	0.658	0.718
Body	1.052	0.715	0.620	0.729
Tip	0.862	0.617	0.603	0.786
Lips	0.887	0.678	0.592	0.736

Results indicate that the normalized EMS error is smaller and the correlation coefficient is higher for articulatory features compared to raw movement data under the same inversion system, suggesting that the proposed palate-referenced features are better choices for representing the vocal tract configuration.

4. Discussion

The results in Figure 3 and Table 2 show that the vertical variance is significantly reduced in the palate-reference feature space, and Figure 4 clearly shows that the proposed features have significantly less overlap between the working space, strongly suggesting that the new features have better discriminatory representations than direct kinematic data. This directly influences the performance of HMM based acoustic-to-articulatory inversion due to increased separation between the observation distributions of different models, as shown by the decreased inversion error and increased correlation to actual feature trajectories. From the inversion results Table 3, the average decrease in RMS error for the vertical dimension is 29% and the increase in correlation is 20%. This improvement implies that the proposed feature is more capable in characterizing vocal tract shapes than the direct measure.

5. Conclusion

This paper introduces a set of palate-referenced articulatory features to characterize vocal tract shapes from EMA measurements, and compares the working space and acoustic-to-articulator inversion accuracy of these new features to that of direct sensor data. Analysis show that the variance of the palate-referenced features is reduced, and even more importantly that the overlap of the vowel spaces characterized by these features is also significantly reduced. The resulting acoustic-to-articulator error is decreased by 29%, while correlation between the estimated and actual feature trajectories increased by 20%. Overall, these results strongly support the hypothesis that palate-referenced articulatory features are significantly more representative of vocal tract structure and acoustic spectral characteristics than direct sensor measures.

6. Acknowledgements

This paper is based upon work supported by the National Science Foundation under Grant No. IIS-1142826 and IIS-1320892.

7. References

[1] Kirchhoff, K., "Robust Speech Recognition Using Articulatory Information", PhD Thesis, University of Bielefeld, 1999.

[2] Frankel, J. and King, S., "ASR-Articulatory Speech Recognition", Proc. of Eurospeech, pp. 599-602, Denmark, 2001.

[3] Deng, L. and Sun, D., "A Statistical Approach to Automatic Speech Recognition Using Atomic Units Constructed from Overlapping Articulaory Features", Journal of Acoustic Society America. 95(5):2702-2719, 1994.

[4] Ling, Z. H., Richmond, K., Yamagishi, J. and Wang, R. H., "Articulatory Control of HMM-bashed Parametric Speech Synthesis Driven by Phonetic Knowledge", Proc. Interspeech, pp. 573-576, Australia, 2008.

[5] Ling, Z. H., Richmond, K., Yamagishi, J. and Wang, R. H., "Integrating Articulatory Features into HMM-based Parametric Speech Synthesis", IEEE Trans. Audio, Speech, Lang. Process, 17(6):1171-1185, 2009.

[6] Ling, Z. H., Richmond, K., Yamagishi, J., “Articulatory Control of HMM-based Parametric Speech Synthesis Using Feature-Space-Switched multiple Regression”, IEEE Trans. Audio, Speech, and Language Processing, 21(1):207-219, 2013.
[7] Levis, J., “Computer Technology in Teaching and Researching Pronunciation”, Annual Review of Applied Linguistics, 27:184-202, 2008.
[8] Toda, T., Black, A. and Tokuda, K., “Acoustic-articulatory Inversion Mapping with Gaussian Mixture Model”, Proc. ICSLP, pp. 1129-1132, Jeju Island, Korea, 2004.
[9] Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P. and Saltzman, E., “Accurate Recovery of Articulatory Position from Acoustics: New Conclusions Based on Human Data”, Journal of the Acoustical Society of America, 100(3):1819-1834, 1996.
[10] Richmond, K., “Estimating Articulatory Parameters from the Acoustic Speech Signal”, PhD thesis, University of Edinburgh, 2001.
[11] Hiroya, S. and Honda, M., “Estimation of Articulatory Movements from Speech Acoustics Using an HMM-based Speech Production Model”, IEEE Trans. Speech Audio Process, 12(2):175-185, 2004.
[12] Zhang, L. and Renals, S., “Acoustic-articulatory Modelling with the Trajectory HMM”, IEEE Signal Processing Letters, 15:245-248, 2008.
[13] Mermelstein, P., “Articulatory model for the study of speech production”, Journal of the Acoustical Society of America, 53(4):1070-1082. 1973.
[14] Coker, C. H., “A Model for Articulatory Dynamics and Control”, Proceedings of the IEEE, 64(4):452-260, 1976.
[15] Birkholz, P., Jackel, D. and Kröger, B. J., “Construction and Control of a Three-dimensional Vocal Tract Model”, ICASSP, pp. 873-876, France, 2006.
[16] Meada, S., “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model”, Speech Production and Speech Modeling, pp. 131–149, 1990.
[17] Ji, A., Johnson, T. M. and Berry, J., “The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) Corpus of Acoustic and 3D Articulatory Kinematic Data”, ICASSP, Italy, 2014.
[18] Ji, A., Johnson, M. T. and Berry, J., "Tracking articulator movements using orientation measurements", International Conference on Audio, Language, and Image Processing, pp, 292-296, Shanghai, China, 2012.
[19] Wahba, G., “Spline models for observational data”, Philadelphia: Society for Industrial and Applied Mathematics, 1990.
[20] Yunusova, Y., Baljko, M., Pintilie, G., Rudy, K., Faloutsos, P. and Daskalogiannakis, J., “Acquisition of the 3D Surface of the Palate by in-vivo Digitization”, Speech Communication, 54:923-931, 2012.
[21] Youssef, A. B., Badin, P., Bailly, G., and Heracleous, P., "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models", Proc. Interspeech, Brighton, UK, 2009.