COMPARING THE PERFORMANCE OF INDIVIDUAL ARTICULATORY FLESH POINTS FOR ARTICULATION-TO-SPEECH SYNTHESIS

Beiming Cao[1], Brian Y. Tsang[1,3], Jun Wang[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]University of Texas at Austin, Austin, Texas, United States
{beiming.cao,wangjun}@utdallas.edu; briantsang2020@utexas.edu

ABSTRACT

Articulation-to-speech (ATS) synthesis has recently shown the potential for silent speech interfaces (SSIs). SSIs are devices for assisting the oral communication for individuals who have lost their voice by mapping their articulatory movement to audible speech. Electromagnetic Articulograph (EMA) is one of the current articulator motion tracking technologies in SSI, which captures the movement of flesh points on articulators. Understanding how well different individual flesh points contribute to ATS performance may help optimize the SSI setup. To our knowledge, this study is the first to explore the individual flesh point's contribution to ATS, where we compared ATS performance using EMA data of different flesh points combinations with a deep neural network (DNN)-based ATS model. Experimental results indicated that more flesh points lead to higher performance generally. However, our perception-based evaluation may suggest the unnecessity of more than one tongue (tip) flesh point for ATS.

Keywords: Articulation-to-speech, deep neural network, silent speech interface.

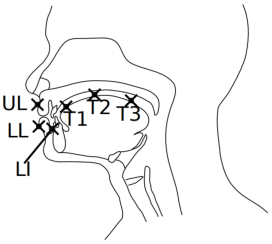
1. INTRODUCTION

Articulatory-to-speech (ATS) [1, 2, 3, 4] synthesis generates speech from articulatory information without textual information. Textual and linguistic analyses of input are not required in ATS, therefore ATS is suitable for a real-time silent speech interface (SSI). Silent speech interfaces (SSIs) are systems enabling speech communication when audible acoustic signal is unavailable [5], which have the potential of recovering speaker's own voice for people who are unable to produce speech sounds but can still articulate. A variety of sensing technologies have been used to capture articulatory movements including electromagnetic articulography (EMA) [6], ultrasound [7], and recently proposed permanent magnet articulography (PMA) [8, 2]. Especially, a new PMA-based device was proposed [9] which captures only tongue tip, and lips motion, and it has been validated for silent speech interface usage [9].

In order to optimize the performance of ATS and utilize articulatory data effectively, it is important to investigate the performance of individual articulatory flesh point for ATS. This study is beneficial to both speech science and technology. Scientifically, speech production is one of the most complex and rapid motor behaviors and involves a precise coordination of over 100 laryngeal, orofacial and respiratory muscles [10]. Exploring the performance of articulators in speech synthesis would help people understand the differences among articulators in speech production [11]. Technologically, this study will be helpful for applications like ATS and silent speech recognition (SSR). In addition, this study can be a reference during articulatory data collection and processing for ATS application. It could help people choose flesh point location during EMA data collection, and for image-based articulatory data like ultrasound and MRI, it could be a reference of flesh points tracking and image feature extraction.

Previous studies [12, 13] explore the performance of flesh points on articulators in silent speech recognition (SSR) [14] and determined an optimal set of flesh points (tongue tip, tongue back, upper lip and lower lip) for SSR. However, speech recognition essentially is a classification process with context information or language model applied. On the other hand, articulation-to-speech (ATS) synthesis based on statistical parameters speech synthesis (SPSS) [15] is a regression procedure which predicts numerical acoustic features from articulatory information. Compared to SSR, ATS is more sensitive to the articulatory input data. Therefore, the optimal sets of flesh points of SSR can only be used as a reference for ATS study rather than a conclusion.

In this study, we investigated the ATS perfor-



label	location	label	location
UL	Upper lip	T1	Tongue tip
LL	Lower lip	T2	Tongue body
LI	Lower incisor	T3	Tongue dorsum

Figure 1: Sensor locations of EMA data in mngu0 (The figure is adapted from [16]).

mances of six flesh points on articulators: tongue tip (TT), tongue body (TB), tongue dorsum (TD), upper lip (UL), lower lip (LL) and jaw. The synthesized speech utterances were evaluated objectively by the prediction accuracies of acoustic features. After that, a subjective testing was conducted to measure the speech intelligibility rate of the conventional TT, TB, UL and LL set and the new TT, UL and LL flesh point set proposed in [9]. A thorough discussion was made based on the results.

2. DATASET

The mngu0 dataset is a corpus of articulatory data of different forms acquired from one male British English speaker [16]. In this paper, we use the EMA subset [16] of mngu0 which consists of audio and EMA data of 1,354 sentences recorded by Carstens AG500 EMA [17]. The total length of the speech data is about 67 mins [16].

The raw EMA data of mngu0 dataset tracks 12 sensor coils in 3D space with two angles of rotation [16]. In this study, we use two-dimensional movement tracks of six sensors (Figure 1): upper lip (UL), lower lip (LL), lower incisor (JAW), tongue tip (TT), tongue body (TB), tongue dorsum (TD) extracted from raw EMA data. The 2D movement includes vertical and front-back directions. The movement of head was subtracted from these sensors' motion data to obtain head-independent articulator movement. The sampling rate of EMA data is 200Hz. The audio data was recorded synchronously with EMA data. The sampling rate of audio data is 16 kHz [16].

3. METHODS

3.1. Articulation-to-speech Synthesis Using Deep Neural Network

In this study, we adopted DNN to map acoustic features from articulatory data. The input of ATS includes sensor position vector in y (front-back) and z (up-down) directions; the outputs are acoustic features which are used for synthesizing speech by the WORLD voice encoder [18].

The input of ATS is 2-dimensional (front-back and up-down) motion vector of individual sensors attached to the flesh point on articulators (tongue, lips and jaw) or their different combinations. In addition, both input articulatory frames and output acoustic feature frames were concatenated with their first and second order of derivatives as the input and output of neural network models. The predicted acoustic features include: mel-cepstral coefficients (MCCs) [19], band aperiodicities (BAP)[20], logarithm of fundamental frequencies (log-F0) and voiced/unvoiced (V/UV) label. Accordingly, the objective evaluation of experimental results is the prediction accuracies of these features, which are mel-coefficient distortion (MCD), band aperiodicities (BAP) distortion, root mean square error of fundamental frequencies (F0-RMSE), and voiced/unvoiced (V/UV) error rate.

3.2. Experimental Setup

The mngu0 dataset provides 1,354 sentences of speech data with both audio and EMA data. The whole dataset was separated to training, development and testing set with 1,226 , 63 and 65 sentences respectively. The DNN used in this study has six hidden layers with 512 nodes. Learning rate was set to 0.003, training batch size and number of epoch were 128 and 25 respectively. We used stochastic gradient descent (SGD) optimizer for training. The experimental parameters are shown in table 1.

In this study, firstly we validated each of the six sensors in DNN-ATS experiment to see their individual performances. After that, they were combined in different sets to explore their performance when working together with others. For future convenience, we use expressions like [UL, LL, TT, TB, TD] to denote a flesh point set on the articulatory

Table 1: Experimental setup.

Acoustic Feature	187-dim. vectors
Mel-Cepstral Coefficients (MCCs)	(60-dim. vectors) + $\Delta + \Delta\Delta$ (180-dim.)
Band Aperiodicities (BAPs)	(1-dim. vectors) + $\Delta + \Delta\Delta$(3-dim.)
Fundamental Frequency on log scale (log-f0)	(1-dim. vectors) + $\Delta + \Delta\Delta$ (3-dim.)
Voiced/Unvoiced (V/UV) label	(1-dim.)
Sampling rate	16000 Hz
Windows length	25 ms
Articulatory Feature	36-dim. vectors
articulatory movement (6 sensors)	(12-dim. vectors) + $\Delta + \Delta\Delta$ (36-dim.)
Common	
Frame rate	5 ms

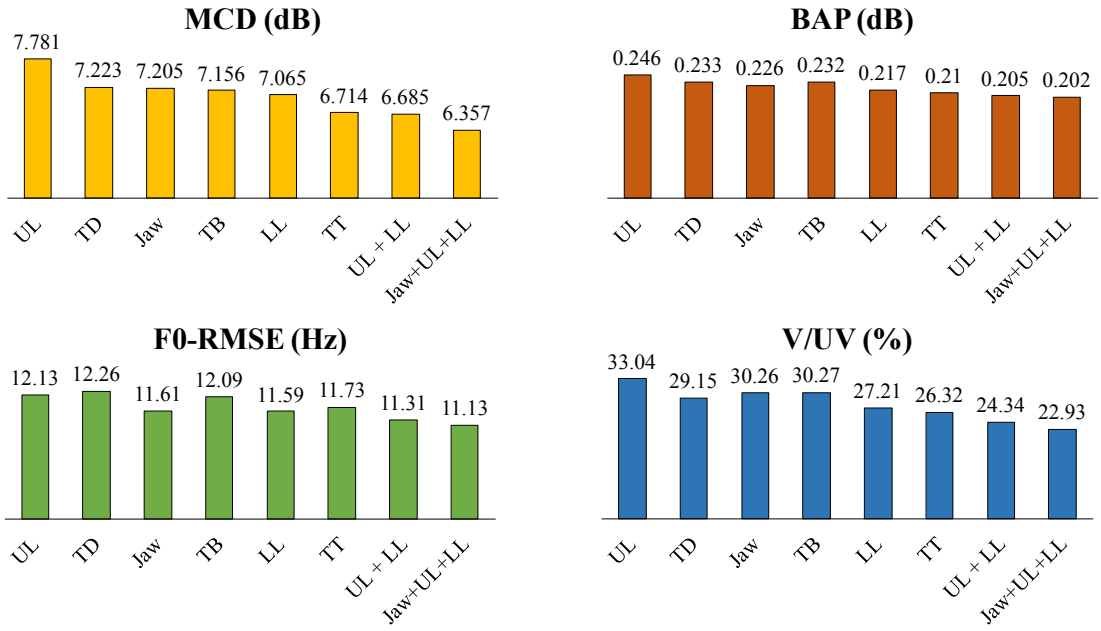


Figure 2: Results of Individual Flesh Points and Combination of Lips and Jaw.

flesh points in the square bracket.

Previous studies [12, 13] successfully proved that [UL, LL, TT, TD] consist an optimal flesh for silent speech recognition, which even outperformed [UL, LL, TT, TB, TD] [13]. Therefore, in this study we validated lips, jaw and lips + jaw combined with all combinations of TT, TB and TD in DNN-ATS.

4. RESULTS AND DISCUSSION

4.1. Individual Flesh Point in ATS

Figure 2 gives the performance of each individual flesh point and the combination of jaw and lips. Here the lower numbers indicate better performance. In Figure 2, for single flesh points, we can see that tongue tip (TT) outperform others in MCD and BAP predictions, LL is the second, whereas upper lip (UL) and tongue dorsum (TD) perform least helpful in all evaluations. Except for in F0-RMSE, lower lip (LL) outperforms other flesh points while jaw is the second. It is important to note that F0 and V/UV are less affected by articulation compared to MCD and BAP. Generally it can be concluded that the performances of six sensors in descending order are: tongue tip (TT), lower lip (LL), tongue body (TB), jaw, tongue dorsum (TD), and upper lip (UL).

4.2. Flesh Point Sets in ATS

The performance of lips, jaw, and lips + jaw combined with tongue sensors (TT, TB, TD) are shown from Table 2 to Table 5. In table 2 and table 3, MCD and BAP strictly follow the trends that [UL, LL, Jaw] $<$ lips $<$ Jaw, and [TT, TB, TD] $<$ [TT, TB] $<$ [TT, TD] $<$ [TB, TD]. In table 4 and 5, F0-RMSE and V/UV shows similar trends in MCD and BAP except for [UL, LL, Jaw, TT, TB] outperformed all six flesh points in both F0-RMSE and V/UV; in addition [Jaw, TT, TD] outperformed [Jaw, TT, TB] in F0-RMSE. Since F0 and V/UV have less relationship to articulation, it can be concluded that generally more articulatory flesh points' movement will generate better performance in ATS. However, from table 2 and 3, it also can be observed that all six points outperform [UL, LL, Jaw, TT, TB] slightly in MCD and BAP by 0.002 dB and 0.001 dB respectively whereas [UL, LL, Jaw, TT, TB] outperform all six points in both F0 and V/UV predictions.

This finding is different to the optimal flesh point set found in silent speech recognition(SSR): [UL, LL, TT, TD] which got rid of TB rather than TD [13]. As mentioned, SSR essentially is a classification process, whereas ATS in this study is frame by frame regression. Therefore, ATS is more sensitive to the articulatory data, and ATS would benefit more from the flesh points which affect speech production more. Based on our result of flesh points on tongue, the closer a flesh point to TT, the better it will perform. Therefore, TD is less important in ATS than in SSR whereas TB is more important in ATS. Another explanation of this is, based on our computation on all mngu0 EMA samples, the

Table 2: *MCD (dB) of Flesh Point Sets.*

	Jaw	Lips	Lips + Jaw
TT + TB + TD	5.644	5.422	5.377
TT + TB	5.758	5.489	5.379
TT + TD	5.826	5.556	5.439
TB + TD	5.933	5.715	5.533

Table 3: *BAP (dB) of Flesh Point Sets.*

	Jaw	Lips	Lips + Jaw
TT + TB + TD	0.182	0.178	0.178
TT + TB	0.186	0.180	0.179
TT + TD	0.186	0.180	0.179
TB + TD	0.192	0.188	0.183

Table 4: *F0-RMSE (Hz) of Flesh Point Sets.*

	Jaw	Lips	Lips + Jaw
TT + TB + TD	10.860	10.520	10.491
TT + TB	11.093	10.630	10.431
TT + TD	10.970	10.862	10.664
TB + TD	11.184	10.839	10.686

Table 5: *V/UV (%) of Flesh Point Sets.*

	Jaw	Lips	Lips + Jaw
TT + TB + TD	19.117	18.287	17.657
TT + TB	20.117	18.478	17.640
TT + TD	20.025	18.885	17.875
TB + TD	21.444	20.255	19.030

Euclidian distances between TT and TB, TB and TD are about 1.76 cm and 1.60 cm respectively. SSR studies [13] used the MOCHA-TIMIT database [21]. In MOCHA-TIMIT, TB was 2-3 cm from TT; TD was 2-3 cm from TB [22]. Therefore the distance between TD and TT in mngu0 is close to distance between TB and TT in MOCHA-TIMIT. Although there are physiological variations between the subjects, this could be considered as a possible explanation for the different performance of tongue flesh points find in SSR and ATS.

Another finding was that the jaw performed better when predicting F0 than predicting MCC and BAP. A possible explanation of this is that tongue body back movement may slightly affect vibration of vocal folds. Jaw movement reflects the movement of people opening and shutting their mouth, these movements squeeze and lose the tongue body back which may affect the vibration of vocal folds. This may explain why jaw performed better in predicting F0 than predicting MCD and BAP.

Table 6: *Word Accuracy (%) by Listeners.*

	Lips + TT	Lips + TT + TB
Listener 1	94.6	92.4
Listener 2	86.9	93.5
Listener 3	95.6	97.1
Listener 4	88.1	88.3
Average	91.3	92.8

4.3. Subjective Testing

A listening test was conducted by four native American English speakers to evaluate the speech intelligibility of ATS using flesh points combination of [TT, UL, LL] and [TT, TB, UL, LL] (Table 6), the former is used in the newly proposed articulatory movement capture device mentioned [9], and the latter is a popular flesh points set for silent speech recognition [12, 13, 14]. The average word accuracies of four listeners are 91.3% for [TT, UL, LL], and 92.8% for [TT, TB, UL, LL]. Therefore, given the slight advantage of using [TT, TB, UL, LL] flesh point and the technical difficulties of attaching one more sensor on tongue dorsum, [TT, UL, LL] might be overall more suitable for ATS application than [TT, TB, UL, LL].

5. CONCLUSIONS

In this study, we first compared individual flesh points for ATS. The results indicated that tongue tip and lower lip are the most helpful flesh points. When combining multiple flesh points, generally more flesh points generated better results. However, the performance of the flesh point set without tongue dorsum but with all other points is only slightly worse than using all six points in predicting MCC and BAP; at the same time it outperforms all six points in predicting F0 and V/UV. Given this result and the difficulties of attaching one more sensor on the tongue, the flesh point set: upper lip, lower lip, jaw, tongue tip, tongue body could be considered as a helpful sensor set in ATS. In addition, the flesh point set (tongue tip and lips) used by the newly proposed device has been compared subjectively to the conventional set. The results suggest that tongue tip and lips might be more suitable than multiple tongue (tip) flesh points for ATS.

6. ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (NIH) under award number R03DC013990 and by the American Speech-Language-Hearing Foundation through a New Century Scholar Research Grant.

7. REFERENCES

[1] P. Palo, "A Review of Articulatory Speech Synthesis," *Master's thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering*, 2006.

[2] J. Gonzalez Lopez, L. A. Cheah, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Evaluation of a Silent Speech Interface based on Magnetic Sensing and Deep Learning for a Phonetically Rich Vocabulary," in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, pp. 3986–3990, ISCA, 2017.

[3] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-based Ultrasound-to-Speech Conversion for a Silent Speech Interface," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3672–3676, 2017.

[4] B. Cao, M. Kim, J. R. Wang, J. Van Santen, T. Mau, and J. Wang, "Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018, pp. 3152–3156, 2018.

[5] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[6] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic Articulography: Use of Alternating Magnetic Fields for Tracking Movements of Multiple Points Inside and Outside the Vocal Tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[7] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips," *Speech Communication*, vol. 52, no. 4, pp. 288 – 300, 2010. Silent Speech Interfaces.

[8] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A Silent Speech System Based on Permanent Magnet Articulography and Direct Synthesis," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[9] M. Kim, N. Sebkhi, B. Cao, K. Okkelberg, M. Ghovanloo, and J. Wang, "Preliminary Test of a Wireless, Portable Magnetic Tongue Tracking System for Silent Speech Interface," *IEEE Biomedical Circuits and Systems Conference (BioCAS)*.

[10] K. Simonyan and B. Horwitz, "Laryngeal Motor Cortex and Control of Speech in Humans," *The Neuroscientist*, vol. 17, no. 2, pp. 197–208, 2011.

[11] J. Wang, J. R. Green, and A. Samal, "Individual Articulator's Contribution to Phoneme Production," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7785–7789, IEEE, 2013.

[12] J. Wang, A. Samal, P. Rong, and J. R. Green, "An Optimal Set of Flesh Points on Tongue and Lips for Speech-Movement Classification," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 15–26, 2016.

[13] J. Wang, S. Hahm, and T. Mau, "Determining an Optimal Set of Flesh Points on Tongue, Lips, and Jaw for Continuous Silent Speech Recognition," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pp. 79–85, 2015.

[14] S. Hahm, J. Wang, *et al.*, "Silent Speech Recognition from Articulatory Movements Using Deep Neural Network," in *Proc. of the International congress of phonetic sciences*, 2015.

[15] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–1229, IEEE, 2007.

[16] K. Richmond, P. Hoole, and S. King, "Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus," in *Interspeech*, pp. 1505–1508, 2011.

[17] M. Stella, A. Stella, F. Sigona, P. Bernardini, M. Grimaldi, and B. G. Fivela, "Electromagnetic Articulography with AG500 and AG501," in *Interspeech*, pp. 1316–1320, 2013.

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[19] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-Generalized Cepstral Analysis-a Unified Approach to Speech Spectral Estimation," in *Third International Conference on Spoken Language Processing*, 1994.

[20] M. Morise, "D4C, a Band-Aperiodicity Estimator for High-Quality Speech Synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[21] A. Wrench, "Mocha-timit," *Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database*, 1999.

[22] A. Wrench and K. Richmond, "Continuous Speech Recognition Using Articulatory Data," 2000.