Interspeech 2018
2-6 September 2018, Hyderabad

Articulation-to-Speech Synthesis Using Articulatory Flesh Point Sensors' Orientation Information

Beiming Cao[1], Myungjong Kim[1], Jun R. Wang[1], Jan van Santen[3], Ted Mau[4], Jun Wang[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders, University of Texas at Dallas, United States
[3]Center for Spoken Language Understanding, Oregon Health & Science University, United States
[4]Department of Otolaryngology - Head and Neck Surgery
University of Texas Southwestern Medical Center, United States

{beiming.cao, myungjong.kim, jun.wang3, wangjun}@utdallas.edu; vansantj@ohsu.edu; ted.mau@utsouthwestern.edu

Abstract

Articulation-to-speech (ATS) synthesis generates audio waveform directly from articulatory information. Current works in ATS used articulatory movement information (spatial coordinates) only. The orientation information of articulatory flesh points has rarely been used, although some devices (e.g., electromagnetic articulography) provide that. Previous work indicated that orientation information contains significant information for speech production. In this paper, we explored the performance of applying orientation information of flesh points on articulators (i.e., tongue, lips and jaw) in ATS. Experiments using articulators' movement information with or without orientation information were conducted using standard deep neural networks (DNNs) and long-short term memory-recurrent neural networks (LSTM-RNNs). Both objective and subjective evaluations indicated that adding orientation information of flesh points on articulators in addition to movement information generated higher quality speech output than using movement information only.

Index Terms: articulation-to-speech synthesis, orientation information, deep neural network

1. Introduction

Articulation-to-speech (ATS) synthesis directly maps articulatory information to speech [1, 2, 3]. In addition to contribute a better understand of how articulatory movements are mapped to speech, ATS has clinical implications as well. For example, ATS can be the software component in silent speech interfaces (SSIs) which are systems enabling speech communication when an audible acoustic signal is unavailable [4]. SSIs will benefit individuals after laryngectomy (a surgical removal of larynx due to the treatment of laryngeal cancer). These individuals lose their voice but they can still articulate. Current treatments (i.e., esophageal, trachea-esophageal puncture, and electrolarynx) for these individuals typically produce mechanical or hoarse sounds, which are difficult to understand. SSIs have a potential of generating synthesized speech with natural sounding voice or even laryngectomee's own voice [5].

ATS has recently gained great interest in SSI [4, 6], because ATS directly generates speech signals from articulatory information with slight delay. Another articulation to speech conversion design in SSI coverts articulation information into text with silent speech recognition (SSR) [7] and then drive a text-to-speech synthesis (TTS) [8, 9]. The SSR+TTS design always causes a delay because SSR takes time for decoding, also TTS typically requires text processing and analysis stage. Although the quality of synthesized speech of ATS is still not as good as text-to-speech synthesis due to lacking of textual information, the speech output [10, 11, 12] of ATS has been recently improved to a level that has the potential for SSI applications.

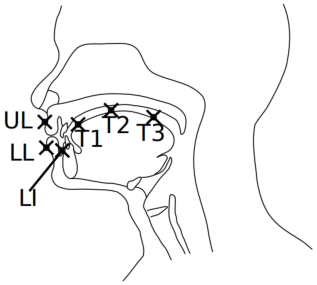
A variety of sensing technologies have been used to capture articulatory movement including electromagnetic articulography (EMA) [13], permanent magnet articulography (PMA) [10, 11, 12], ultrasound [14], and surface electromyography (sEMG) [15]. Most of current ATS works used only the articulatory movement information (spatial coordinates), although magnetic tracking technologies provide orientation information of sensors attached to the articulators (e.g., tongue, lips, and jaw) and recent studies suggest the sensor orientation information is significant in speech production [16].

EMA generates a magnetic field and tracks small flesh point sensors attached to articulators in the electromagnetic field. Modern 3D EMAs such as Wave (NDI Inc., Waterloo, Canada), Carstens AG500, and AG501 (Carsens Medizinelectronik, Lenglern, Germany) obtain 3-dimensional position data and 2-dimensional orientation data including rotations around lateral axis (pitch) and longitudinal axis (roll). Two-dimensional rotations (pitch and roll) are capable of defining a three-dimensional orientation vector with fixed length using spherical coordinates.

We assumed that orientation information of EMA sensors may provide information [17] for articulatory speech synthesis. A previous study [18] indicated that using two tongue sensors with orientation and position data provides the equivalent amount of information as four sensors with only 2D (x and y) positional data [19]. Thus, orientation information of articulatory flesh point sensors may capture useful information to model the relationship between articulatory movements and acoustic features. However, orientation EMA data has rarely been used in ATS.

In this paper, we investigated the effectiveness of using orientation information alone and together with articulatory movement information in articulation-to-speech (ATS) synthesis. First, we explored the performance of orientation components in each of three directions (x, y, and z) and their combinations in ATS. Then, the performance of ATS using both orientation and position information was validated. Two neural network models were used in the experiments: standard deep neural networks (DNNs) and long short-term memory-recurrent neural networks (LSTM-RNNs). The synthesized speech utterances were evaluated both objectively and subjectively. The ob-

10.21437/Interspeech.2018-2484



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 dataset (The picture is adapted from [20]).*

jective evaluation is the accuracies of acoustic parameters prediction. The subjective evaluation is the average preference scores in naturalness and speaker voice identity (similarity to original voice) by 10 listeners.

2. Dataset

The mngu0 dataset is a corpus of articulatory data of different forms acquired from one male British English speaker [20]. We used the EMA subset [20] of mngu0 which consists of audio and EMA data with 1,354 sentences recorded using Carstens AG500 [21]. The total duration of the speech data is about 67 mins [20]. Based on the training, development and testing file list provided by the dataset, the whole dataset was separated to a training set with 1,226 sentences, a development set with 63 sentences, and a testing set with 65 sentences. There is no overlap between the training, development, and testing sets.

The raw EMA data of mngu0 dataset tracks 12 sensor coils in 3D space with two angles of rotation [20]. In this study, we used the 3D position vector $P[xyz]$ of 6 sensors (Figure 1): upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD) extracted from raw EMA data, and the 3D orientation vector of each sensor provided by the dataset. Here, x is left-right, y is anterior-posterior, and z is vertical. The movement of the head was subtracted from these sensors' position data to obtain head-independent articulatory movement. The sampling rate of EMA data is 200 Hz. The 3D orientation vector $O[xyz]$ is a unit length vector which represents the orientation of a sensor in 3D space (Figure 2). The audio data with a sampling rate of 16 kHz was simultaneously recorded with articulatory data [20].

3. Methods

3.1. Articulation-to-Speech Synthesis

In this study, we implemented ATS models that predict acoustic parameters from articulatory position and orientation data with a trained DNN or LSTM-RNN. The predicted acoustic features include: Mel-generalized cepstral coefficients (MGCs) [22], band aperiodicities (BAPs) [23], logarithm of fundamental frequencies (logF0), and voiced/unvoiced (V/UV) labels. Accordingly, the objective evaluation of experimental results is the prediction accuracies of these features, which are mel-coefficient distortion (MCD), band aperiodicities distortion, root mean square error of fundamental frequencies (F0-RMSE), and voiced/unvoiced error rate. The input of ATS includes sensor position vector P, orientation vector O, and their combinations. All the articulatory data were concatenated with their first and second order derivatives as the input of the neural network models.

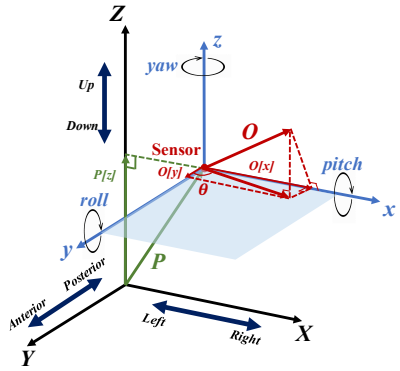


Figure 2: *Demonstration of orientation and position vectors of an EMA sensor.*

Table 1: *Experimental setup.*

Acoustic Feature	187-dim. vectors
Mel-generalized coefficients (MGCs)	(60-dim. vectors) + Δ + ΔΔ (180-dim.)
Band aperiodicities (BAPs)	(1-dim. vectors) + Δ + ΔΔ (3-dim.)
log-F0	(1-dim. vectors) + Δ + ΔΔ (3-dim.)
Voiced/Unvoiced (V/UV) label	1-dim.
Sampling rate	16000 Hz
Windows length	25 ms
Articulatory Feature	
Articulatory position (6 sensors)	(18-dim. vectors) + Δ + ΔΔ (54-dim.)
Articulatory orientation (6 sensors)	(18-dim. vectors) + Δ + ΔΔ (54-dim.)
Common	
Frame rate	5 ms
DNN Topology	
Input	**Orientation**: From 18-dim. to 54-dim. **Position**: 54-dim **Combined**: From 72-dim. to 108-dim.
Output.	187-dim. acoustic feature
No. of nodes each hidden layer	512
Depth	6-depth hidden layers
Learning rate	0.003
Batch size	128
Epoch	25
Optimizer	SGD
LSTM Topology	
Input	**Position**: 54-dim **Combined**: 90-dim.
Output	187-dim. acoustic feature
No. of nodes each hidden layer	256
Depth	3-depth hidden layers
Learning rate	0.003
Batch size	1024
Epoch	50
Optimizer	Adam
Vocoder	WORLD

Long short-term memory-recurrent neural networks (LSTM-RNNs) can model long-range temporal information by overcoming the vanishing gradient problem in conventional recurrent neural networks (RNNs). LSTM-RNN based models have been successfully used in ATS with only articulatory position information by outperforming DNN-based ATS [12, 10, 24]. Therefore, we adopted LSTM-RNN-based ATS to model the long-range temporal relationship between acoustic parameters and both the articulatory position and orientation

Table 2: Results of using orientation (O) and positional information (P) separately.

	$O[x]$	$O[y]$	$O[z]$	$O[xy](Yaw)$	$O[xz](Roll)$	$O[yz](Pitch)$	$O[xyz]$	P
MCD (dB)	6.57	5.59	5.55	5.29	5.21	5.31	5.19	4.895
BAP (dB)	0.193	0.175	0.176	0.165	0.165	0.168	0.163	0.159
F0-RMSE (Hz)	10.80	10.57	11.00	10.24	10.31	10.38	10.14	10.162
V/UV (%)	23.49	17.88	18.08	16.03	15.92	16.44	15.93	14.378

information.

After the acoustic parameters were predicted, the WORLD [25] voice encoder was used to generate speech waveform. The neural network models and objective evaluations were implemented with Merlin toolkit [26]. More detailed information of experimental setup is shown in Table 1.

3.2. Orientation Information for ATS

To understand if orientation information alone of the EMA sensor data is significant in ATS, we first conducted experiments using only three-dimensional orientation vector provided by the dataset. As shown in Figure 2, for each sensor, the vector $P[xyz]$ denotes the position vector P from the head sensor to it, the vector $O[xyz]$ defined a unit vector which represents the orientation of the sensor. To clarify, O is numerically independent of P. In other words, changes in orientation only correlated to sensors' self rotation rather than sensors' movement direction. $O[x]$, $O[y]$, and $O[z]$ denote the projection of sensor rotation amount in the x, y, and z axes, respectively. $O[xy]$, $O[xz]$, and $O[yz]$ are the concatenations of $O[x]$, $O[y]$, and $O[z]$, providing the rotation information of sensors in the $x-y$ (yaw), $x-z$ (roll), and $y-z$ (pitch) planes (Figure 2), respectively. $O[xyz]$ is the orientation of sensors in the real world which combines rotations in all three dimensions. In addition, $O[x]$ is the vector sum of yaw (rotation around the z axis) and roll (rotation around the y axis) rotation's projection on the x axis. Although $O[x]$ contains projections of both yaw and roll rotations, it contains less information than directly concatenating them. Similarly, $O[y]$ contains pitch and yaw, $O[z]$ contains roll and pitch.

3.3. Combination of Position and Orientation Information for ATS

To determine if flesh point articulatory orientation information can be complementary to position information in ATS, we conducted experiments using positional data with and without orientation information. First, to obtain the baseline results, we used DNN to evaluate the ATS performance of using only 3D position vector ($P[xyz]$). Then, DNN-based ATS experiments using all combinations of orientation components ($O[x]$, $O[y]$ and $O[z]$) along with P were conducted to find the best combination of orientation components.

After that, we fed the best combination of position and orientation components into an LSTM-RNN-based ATS to determine if LSTM-RNN outperforms DNN. We hypothesized that LSTM-RNN might show better performance than DNN for the combination of orientation and position information [12].

4. Results and Discussion

4.1. ATS Using Orientation and Positional Data Separately

Table 2 shows the objective measures (MCD, BAP, F0-RMSE, and V/UV%) of using the individual components of articulatory orientation vectors: $O[x]$, $O[y]$, $O[z]$, and their combinations in DNN-based ATS. Interestingly, the performance with all the orientation components ($O[xyz]$) is comparable to that using positional information ($P[xyz]$) in terms of F0-RMSE, BAP, MCD, and V/UV error rates. The results indicate even orientation information alone is useful in ATS.

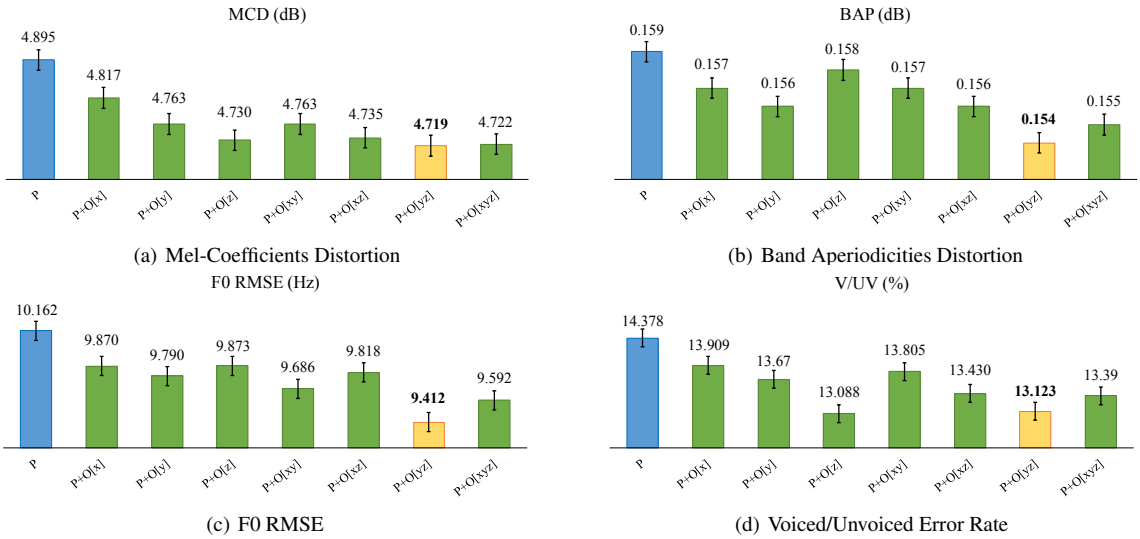
Regarding the performance of individual component of the orientation vectors, we observed that $O[xyz]$ outperformed other combinations in all evaluations (Table 2). This result is not surprising, because higher dimensions may contain more information. Also, all the orientation vectors with two components such as $O[xy]$ produced better results than all the orientation vectors with single component. For single component, $O[z]$ outperformed $O[y]$, whereas $O[y]$ was better than $O[x]$.

4.2. ATS Using Combined Orientation and Positional Data

Figure 3 shows the results of using orientation combined with three-dimensional position vector P in DNN-based ATS (Lower values indicate better performances, the yellow bars are the best performed in the figure). When orientation information was added, the ATS performance was significantly improved compared to positional information alone. $P + O[yz]$ outperformed all other combinations in all evaluations in DNN-based ATS. For P with single component of O, $P + O[z]$ performed the best in MCD and V/UV prediction, whereas $P + O[y]$ performed the best in F0 and BAP prediction. Besides, except for F0 RMSE, $P + O[xz]$ outperformed $P + O[xy]$ in other three objective evaluations.

Next, we evaluated the combination of $P[xyz] + O[yz]$ in LSTM-RNN-based ATS and compared the results to LSTM-RNN-based ATS with P only (Figure 4). The results indicated that LSTM-RNN-based ATS follows the trend that adding $O[yz]$ to $P[xyz]$ would improve the performance. Moreover, LSTM-RNN outperformed DNN in ATS using P both with and without $O[yz]$. These results show the effectiveness of applying orientation information in ATS based on LSTM-RNN.

We observed $O[yz]$ was the most helpful set among all the combinations when using both position and orientation data (Figure 3). According to speech science, healthy adult speakers rarely intend to move their tongue left or right during speech [27], except for producing lateral sounds like "/l/" and "/r/" which may require oriention change in left and right direction. However, all sensors were attached to the middle of articulators in this study, producing lateral sound still don't led to left or right rotations. Therefore, the rotation information in the $y-z$ plane is more correlated with articulatory movement than those in other two planes. Thus, we think the $O[yz]$ component (pitch, rotation around x) contains the most information of speech production. In contrast, it is also observed that $O[xyz]$ was better than $O[yz]$ when only orientation data was used (Table 2). Moreover, other orientation combinations performed differently in ATS using orientation only and ATS using both orientations and positions. Given the different performance patterns in Table 2 and Figure 3, we believe that there are correlations between the orientation components and position informa-



(a) Mel-Coefficients Distortion

(b) Band Aperiodicities Distortion

(c) F0 RMSE

(d) Voiced/Unvoiced Error Rate

Figure 3: *Results of objective evaluations using positional data with or without orientation information.*

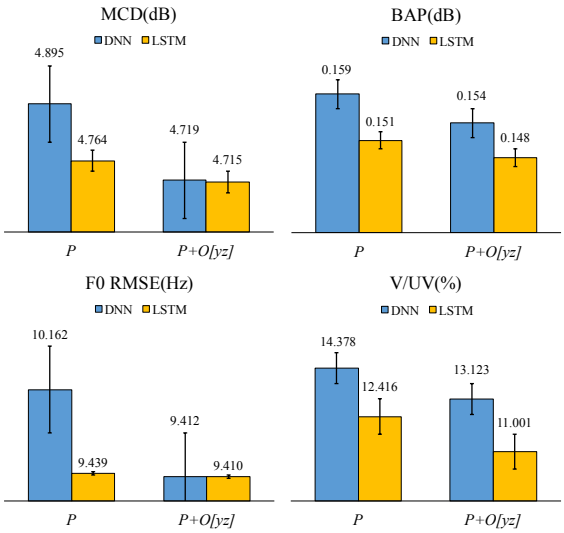


Figure 4: *Objective evaluation of LSTM-RNN.*

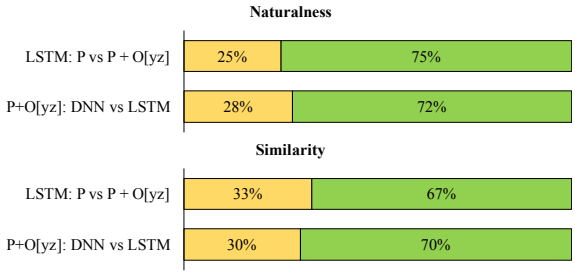


Figure 5: *Results of subjective evaluations.*

tion which need to be discovered in the future.

Finally, subjective evaluations were conducted using positional information (P) with and without orientation information ($O[yz]$) on DNN-based and LSTM-RNN-based ATS methods. Figure 5 gives the average preference scores in naturalness and similarity to original voice given by 10 listeners (20 sentences of 65 testing samples were evaluated by listeners). For the evaluation, firstly we asked listeners to choose their preferences in terms of naturalness and similarity between speech samples generated from LSTM-RNN-based ATS using P and those using $P + O[yz]$. Then, given $P + O[yz]$, we let listeners choose their preferences between LSTM-RNN and DNN-based ATS. Our results show that adding $O[yz]$ to P outperformed P alone in both naturalness and similarity. In addition, LSTM-RNN-based ATS performed better than DNN-based ATS in both subjective evaluations.

Limitation. Although the experimental results confirmed our hypothesis that adding EMA sensors' orientation information to positional information would improve the ATS performance, this study is still preliminary with only one subject. Further studies with EMA data from multiple subjects are needed to verify these findings particularly on the performances of the individual orientation components in the x, y, and z directions.

5. Conclusions

This study investigated the effectiveness of applying orientation information of sensors attached to flesh points on articulators (tongue, lips, and jaw) in articulation-to-speech (ATS) synthesis. EMA sensors' position information with and without orientation information were used as input to DNN-based ATS and LSTM-RNN-based ATS. The experimental results proved the effectiveness of applying orientation information in ATS. Adding orientation information representing pitch rotation ($O[yz]$) produced the best ATS results on both the DNN-based ATS and LSTM-RNN-based ATS. In addition, LSTM-RNN-based ATS outperformed DNN-based 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. We thank all volunteering listeners (judges) in the subjective evaluation.

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] D. Qinsheng, Z. Jian, W. Lirong, and S. Lijuan, "Articulatory speech synthesis: a survey," in *Computational Science and Engineering (CSE), IEEE 14th International Conference on*. IEEE, 2011, pp. 539–542.

[3] S. Parthasarathy, J. Schroeter, C. Coker, and M. Sondhi, "Articulatory analysis and synthesis of speech," in *TENCON'89. Fourth IEEE Region 10 International Conference*. IEEE, 1989, pp. 760–764.

[4] 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.

[5] J. M. Gilbert, J. A. Gonzalez, L. A. Cheah, S. R. Ell, P. Green, R. K. Moore, and E. Holdsworth, "Restoring speech following total removal of the larynx by a learned transformation from sensor data to acoustics," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. EL307–EL314, 2018.

[6] R. Li and J. Yu, "An audio-visual 3d virtual articulation system for visual speech synthesis," in *Haptic, Audio and Visual Environments and Games (HAVE), IEEE International Symposium on*. IEEE, 2017, pp. 1–6.

[7] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2323–2336, 2017.

[8] 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 *SSW*, 2007, pp. 294–299.

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] 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.

[11] 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*. ISCA, 2017, pp. 3986–3990.

[12] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[13] 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.

[14] 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.

[15] G. S. Meltzner, J. T. Heaton, Y. Deng, G. D. Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2386–2398, Dec 2017.

[16] A. J. Kolb, M. T. Johnson, and J. Berry, "Interpolation of tongue fleshpoint kinematics from combined EMA position and orientation data," in *INTERSPEECH*. International Speech and Communication Association, 2015, pp. 2177–2181.

[17] J. Berry, A. Kolb, J. Schroeder, and M. T. Johnson, "Jaw rotation in dysarthria measured with a single electromagnetic articulography sensor," *American journal of speech-language pathology*, vol. 26, no. 2S, pp. 596–610, 2017.

[18] P. Hoole, A. Zierdt, and C. Geng, "Beyond 2D in articulatory data acquisition and analysis," in *Proceedings of the fifteenth international congress of phonetic sciences, Barcelona*, 2003, pp. 265–268.

[19] A. J. Kolb, "Software tools and analysis methods for the use of electromagnetic articulography data in speech research," Ph.D. dissertation, Marquette University, 2015.

[20] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus." in *Interspeech*, 2011, pp. 1505–1508.

[21] M. Stella, A. Stella, F. Sigona, P. Bernardini, M. Grimaldi, and B. G. Fivela, "Electromagnetic articulography with ag500 and ag501." in *Interspeech*, 2013, pp. 1316–1320.

[22] 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.

[23] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[24] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1613–1616.

[25] 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.

[26] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[27] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.