# IS ARTICULATION-TO-SPEECH SYNTHESIS LANGUAGE INDEPENDENT? A PILOT STUDY

Beiming Cao[1], Alan Wisler[1], 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
{beiming.cao, alan.wisler, wangjun}@utdallas.edu

## ABSTRACT

Articulation-to-speech (ATS) synthesis is to directly synthesize speech from articulatory information, which does not require textual input. ATS has recently shown the potential for assistive technologies such as silent speech interfaces (SSIs). ATS is theoretically language-independent, since there is no dictionary involved. However, to our knowledge, there is no data-based experiment has been conducted to answer this question, due to lack of the multi-language, articulatory movement data from the same speakers. In this study, we conducted speaker-dependent ATS experiments using data collected from bilingual speakers, who speak two of the three languages: English, Spanish, and Korean. The experimental results indicated the performance was degraded if ATS was trained with a language and tested with another language. Interestingly, we observed the performance of ATS for one language could be improved if some samples of another language were added to the training dataset.

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

## 1. INTRODUCTION

Silent speech interfaces (SSIs) [1] are devices to facilitate speech communication for individuals that are unable to properly vocalize speech sounds. For individuals like laryngectomees (people who have their larynx removed due to the treatment of laryngeal cancer), SSIs provide a way of recovering normal communication. SSIs capture the biosignals with different technologies such as electromagnetic articulograph (EMA) [2], permanent magnetic articulography (PMA) [3, 4, 5, 6], ultrasound [7], non-audible murmur (NAM) [8, 9]. Based on the model is directly or indirectly mapping from articulatory bio-signals to speech, the software design of SSIs typically falls under two major categories. The indirect-mapping design (recognition-and-synthesis approach) includes two steps: a silent speech recognition (SSR) [10, 11, 12, 13] stage for converting non-audio articulatory signals to text, and a text-to-speech (TTS) [14, 15] synthesis stage for converting text into speech. The direct-mapping SSI design is articulation-to-speech (ATS) synthesis [3, 4, 5, 7, 16, 17], which directly maps articulatory information to speech (Figure 1). Compared to the indirect-mapping design, ATS-based SSI has the benefit of real-time and easy implementation. In addition, ATS is independent of textual input whereas the performance of the indirect-mapping SSI highly depends on the accuracy of speech recognition. Because of these advantages, ATS has recently gained increasing attention from researchers in SSI.

Another benefit of direct-mapping method (ATS) is the reduced level of language dependency that is inherent to their design. Whereas indirect-mapping SSI method utilizes SSR systems that are heavily language dependent, no part of a direct ATS system that is explicitly language dependent. This does not mean we should expect direct ATS systems to be completely language independent. While we hypothesize that the mapping between articulatory movements and the sounds that they produce is independent of language, the subset of common motions pertaining to a specific language is not. This means that if an ATS system is asked to synthesize a phoneme that doesn't exist in the language it is trained on, that system is less likely to produce the appropriate speech sound.

While the idea that direct ATS systems are language-independent has been suggested in the literature [18]. It has not been investigated due to lack of multilingual articulatory data from same speakers. In this paper, we utilized data from two bilingual speakers (one English/Spanish speaker, and one English/Korean speaker) to develop ATS systems and examined the degree to which these systems are capable of generalizing to new languages using little or

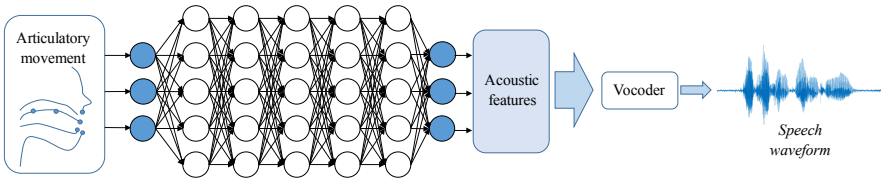



**Figure 1:** Articulation-to-speech Synthesis Model.

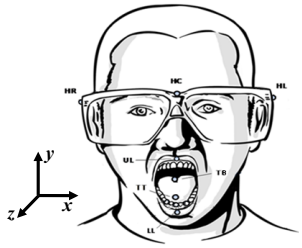



**Figure 2:** Sensor Locations.

no data from the language to be synthesized. Additionally, by examining two different language pairs, English-Spanish and English-Korean, we hope to identify the degree to which phonetic differences between languages inhibit the ability of ATS systems to generalize to new languages.

## 2. DATA COLLECTION

Electromagnetic articulograph (EMA) data from two speakers was used for this study. One female speaker who speaks Spanish and English, and one male speaker who speaks Korean and English participated in the data collection sessions. Each subject repeated a sequence of 132 phrases in each of two languages he or she speaks at their habitual speaking rates. 264 phrases were recorded from each subject. The 132 phrases were selected from phrases which are frequently spoken by the users of augmentative and alternative communication (AAC) devices [19].

### 2.1. Tongue motion tracking device

The EMA and audio data were collected using the Wave system (Northern Digital Inc., Waterloo, Canada). During data recording, the movement of articulators were captured by four sensors attached (two on the tongue and two on the lips). One additional sensor was attached to the forehead for the correction. Two tongue sensors were attached to the tongue tip (TT, 5-10 mm to tongue apex) and tongue back (TB, 20-30mm back from TT) with dental glue (Peri-Acry1 90, GluStitch). Two lip sensors were attached to the middle of upper lip (UL) and lower lip (LL) with tapes. The locations of the sensors are shown in Figure 2. The sampling rate of EMA data is 100Hz. The spatial precision of movement tracking is about 0.5mm [20]. Before the data collection, a three-minute training session was conducted to help participants adapt to the speaker with sensors on their tongue and lips. For each subject, the data collection was completed in one session to ensure that sensor locations are identical for the different language recordings.

To obtain head-independent articulatory movement, first the head motion collected from the forehead sensor was subtracted from the motion of articulator sensors. The derived Cartesian coordinates system is shown in Figure 2: $x$ is lateral direction, $y$ is superior-inferior direction, and $z$ is anterior-posterior direction. All of the EMA data were upsampled to 200 Hz using a spline interpolation procedure to match the 5 milliseconds frame rate of acoustic features.A three-minute adaptation session was conducted to help participants adapt to speaking with sensors on their tongue and lips

**Table 1:** Experimental setup.

| | |
|---|---|
| **Acoustic Feature** | 187-dim. vectors |
| Mel-Cepstral Coefficients (MCCs) | (60-dim. vectors) + Δ + ΔΔ (180-dim.) |
| Band Aperiodicities (BAPs) | (1-dim. vectors) + Δ + ΔΔ(3-dim.) |
| Fundamental Frequency on log scale (log-F0) | (1-dim. vectors) + Δ + ΔΔ (3-dim.) |
| Voiced/Unvoiced (V/UV) label | (1-dim.) |
| Sampling rate | 22050 Hz |
| Windows length | 25 ms |
| **Articulatory Feature** | 36-dim. vectors |
| articulatory movement (4 sensors) | (12-dim. vectors) + Δ + ΔΔ (36-dim.) |
| **Common** | |
| Frame rate | 5 ms |
| **DNN Topology** | |
| Input | Articulatory movement: 36-dim. 3D motion of 4 sensors + Δ + ΔΔ |
| Output. | 187-dim. acoustic feature |
| No. of nodes each hidden layer | 512 |
| Depth | 6-depth hidden layers |
| Learning rate | 0.0035 |
| Batch size | 128 |
| Epoch | 25 |
| Optimizer | SGD |

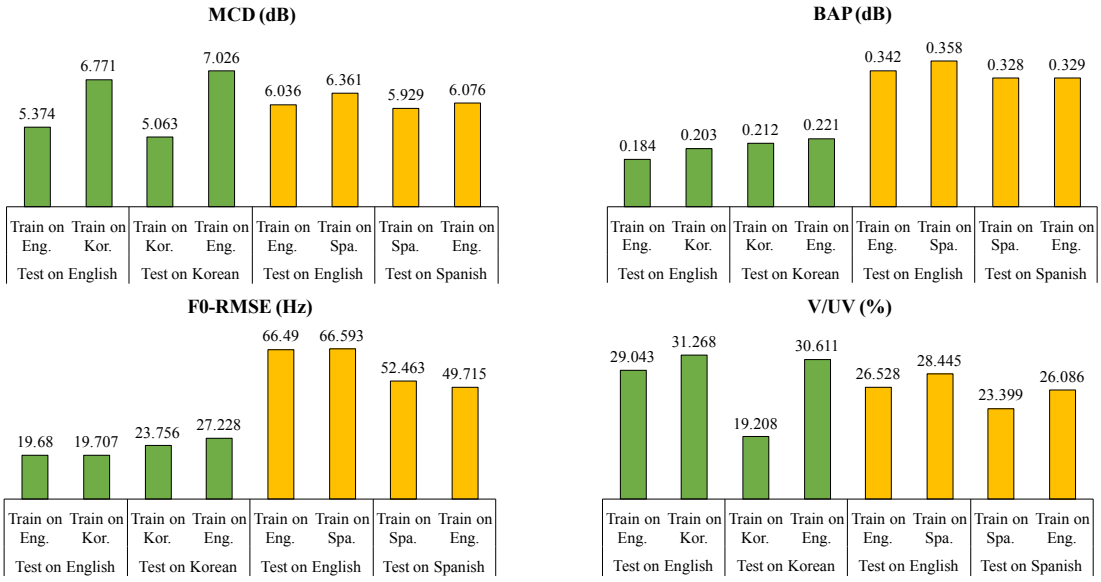



**Figure 3:** Results of Experiment Session 1.

## 3. METHODS

### 3.1. Articulation-to-speech synthesis

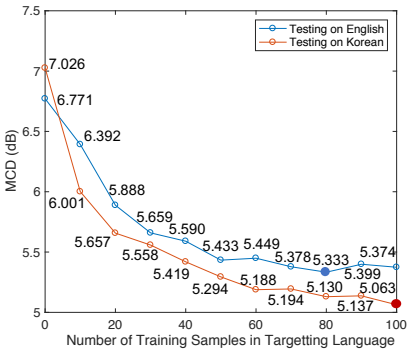
In this study, we implemented an ATS system which uses deep neural network (DNN) to predict acoustic features from the EMA data (Figure 1). The input of DNN is the articulation sensor data concatenated with its first and second order derivative. The output of DNN is the concatenation of acoustic features to be fed to a voice encoder (Vocoder) for speech synthesis. The acoustic features used include: mel-cepstral coefficients (MCCs) [21], band aperiodicities (BAPs) [22], logarithm of fundamental frequencies (Log-F0s), and a voiced/unvoiced flag. These acoustic features are sent to the Vocoder for synthesizing speech. The measurement of ATS performance was the accuracies of acoustic feature prediction. The Vocoder used in this study is the World Vocoder [23]. The detailed experimental setup is shown in Table 1.

### 3.2. Experimental Setup

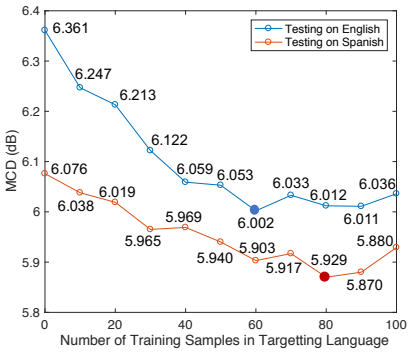
As mentioned, 132 sentences were recorded for each of two languages spoken by each of subject. Each 132-sentence dataset was separated as a training set of 100 sentences, a validation set of 16 sentences and a testing set of 16 sentences. We conducted two major experiment sessions in this study. In the first session, firstly we conducted speaker and language-dependent ATS experiments as the baseline, which means both training and validating data are from same speakers and for same languages. Then we tested the trained ATS models with a different language from same speakers. This goal of this session is to gauge how well single-language ATS systems can generalize to new languages for two different language pairs. In the second session, we used the mixtures of two languages as training data and tested the model on each language. In this experiment, we keep the total number of sentences in training 100, and vary what percentage of the training data is drawn from each language. The goal of this experiment is to measure the degree to any performance degradation related to the language-dependency of the ATS system can be mitigated by the inclusion of some portion of in-language training data.

## 4. RESULTS AND DISCUSSIONS

The results of the first experiment session are presented in Figure 3. Here the lower numbers indicate better performance of ATS. The green color represents the English-Korean speaker, and the yellow color represents the English-Spanish speaker. We can see that all of four objective measurements of ATS: mel-cepstral distortion (MCD), band aperiodicities distortion (BAP), RMSE of fundamental frequency (F0-RMSE) and voiced/unvoiced (V/UV) error rate were increased when the ATS was applied to testing data of different languages. Except for the F0-RMSE was decreased when the ATS trained on English was tested on Spanish. But since F0 is less related to articulatory movement, improvement in F0 estimation does not necessarily indicate a better ATS. We still consider the performance of ATS




(a) English-Korean




(b) English-Spanish

**Figure 4:** Results of Experiment Session 2.

was decreased since both MCD and BAP distortion were increased.

Moreover, we can observe that English-Korean ATS and English-Spanish ATS performed differently. When the ATS trained with English was tested with Korean, the performance was decreased significantly (5.37 dB to 7.03 dB in MCD), and vice versa (5.06 dB to 6.77 dB in MCD). However for the ATS trained with English was tested with Spanish, the MCD was only increased by 0.04 dB (6.04 dB to 6.08 dB), and 0.43 dB MCD was increased when trained with Spanish and tested with English. These observations indicated that Spanish has a higher similarity in articulatory patterns to English than Korean.

Figure 4 shows the mel-cepstral distortions (MCDs) of the second session which is testing ATS trained with the mixtures of two languages with both two languages. Here we present MCD only because it is more correlated to articulatory movement than other acoustic features. The x-axis is the number of the training sentences in testing language. For example, for the Testing on English curves in figures, 90 on x-axis means the ATS was trained with dataset contains 90 English sentences, and 10 the other language sentences.

As can be seen in Figure 4(a), for Korean ATS, the best performance was achieved by using 100% Korean sentences as training data. However, for English the best performance was achieved by using 80% English and 20% Korean. In Figure 4(b), the best performance of Spanish was achieved by using 80% Spanish and 20% English. For English, it is 60% English and 40% Spanish. These numbers indicated that there are similar articulatory patterns between Korean and English, Spanish and English since each of them has improved the performance of ATS for another language or been improved performance of ATS by another language.

## 5. CONCLUSION AND FUTURE WORK

This study investigated the performance of ATS systems when synthesizing speech from languages other than the primary language of the training data. The experiments conducted in this paper yielded three main findings. First, we observe a reduction in the performance of ATS systems when synthesizing languages outside of the training set. Second, we found that this reduction in performance is heavily dependent on the degree of similarity between the language being synthesized and the language used for training. While there exists a large performance loss when using an English-trained ATS system to synthesize Korean, the performance loss when using an English-trained system to synthesize Spanish was found to be relatively minor. Third, it can be observed that at below a certain proportion of inclusion of the second language the results flatten out, with the only difference is beyond that point being attributable to experimental fluctuations.

Although the results of the experiments conducted in this paper are largely in agreement with our prior hypotheses, the reliability of these findings is inherently limited by the data size. Because the data used was collected from only two participants, our findings are hard to generalize to other speakers. Furthermore, as the systems examined in this paper are all speaker-dependent, it remains an open question whether or not these findings will be consistent with those for speaker independent systems. Future work will focus on expanding this study with more speakers.

## 6. ACKNOWLEDGEMENT

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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)*.

[7] 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.

[8] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pp. V–708–11 vol.5, 2003.

[9] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Non-audible Murmur (NAM) Speech Recognition Using a Stethoscopic NAM Microphone," in *INTERSPEECH 2004 - Icslp, International Conference on Spoken Language Processing, Jeju Island, Korea, October*, 2004.

[10] 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*, pp. 1–5, 2015.

[11] M. Kim, B. Cao, T. Mau, and J. Wang, "Multiview Representation Learning via Deep CCA for Silent Speech Recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 2769–2773, 2017.

[12] M. Kim, B. Cao, T. Mau, and J. Wang *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[13] J. Wang and S. Hahm, "Speaker-Independent Silent Speech Recognition with Across-Speaker Articulatory Normalization and Speaker Adaptive Training," *Interspeech*, pp. 2415–2419, 01 2015.

[14] R. W. Sproat and J. P. Olive, "Text-to-Speech Synthesis," *AT&T technical journal*, vol. 74, no. 2, pp. 35–44, 1995.

[15] M. S. Heiga Zen, Andrew Senior, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7962–7966, May 2013.

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

[17] 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.

[18] T. S. Lorenz Diener, "Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018, pp. 3162–3166, 2018.

[19] D. Beukelman and P. Mirenda, "Augmentative and Alternative Communication," 2005.

[20] J. J. Berry, "Accuracy of the NDI Wave Speech Research System," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–1301, 2011.

[21] 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.

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

[23] 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.