A Wearable Electromagnetic Articulograph (EMA) for Silent Speech Interface

Shravan Ravi, Beiming Cao, Nordine Sebkhi, Arpan Bhavsar, Omer T. Inan, Wen Xu, Jun Wang

Introduction

Oral communication plays an integral role in our daily life. To that effect, significant focus has been placed in alleviating the hardships of patients after laryngectomy. Progress has been made in the development of silent speech interfaces (SSIs) which convert articulatory movement data into synthesize speech [1, 2]. Despite recent innovations, there is still no available wearable device for everyday use.

This study addresses the gap by developing a wearable device based on permanent magnet localization (PML) with an inertial measurement unit (IMU) to track tongue motion [3]. The IMU's magnetometer measures the local magnetic field which compensates for a small ($6 \times 6 \times 0.8$ mm^3) tracer's orientation using the IMU's accelerometer and gyroscope.

To test the efficacy of this new SSI prototype, we conducted an experiment for speaker dependent, isolated vowel classification using a support vector machine (SVM), deep neural network (DNN), and convolutional neural network (CNN). Our preliminary results show similar levels of accuracy to previously reported results [6] with commercial EMA systems (AG500) at the classification task, demonstrating the potential of the wearable SSI.

Data Collection

This study was approved by the IRBs at the University of Texas at Austin and Georgia Institute of Technology. Due to COVID19, two researchers participated in the study for data collection. Participants were able to freely move their head and body, and recorded measurements both sitting and standing at home. No prior speech, language, hearing, or cognition history was reported.

Figure 1 shows the prototype, where a localized magnetic field is generated from a magnetic strip in the patient's eyewear and tracks an IMU placed on the patient's tongue. This IMU data is processed using a localization algorithm to determine the sensors movement and supplies researchers with even more than x, y, and z coordinates (e.g., sensor orientation, quaternion, and magnetic field data) [3]. Our prior work showed that a single sensor on the tongue tip could be sufficient for SSI to produce intelligible speech [4]. Therefore, use one sensor on the tongue tip in this SSI application [5].

Fig.1. Wearable EMA device

Eight English vowels in consonant-vowel-consonant (CVC) context - /bab/, /bib/,/beb/, /bæb/, /bʌb/, /bɔb/, /bob/, /bub/ - were used as vowel stimuli. These eight selected syllables were utilized in previous work [6] when originally proving the efficacy of standard EMA devices due to the vowels range of articulatory motion and the consonant context's ability to reduce coarticulation effects.

To leverage the additional data provided by our device, four combinations of the available EMA readings were utilized and augmented to maximize the accuracy of each algorithm (XYZ, XYZ + RPY (orientation), XYZ + quaternion, XYZ + RPY + magnetic field data, XYZ + quaternion + magnetic field data).

Results

Figure 2 gives the average cross-validation accuracy of the speaker dependent models for every model-feature combination.

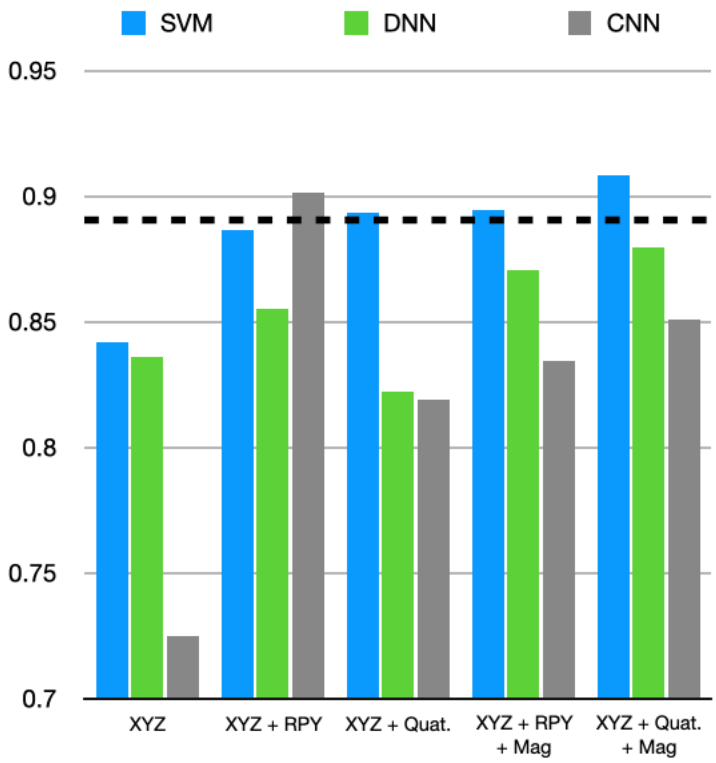

Fig. 2. Cross-validation accuracy for each model-feature combination. Dotted black line delineates reported SVM accuracy from a commercial EMA (AG500) [6]

The highest average classification accuracy across the feature combinations was 90.85%, using SVM, which is comparable to the reported accuracy in [6] (89.05%), which also used SVM on data collected using a commercial EMA (AG500).

Discussion

The preliminary results are encouraging in demonstrating the potential of our new device for SSI applications, although there are still limitations including small data size and few number of subjects. Moreover, only isolated vowels were used as stimuli.

Although there is only one tracer in the current prototype, the proposed system can scale to incorporate more, when needed [7] [8], for other applications such as vision-based speech therapy [9], and even secondary language learning. Future analysis is necessary containing a large, more representative population with other stimuli (e.g., consonants and phrases) using both our device and commercial EMA. We are currently actively collecting new data and expect more robust results by the time of the conference (February 2022).

References

[1] Lee, W., Seong, J. J., Ozlu, B., Shim, B. S., Marakhimov, A., & Lee, S. (2021). Biosignal sensors and DEEP Learning-Based speech Recognition: A review. *Sensors*, *21*(4), 1399. https://doi.org/10.3390/s21041399

[2] Kim, M., Sebkhi, N., Cao, B., Ghovanloo, M., & Wang, J. (2018). Preliminary test of a wireless magnetic tongue tracking system for silent speech interface. *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. https://doi.org/10.1109/biocas.2018.8584786

[3] Sebkhi, N., Sahadat, N., Hersek, S., Bhavsar, A., Siahpoushan, S., Ghoovanloo, M., & Inan, O. T. (2019). A deep neural network-based permanent MAGNET localization for Tongue Tracking. *IEEE Sensors Journal*, *19*(20), 9324–9331. https://doi.org/10.1109/jsen.2019.2923585

[4] Cao, B., Tsang, B., & Wang, J. (2019). Comparing the performance of individual articulatory flesh points for articulation-to-speech synthesis. In Proceedings of the 19th International Congress of Phonetic Sciences, pp. 3041-3045.

[5] Sebkhi, N., Bhavsar, A., Anderson, D. V., Wang, J., & Inan, O. T. (2021). Inertial measurements for tongue motion tracking based on magnetic localization with orientation compensation, IEEE Sensors Journal, 21(6), 7964-7971.

[6] Wang, J., Green, J. R., Samal, A., & Yunusova, Y. (2013). Articulatory distinctiveness of vowels and consonants: A data-driven approach. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1539–1551. https://doi.org/10.1044/1092-4388(2013/12-0030)

[7] Wang, J., Hahm, S., & Mau, T. (2015). Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition, Workshop on Speech and Language Processing for Assistive Technologies, pp. 79-85.

[8] Wang, J., Samal., A., Rong, P., & Green, J. R. (2016). An optimal set of flesh points on tongue and lips for speech-movement classification, Journal of Speech, Language, and Hearing Research, 59, 15-26.

[9] Katz, W., Campbell, T. F., Wang, J., Farrar, E., Eubanks, C., Balasubramanian, A., Prabhakaran, B. & Rennaker, R. (2014). Opti-Speech: A real-time, 3D visual feedback system for speech training, Proc. Interspeech, pp. 1174-1178.