


Research Note

MagTrack: A Wearable Tongue Motion Tracking System for Silent Speech Interfaces

Beiming Cao,^{a,b} Shravan Ravi,^c Nordine Sebkhii,^d Arpan Bhavsar,^d Omer T. Inan,^d Wen Xu,^e and Jun Wang^{b,f} 

^aDepartment of Electrical and Computer Engineering, The University of Texas at Austin ^bDepartment of Speech, Language, and Hearing Sciences, The University of Texas at Austin ^cDepartment of Computer Science, The University of Texas at Austin ^dSchool of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta ^eDivision of Computer Science, Texas Woman's University, Denton ^fDepartment of Neurology, The University of Texas at Austin

ARTICLE INFO

Article History:

Received June 1, 2022

Revision received September 6, 2022

Accepted February 20, 2023

Editor-in-Chief: Mili Kuruvilla-Dugdale

Editor: Elaine Kearney

https://doi.org/10.1044/2023_JSLHR-22-00319

ABSTRACT

Purpose: Current electromagnetic tongue tracking devices are not amenable for daily use and thus not suitable for silent speech interface and other applications. We have recently developed MagTrack, a novel wearable electromagnetic articulograph tongue tracking device. This study aimed to validate MagTrack for potential silent speech interface applications.

Method: We conducted two experiments: (a) classification of eight isolated vowels in consonant–vowel–consonant form and (b) continuous silent speech recognition. In these experiments, we used data from healthy adult speakers collected with MagTrack. The performance of vowel classification was measured by accuracies. The continuous silent speech recognition was measured by phoneme error rates. The performance was then compared with results using data collected with commercial electromagnetic articulograph in a prior study.

Results: The isolated vowel classification using MagTrack achieved an average accuracy of 89.74% when leveraging all MagTrack signals (x , y , z coordinates; orientation; and magnetic signals), which outperformed the accuracy using commercial electromagnetic articulograph data (only y , z coordinates) in our previous study. The continuous speech recognition from two subjects using MagTrack achieved phoneme error rates of 73.92% and 66.73%, respectively. The commercial electromagnetic articulograph achieved 64.53% from the same subject (66.73% using MagTrack data).

Conclusions: MagTrack showed comparable results with the commercial electromagnetic articulograph when using the same localized information. Adding raw magnetic signals would improve the performance of MagTrack. Our preliminary testing demonstrated the potential for silent speech interface as a lightweight wearable device. This work also lays the foundation to support MagTrack's potential for other applications including visual feedback–based speech therapy and second language learning.

Spoken language is conveyed via well-coordinated speech movements (Gafos & van Lieshout, 2020). Articulatory movement during speech is key to studying the underlying mechanisms of speech production. Scientifically, for example, whether the principles that apply to limb

movements (e.g., Fitts' law) hold true for speech movement is still unsure (Gafos & van Lieshout, 2020). How exactly the tongue and lip motions are mapped to speech outcomes is still poorly understood (Green et al., 2013). Clinically, it is important to understand how tongue and lip motion patterns are impacted by specific disorders. A better understanding of articulatory movement could improve early disease detection, monitor disease progression, and optimize the efficacy of therapeutic drug trials for neurological disorders such as amyotrophic lateral sclerosis (Green et al., 2013; Hahn et al., 2015; J. Wang, Kothalkar, et al., 2016).

Correspondence to Jun Wang: jun.wang@austin.utexas.edu. **Publisher Note:** This article is part of the Special Issue: Select Papers From the 2022 Conference on Motor Speech—Basic Science and Clinical Innovation. **Disclosure:** The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

Practically, articulatory movement is used in some assistive devices such as silent speech interface (SSI; Denby et al., 2010; Gonzalez-Lopez et al., 2020; Schultz et al., 2017) and tongue-controlling systems (e.g., tongue-controlling wheelchair; Sebkhi et al., 2022). Articulatory movement can also be used for providing visual feedback of tongue motion in speech therapy (Katz et al., 2014) and in language and pronunciation learning (Benway et al., 2021; Katz & Mehta, 2015).

SSIs are devices that convert articulatory movement to speech and have the potential of recovering the speech ability for people who lost their voice but can still articulate such as laryngectomees (Denby et al., 2010; Gonzalez-Lopez et al., 2020; Schultz et al., 2017). Laryngectomees are individuals who have their larynx surgically removed due to the treatment of cancer (Bailey et al., 2006). Without their larynx, laryngectomees are unable to produce speech sounds. Laryngectomees currently use three types of speech modes in their daily communication (called *alaryngeal speech*): electrolarynx (Kaye et al., 2017), tracheo-esophageal puncture speech (Chen et al., 2001), and esophageal speech (Nijdam et al., 1982). Electrolarynx speech relies on a battery-powered, external electromechanical device that produces either pharyngeal or oral cavity vibrations (Kaye et al., 2017). Tracheo-esophageal puncture speech requires an additional surgery that makes a one-way valve from the trachea to the esophagus, which allows airflow from the trachea to drive the vibration of the throat wall (Chen et al., 2001). Esophageal speech involves ingesting air into the esophagus and then expelling it to drive throat wall vibration to produce sound (Cao et al., 2021). Alaryngeal speech typically results in an unnatural-sounding voice (extremely hoarse or robotic), which discourages speakers' willingness to communicate and often results in social isolation and depression (Eadie et al., 2016). Although the articulation patterns of laryngectomees are different from those of healthy speakers (e.g., longer duration and more lateral movement; Teplansky et al., 2020), their patterns are consistent in producing the same speech. SSIs have been recently demonstrated to be able to generate more natural-sounding voice for laryngectomees (Cao et al., 2021).

SSIs are typically implemented using two types of algorithm designs. The first is the recognition and synthesis design (Cao et al., 2021; Denby et al., 2010; J. Wang et al., 2014), which recognizes text from articulation (Fagan et al., 2008; Kim et al., 2017; Meltzner et al., 2018) and then employs text-to-speech synthesis (Taylor, 2009) to convert the recognized text to speech. The process to recognize text from articulation is called *silent speech recognition* (SSR). The second SSI design is the direct synthesis design (Cao et al., 2018; Diener et al., 2019; Gonzalez et al., 2017; Shandiz et al., 2021), which

directly maps articulation to speech. Significant progress has been made by studies using both of the SSI algorithm designs. Cao et al. (2021) implemented SSIs for laryngectomees using the first design (recognition and synthesis), from which higher naturalness but lower intelligibility audio samples were generated, compared with alaryngeal speech. Kim et al. (2017) explored multiple speaker normalization approaches to improve the speaker-independent SSR. End-to-end automatic speech recognition models have been investigated in SSR as well (Kimura et al., 2020). Recently, representation learning has been demonstrated to be effective in SSR (H. Wang et al., 2021). For the second design of SSI (direct synthesis design), two recent studies have demonstrated that the direct synthesis-based SSIs have the potential of generating audio samples with high intelligibility (Cao et al., 2019; Gonzalez & Green, 2018). To mitigate the effort in synchronous articulatory and acoustic data collection, Gonzalez et al. (2022) proposed an algorithm based on multiview-based time warping (Gonzalez-Lopez et al., 2022) for aligning the separately collected articulatory and acoustic data. Cao et al. (2022) proposed an approach of converting other speakers' audio data to the target speaker for SSI training for maintaining speaker identity of the output speech.

Other than the software designs, a key challenge for SSI development is to track the tongue motion patterns during speech using wearable devices suitable for daily use. Several articulation tracking devices for SSI have been developed and used. These devices include electromagnetic articulograph (EMA; Cao et al., 2018; Kim et al., 2017; Rebernik, Jacobi, Jonkers, et al., 2021), permanent magnetic articulograph (Gonzalez et al., 2017), surface electromyography (Diener et al., 2019), and ultrasound imaging (Shandiz et al., 2021). Most of the devices have shown their potential for SSI. Our team has been using EMA for SSI algorithm development (Cao et al., 2018, 2021; J. Wang et al., 2013, 2015).

There were two commercially available EMA devices: AG series (AG500, AG501) by Carstens (Yunusova et al., 2009) and Wave (renamed as Vox for the next generation) by Northern Digital Inc. (NDI; Berry, 2011). Unfortunately, NDI discontinued their devices in 2020 (Denny, 2020), and the AG series is currently the only commercially available EMA (see AG501 in Figure 1a and Wave in Figure 1b). Both Carstens and NDI EMA devices have high spatial accuracies (0.12–1.37 mm for Wave, 0.5 mm for AG 500, and claimed 0.3–1.0 mm for AG501; Berry, 2011; Rebernik, Jacobi, Tiede, & Wieling, 2021; Savariaux et al., 2017; Yunusova et al., 2009). Despite our progress in algorithm development for SSI using data collected using both Carstens AG series and NDI Wave, current EMAs are limited to lab use and are not wearable for use in everyday living, as is required for SSIs.

Figure 1. (a) Carstens AG501 system. (b) NDI Wave system.

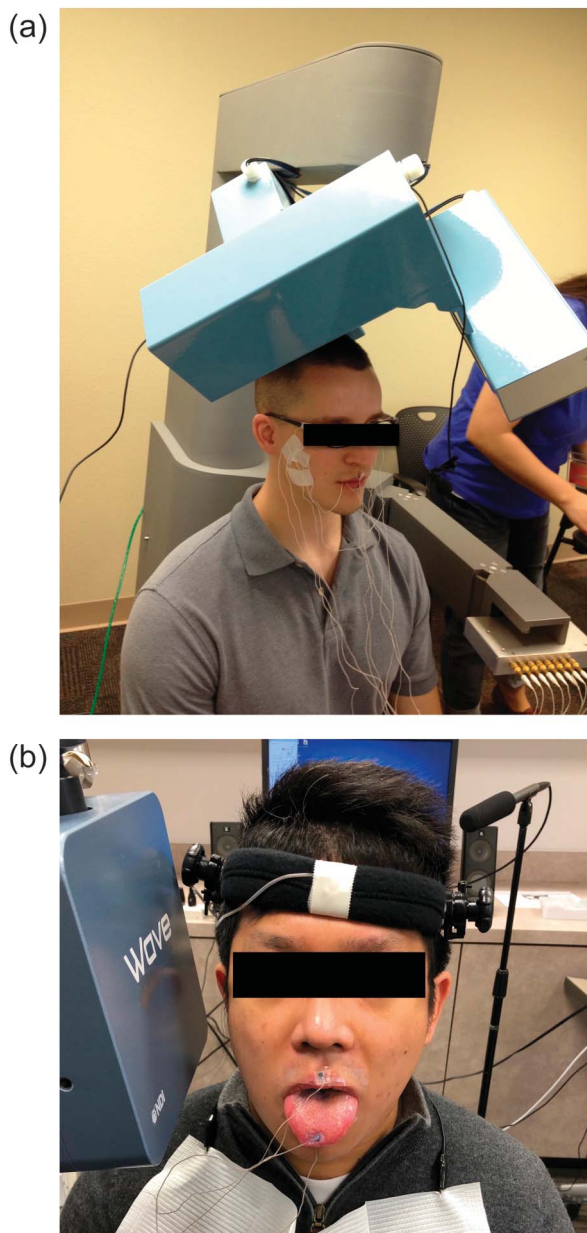
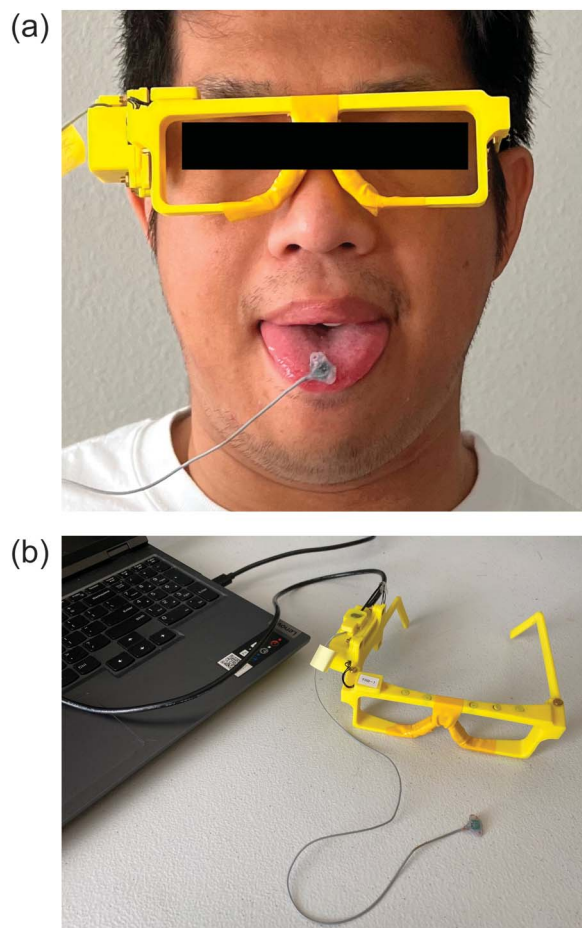


Figure 2. (a) Eyeglasses frame and the tongue motion sensor of the MagTrack system. (b) USB connection between the MagTrack and the recording laptop.



Recently, we developed a novel, lightweight, and wearable EMA device: MagTrack (Sebkhi et al., 2021). MagTrack contains articulation tracking hardware and integrated software. The device (see Figure 2a) consists of a small inertial measurement unit as a sensor attached to the tongue, with a size of $6 \times 6 \times 0.8 \text{ mm}^3$, and an eyeglasses frame for magnetic field generation. During speech, the sensor moves in a local magnetic field generated by the glasses frame (see Figure 2a), and the variation of the magnetic field is captured by the sensor and fed to the MagTrack software. The software localizes

(converts) the three-dimensional (3D) raw magnetic signals into the 3D positional (xyz) and the two-dimensional (2D) orientational (pitch and roll) information, with a pre-trained deep neural network (DNN; Sebkhi et al., 2021). Therefore, MagTrack returns raw magnetic, positional, and orientational signals. The coordinate origin is a fixed point located on the frame of the glasses. The software also provides a user interface for real-time visualization of tongue movement.

Compared to commercial EMA devices, MagTrack has key advantages of being lightweight and wearable. In addition, MagTrack is easy to set up with the procedures of software installation and connecting hardware to a computer with a USB cable (see Figure 2b). Unlike AG series and Wave, MagTrack has no extra hardware for data capture and processing (data are preprocessed in a chip in the glasses and are localized by software on a computer). Despite these advantages, MagTrack has demonstrated a lower localization spatial accuracy (1.6–2.4 mm)

compared to EMA devices (about 0.5 mm; Berry, 2011; Sebkhi et al., 2021; Yunusova et al., 2009). Thus, whether tongue motion information captured by MagTrack is sufficient for SSI application is undetermined.

In this study, we performed SSR experiments to validate the use of MagTrack for SSI. Before the SSR experiments, we directly compared the articulatory data collected from MagTrack with a commercial EMA (NDI Wave). Then two experiments were conducted: isolated vowel recognition (consonant–vowel–consonant [CVC] classifications on eight English vowels) and continuous SSR. The vowel recognition used support vector machines (SVMs; Cortes & Vapnik, 1995) as the classifiers with data collected from two male subjects using MagTrack. The experimental results were measured by classification accuracies and compared with our previous similar work done by J. Wang, Samal, et al. (2016). The continuous SSR experiments used data collected from two different male subjects using MagTrack. Two conventional hybrid speech recognizers, the hidden Markov model (HMM)–Gaussian mixture model (GMM) and HMM–DNN, were used (Juang & Rabiner, 1991) and measured by phoneme error rates (PERs). In addition, one of the subjects in the current MagTrack experiment had also participated in our prior commercial EMA (NDI Wave) study, which used the same speech stimuli. A comparison between MagTrack with the commercial EMA data from the same subject was performed in the continuous speech recognition experiment.

Method

Data Collection

Participants and Stimuli

Four of our male researchers participated in this study as subjects. All MagTrack data collection was conducted at the participants' homes during the COVID-19 pandemic while the participating university campuses were closed. They were instructed to speak at their normal pace and loudness. Two of the participants (Subjects 1 and 2, age: 35 and 25 years) collected isolated vowel data (with MagTrack), and another two researchers (Subjects 3 and 4, age: 43 and 33 years) participated in the continuous speech data collection (with MagTrack). As mentioned, one of them (Subject 4) also collected EMA data in the lab at the University of Texas at Dallas (J. Wang's previous institution) before COVID-19. No history of prior speech, language, hearing, or cognition difficulty was self-reported from any of the participants. This study was approved by the institutional review boards at The University of Texas at Austin, the Georgia Institute of Technology, and the University of Texas at Dallas.

For the isolated vowel data, eight English vowels in CVC syllables (/bab/, /bib/, /beb/, /bæb/, /bʌb/, /bɒb/, /bob/, and /bub/) were used as vowel stimuli. These eight selected syllables were utilized in previous works (J. Wang et al., 2013, 2015; J. Wang, Samal, et al., 2016). We call them “isolated vowels” (rather than isolated CVCs) because they have the same consonant context and also for simplicity. By utilizing the same stimuli, we are able to evaluate the MagTrack in comparison to commercial EMA devices.

For continuous speech recognition, two phrase lists were used as the stimulus. One phrase list includes a total of 432 phrases. The first 132 phrases in the list were selected from phrases that are frequently spoken by the users of augmentative and alternative communication devices (Beukelman & Mirinda, 1998). Then 300 additional phrases were added, which included sentences frequently used in daily communication. The other stimuli were a phoneme-balanced list of 400 sentences developed by Kalikow et al. (1977).

Procedure

Two simple steps were needed to set up MagTrack before data collection. We first installed the MagTrack software on a laptop computer running Microsoft Windows 11 operating system. Then, we connected the MagTrack device to the computer with a USB cable (see Figure 2b). After that, the sensor was attached to the tongue tip with dental glue (Peri-Acryl 90, GluStitch). Our prior work showed that a single sensor on the tongue tip and two sensors on the upper and lower lips could be sufficient for SSI to produce intelligible speech (Cao et al., 2019). Therefore, we leverage the use of one sensor on the tongue tip (1 cm from the apex) for SSI application in this study. As introduced, MagTrack returns the captured raw magnetic data (3D), positional data (3D), and orientational data (2D). The 3D positional data include left–right (x), superior–inferior (y), and anterior–posterior (z). Tongue motion and speech audio were recorded synchronously, which was controlled by the software provided with the device (Sebkhi et al., 2021). The sampling rate of articulatory data was 250 Hz.

The commercial EMA device used in this study was the NDI Wave system (see Figure 1b). The articulatory movement and the speech audio were recorded synchronously. The 3D positional (xyz) and 3D quaternion (representing roll and pitch) signals were captured by each of the sensors (sampling rate = 100 Hz). Consistent with MagTrack, 3D positional data include left–right (x), superior–inferior (z), and anterior–posterior (y) dimensions. Four sensors were attached to the tongue tip (5–10 mm to tongue apex), tongue back (20–30 mm back from tongue tip), middle of the upper lip, and lower lip. The tongue sensors were attached with the same dental

glue (Peri-Acryl 90, GluStitch), and the lip sensors were attached with tape. Only data from the tongue tip were used in this study for a comparison with the MagTrack. We used both positional and orientational information of tongue tip motion captured by the commercial EMA (NDI Wave) in these experiments. Orientational information has been demonstrated helpful in SSI application (Cao et al., 2018).

The CVC data collection was completed over multiple sessions, in which the subjects took breaks between each session (20 recordings per session, eight CVC samples per recording). Both Subjects 1 and 2 completed data collection in 2 days. Subject 1 recorded three and two sessions on Day 1 and Day 2, respectively (100 recordings, 800 samples in total). Subject 2 recorded two sessions on both Days 1 and 2 (80 recordings, 640 samples in total). In continuous speech data collection, Subject 3 recorded 300 phrases from the phonetic-balanced phrase lists (Kalikow et al., 1977) with the repetitions of some phrases, which provided 6,539 phonemes in total. Subject 4 recorded 432 phrases with MagTrack using the same stimuli with the commercial EMA (8,824 phoneme samples in 432 phrases). There are 39 unique phonemes (from the Carnegie Mellon University pronouncing dictionary), and silences were indicated at the beginning and end of each phrase in the transcriptions.

Data Analysis

We performed a direct comparison experiment of tongue motion trajectories of commercial EMA and MagTrack and two SSR experiments (vowel classification and continuous speech recognition). Both vowel classification and continuous speech recognition experiments were speaker dependent, in which the training, validation, and testing sets are from the same speakers.

Direct Comparison of Trajectories

We first performed a direct comparison on the tongue tip trajectories of the commercial EMA and our

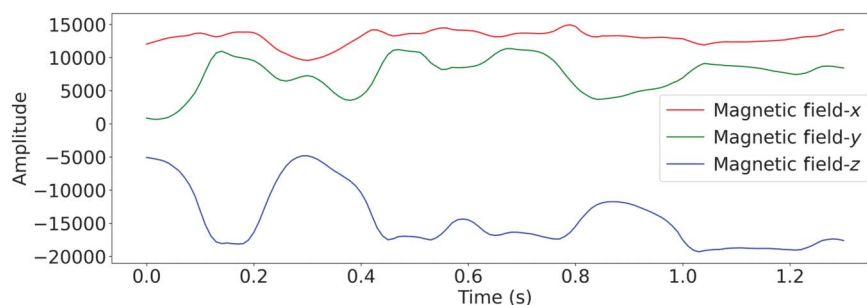
MagTrack. Here, only y and z dimensions (superior–inferior and anterior–posterior) were considered as they are more significant in speech production.

A nonlinear alignment technique, dynamic time warping (DTW; J. Wang et al., 2014), was used to align the data (same stimuli) in this experiment. First, the mean of each dimension was subtracted. Then, DTW was first applied to the parallel (same stimuli) EMA and MagTrack data samples from Subject 4. After DTW alignment, the Pearson correlations were computed on the two dimensions. A higher correlation value indicates higher similarity.

We did not use linear alignment in this experiment because the data are collected from two sessions and the starting points of these data samples are not consistent. A manual segmentation (to indicate the real starting point of each data sample) is needed, which is time-consuming for these data (more than 800 phrases). As indicated in recent literature (Wisler et al., 2022), nonmatched starting points may significantly affect the linear alignment results. Thus, we did perform manual adjustment of the starting points in the vowel classification experiment below. The manual adjustment of the starting points was not needed in the continuous SSR experiment because the algorithm will align them automatically.

Besides the localized signals (spatial coordinates), raw magnetic signals are also provided by MagTrack as mentioned. The example of raw magnetic signals is presented in Figure 3, which shows the raw output of the magnetometer for the following utterance: “I don’t understand.” There is one magnetic output per axis, and this output is represented as a 16-bit value that measures the magnetic intensity. These magnetic measurements are fed into a machine learning model that predicts the spatial position and orientation of the sensor (x , y , z coordinates and orientations). More details about the magnetic localization can be found in Sebkhii et al. (2021). As raw magnetic data are not provided in commercial EMAs, we did not perform the comparison of the magnetic signals from MagTrack and the commercial EMA. The two experiments

Figure 3. An example of raw magnetic signals when the subject was producing “I don’t understand.”



below were performed with and without using raw magnetic signals in MagTrack data as input to determine the usefulness of the raw magnetic data when combined with positional and orientational data.

Vowel Classification

In this experiment, we performed similar CVC classification experiments with SVM produced by J. Wang, Samal, et al. (2016). SVM (Cortes & Vapnik, 1995) is a supervised machine learning algorithm that could be used for classification and regression. SVM classifiers were more commonly used, which learn by finding hyperplanes that best separated the data space. The support vectors are the data points that are closest to the hyperplanes. By employing kernel methods (Hofmann et al., 2008), SVM can also perform a nonlinear separation. As a non-deep learning algorithm, SVM has been demonstrated to be powerful, especially in some cases that are not suitable for deep learning models (e.g., insufficient training data). DNNs were explored in the preliminary experiments but demonstrated lower performance, possibly due to the small size of data; thus, they were not used in this vowel classification experiment.

We first used only the 2D positional signals (y and z without left-right direction) as that in the study of J. Wang, Samal, et al. (2016) for comparison, then added x -direction position, orientation, and magnetic signals to explore the highest performance. The collected CVC MagTrack data were manually parsed into clips of whole CVC syllables. We followed the procedure by J. Wang, Kothalkar, et al. (2016) to down-sample each of the CVC clips to 10 frames to reduce feature dimension. Then, the concatenation of the 10 data points from each dimension (possibly include xyz position, roll, pitch, magnetic, and their combinations) was used as the input of SVM (the maximum dimension is [3D position +2D orientation +3D magnetic] \times 10 data points = 80D) for eight-class classification (/bab/, /bib/, /beb/, /bæb/, /bAb/, /bOb/, /bob/, and /bub/). The performance was measured by the classification accuracies, which were computed by the number of correctly classified samples divided by the total number of samples, where a sample is a production of CVC. Five-fold cross-validation experiments were performed on each of the two subjects (Subjects 1 and 2). The whole data were first partitioned into five folds/parts. In each fold, one fifth of the data were used as testing data, and the rest were used as training data. The averaged accuracies over the five-fold were reported as the final performance. We used the Sklearn toolkit in Python for the SVM implementation in this experiment.

Continuous SSR

Continuous SSR is to recognize phoneme (or word) sequences from articulatory speech data. In this study, we

performed phoneme-level recognition since the phoneme sequence output is more convenient for the following text-to-speech stage in SSIs (Cao et al., 2021). All the sentences were first transcribed to phoneme sequences (39 unique phonemes and silence) based on the Carnegie Mellon University pronouncing dictionary. We implemented HMM-based automatic speech recognition models (Juang & Rabiner, 1991; Kim et al., 2017), with two different speech recognizers: GMM-HMM and DNN-HMM. The speech recognizers adopted HMM to model the temporal variations during speech such as varying durations of the same phonemes. A single phoneme would be represented by a three-state left-to-right HMM (begin-middle-end subphones); the time variation of this phoneme would be modeled by intra- or interstate transition (stay in the same state or go to the next state). The GMM and DNN were used for modeling the probability distribution of the observed articulatory signals (MagTrack or commercial EMA data frames) given the current phonemes.

For the MagTrack data, the input to the speech recognition models includes positional and orientational signals with or without raw magnetic signals included. The orientational information of MagTrack includes the returned roll and pitch signals. For the commercial EMA (NDI Wave), the orientational information also includes roll and pitch but was represented by the 3D quaternion. The raw magnetic signals returned by MagTrack were validated by comparing the experimental results with and without using them as input. All MagTrack signals were down-sampled to 100 Hz from 250 Hz, which were the same to the commercial EMA signals. The performance was measured by the PERs, which were computed by the sum of insertion, deletion, and substitution errors divided by the number of the phonemes tested. The insertion recognition errors are inserting phonemes that do not exist in the ground truth phoneme sequences. The deletion errors are missing some phonemes. Substitution errors occur when phonemes are mistakenly recognized as different ones. PER could be larger than 100%, if there are too many insertion errors. In order to leave sufficient testing (and training) data for satisfying phoneme distributions in each of the cross-validations, for Subject 3, we performed five-fold cross-validation experiments on the recorded 300 phrases (6,539 phonemes); each validation used 240 and 60 phrases for training and testing, respectively. For Subject 4 (432 MagTrack and commercial EMA phrases, 8,824 phonemes), we performed eight-fold cross-validation, in which for each validation, the models were trained and validated on 378 phrases and tested with the remaining 54 phrases. The Kaldi speech recognition toolkit was used for this experiment (Povey et al., 2011). More technique details are provided in the Appendix.

Statistical Analysis

One-way analysis of variance (ANOVA) tests and pairwise two-tailed t tests were performed to compare the PERs in all cross-validations in the continuous SSR experiments. The main comparisons include (a) MagTrack with versus without using magnetic signals and (b) MagTrack with magnetic signals versus commercial EMA. A p value less than .05 is required to be considered as significantly different.

Results

Direct Comparison of Trajectories

In the direct comparison experiment, the Pearson correlations between the data collected using our MagTrack and the commercial EMA after alignment with DTW were .72 and .63 for the superior–inferior (y) and anterior–posterior (z) dimensions, respectively. Figure 4 gives examples of the tongue tip movement trajectories (MagTrack and commercial EMA) of Subject 4 when producing “I don’t understand.” Figures 4a and 4b are the original tongue tip motion trajectories in y and z dimensions from the MagTrack and commercial EMA (NDI Wave), respectively. The tongue tip trajectory patterns from the two devices look similar, although the starting points are different (possibly due to session difference, as mentioned in the Method section). Figures 4c and 4d are DTW aligned trajectories for y and z dimensions, respectively, which visually demonstrate the high similarity of the tongue tip trajectories after the DTW alignment of data in Figures 4a and 4b, respectively.

Vowel Classification

Figure 5 shows the accuracies of vowel classification of Subjects 1 and 2. The average accuracy is 78.46% when using 2D (yz) positional information only, as in our previous work (J. Wang, Samal, et al., 2016), which was lower than the average accuracy of 81.6% achieved by commercial EMA in J. Wang, Samal, et al. (2016). When the x -dimension was used, the accuracy was increased to 83.98%. After that, as the orientational and magnetic information was added, the accuracies were improved to 88.07% and 89.74%, which are higher than the commercial EMA results in the study of J. Wang, Samal, et al. (2016).

Continuous SSR

Figure 6 shows the PERs of continuous SSR using MagTrack data collected from Subjects 3 and 4, where two models, GMM and DNN, were used with or without magnetic signals. Unlike accuracy, a lower PER indicates a higher performance. One-way ANOVA test results

showed that there was a statistical significance between at least two groups for both Subject 3, $F(2, 2) = 3.33$, $p = .046^*$, and Subject 4, $F(2, 2) = 3.81$, $p = .021^*$ (significant results are marked with an asterisk). Table 1 provides the two-tailed t -test results for selected comparisons under different experimental setups (using GMM or DNN with or without using magnetic data) from the same subjects. Generally, including magnetic signals (orange bars in Figure 6) showed higher performance (lower PERs) than without using magnetic signals (blue bars) except Subject 3 using GMM. The improvement of using magnetic signals in GMM for Subject 4 is not significant (see Table 1). When comparing the two models, DNN outperformed GMM either with or without using magnetic signals, except for Subject 3 without magnetic signals. For both subjects, the best results were achieved using DNN with magnetic signals.

Figure 7 shows the PERs of continuous speech recognition using GMM and DNN with MagTrack (x , y , z coordinates; orientation; and magnetic signals) and commercial EMA data (x , y , z coordinates and orientation signals only, as magnetic data were not available) from the same participant (Subject 4). One-way ANOVA test showed there was a statistical significance between at least two groups, $F(2, 2) = 8.55$, $p = .0003^*$. The two-tailed t -test results for selected comparisons under different experimental setups (using GMM or DNN with MagTrack or commercial EMA data) are provided in Table 2. The PERs from the commercial EMA were generally lower than those from MagTrack, although the difference was not statistically significant when using GMM (see Table 2). When comparing the two models, DNN statistically outperformed GMM.

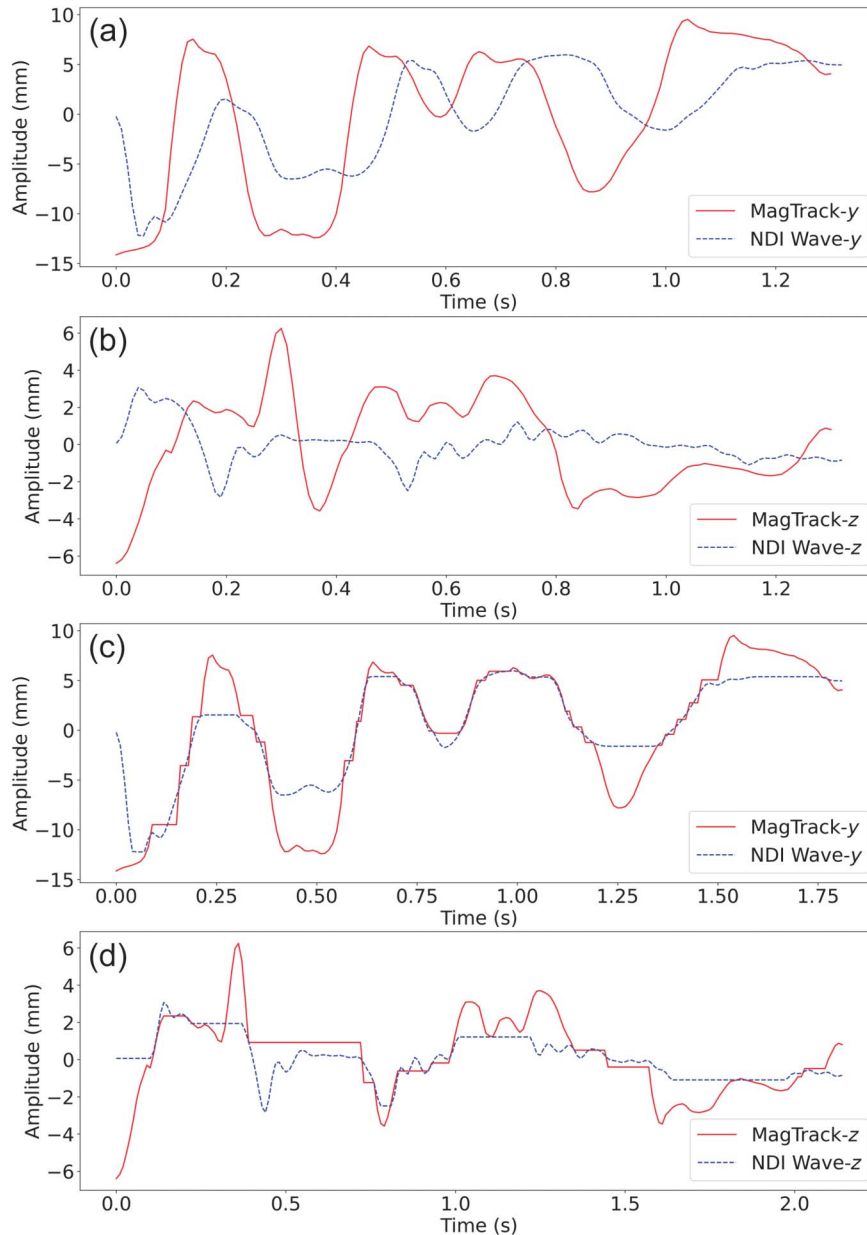
Table 3 gives the percentages of different types of errors (i.e., substitution, deletion, and insertion errors) of the DNN–HMM experiment on Subject 4. Substitution and deletion errors dominated the errors consistently across different experimental setups (using different input data).

Table 4 lists the occurrences of the 10 most frequent specific substitution, deletion, and insertion errors of the DNN–HMM experiment on Subject 4. The most common substitution error observed was the voiced stop /d/ being substituted by the voiceless stop /t/ across various speech recording setups. The most common deletion error was the deletion of silence, which was also consistent across all setups. The most common insertion error was /ih/ when using EMA data and MagTrack data with magnetic signals. Instead, the most insertion error was /ah/ when using MagTrack data without magnetic signals.

Discussion

The direct comparison experiment clearly demonstrated the similarity of the tongue tip trajectories captured

Figure 4. Examples of tongue tip motion trajectories, original and after DTW alignment, captured by MagTrack and NDI Wave from Subject 4 when producing “I don’t understand.” (a) Example of superior–inferior (y) dimension of original tongue tip motion trajectories captured by MagTrack and NDI Wave. (b) Example of anterior–posterior (z) dimension of original tongue tip motion trajectories captured by MagTrack and NDI Wave. (c) Example of superior–inferior (y) dimension of DTW-aligned tongue tip motion trajectories captured by MagTrack and NDI Wave. (d) Example of anterior–posterior (z) dimension of DTW-aligned tongue tip motion trajectories captured by MagTrack and NDI Wave. DTW = dynamic time warping.

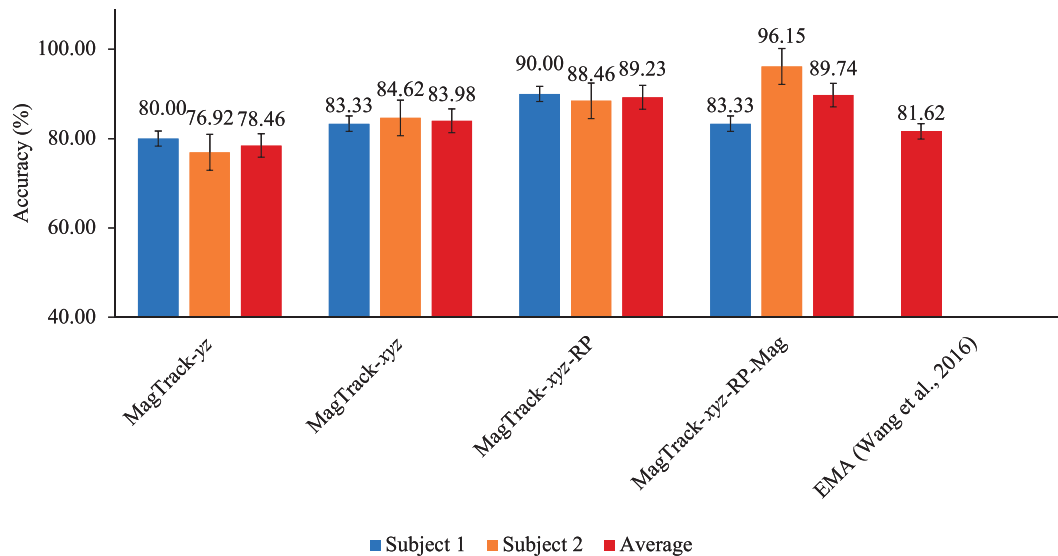


by the two devices via moderate-to-strong Pearson correlations (after DTW alignment). In addition, Figure 4 visually shows the similarity of the trajectories (before and after aligned). These results suggested our MagTrack is suitable for general speech articulation studies.

For vowel recognition tasks, MagTrack positional data have demonstrated slightly lower performance than

previous similar commercial EMA under the same data setup (use y and z position only; J. Wang, Samal, et al., 2016) due to the lower localization accuracies. As the other information was added (x position, orientation, and magnetic), the accuracies could be higher than commercial EMA. Although the experimental setups were similar (e.g., articulatory information, data amount, and classification model), only two researchers participated in the

Figure 5. Accuracies using MagTrack data averaged from the two subjects in the isolated vowel recognition experiment, compared with the accuracy using EMA in the study of J. Wang, Samal, et al. (2016). Error bars indicate standard errors. MagTrack-*yz* = using *yz* (2D) positional data; MagTrack-*xyz* = using 3D positional data; MagTrack-*xyz*-RP = using 3D positional + roll and pitch; MagTrack-*xyz*-RP-Mag = using raw magnetic data long with 3D positional and roll-pitch orientational data; EMA = electromagnetic articulograph.



current experiments, whereas the commercial EMA results were from 13 subjects. These discrepancies may also affect the comparison results.

In addition, it is surprising that adding *x*-axis improved performance here as the literature suggested that *x*-dimension (lateral movement) is not significant at least in typical speech production (Beautemps et al., 2001;

Westbury, 1994). One possible explanation may be that machine learning is able to find subtle pattern differences that these conventional approaches could not detect. Our team found *x*-dimension could be significant in dysarthric speech due to amyotrophic lateral sclerosis in a preliminary study (unpublished). However, these observations need further validation with larger data sets from more subjects, which will be conducted in the future.

Figure 6. PERs of Subjects 3 and 4 in continuous silent speech recognition experiments using GMM-HMM and DNN-HMM speech recognizers. Error bars indicate standard errors. Significant differences ($p < .05$) are marked with “*”; nonsignificant difference is marked with “n.s.” PER = phoneme error rate; GMM = Gaussian mixture model; HMM = hidden Markov model; DNN = deep neural network; MagTrack-*xyz*-RP indicates using 3D positional + roll and pitch data; MagTrack-*xyz*-RP-Mag indicates using raw magnetic data along with 3D positional and roll-pitch orientational data.

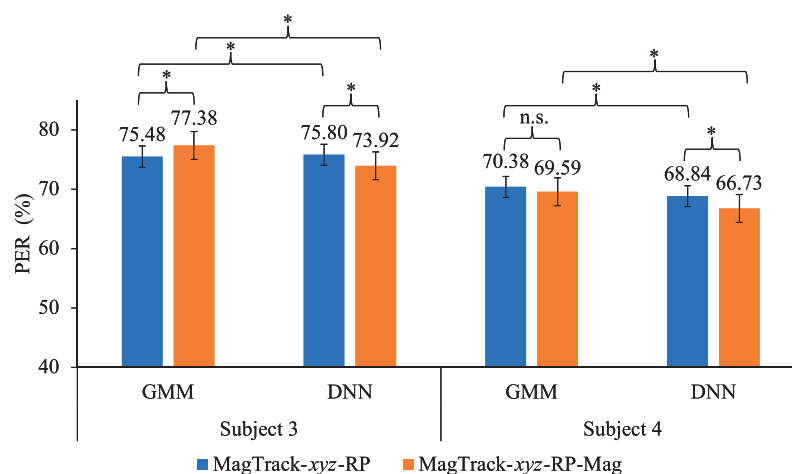


Table 1. The results (t values and p values) of two-tailed t tests on Subjects 3 and 4 under different experimental setups using MagTrack data.

Comparisons	Subject 3	Subject 4
With Mag vs. without Mag (GMM)	$t(8) = 5.37, p = .006^*$ $d = 1.03$	$t(14) = -1.87, p = .1$ $d = 0.39$
With Mag vs. without Mag (DNN)	$t(8) = -3.75, p = .02^*$ $d = 1.16$	$t(14) = -4.71, p = .002^*$ $d = 0.84$
GMM vs. DNN (with Mag)	$t(8) = 7.35, p = .002^*$ $d = 2.09$	$t(14) = 6.06, p = .0005^*$ $d = 1.25$
GMM vs. DNN (without Mag)	$t(8) = -3.75, p = .02^*$ $d = 0.18$	$t(14) = 2.78, p = .03^*$ $d = 0.68$

Note. Subject 3 has five samples (cross-validations) from each of the two groups ($df = 8$). Subject 4 has eight samples (cross-validations) from each of the two groups ($df = 14$). Significant results are marked with an asterisk. With Mag = with magnetic signals included; d = Cohen's d effect size; GMM = Gaussian mixture model; DNN = deep neural network.

For continuous SSR, the PERs are high compared to those from audio speech recognition, which is not surprising. Audio speech recognition could achieve lower than 10% PER because of rich information in the acoustics and larger data size (Baevski et al., 2020). Particularly, SSR lacks acoustic information that helps distinguish voiced and unvoiced phonemes. As indicated in Table 3, deletion and substitution errors dominated the errors, which was likely due to lack of phonation information. Specifically, as shown in Table 4, the most substitution errors occurred in consonant cognates (voiced and voiceless consonant pairs, e.g., /d/ and /t/). Deletion of the silence also contributed a significant portion of the recognition errors likely for the same reason. These errors are expected to be reduced at the word- or subword-level recognition since more contextual information can be

embedded in the recognition tokens (words or subwords). Future studies will explore word- or subword-level recognition. Another reason for the high PER of SSR is the data size and the data coverage. The data size in this study is relatively small, and there is tongue tip motion only, which is insufficient to distinguish some of these phonemes that have significant tongue dorsum or back movement. Recasens (2002) suggested that the tongue tip and tongue dorsum act more independently for more anterior consonantal productions. Our previous study also found tongue tip and tongue back are an actually optimal combination for speech classification (J. Wang, Samal, et al., 2016). However, as a starting point to test this new device, we think these PERs are actually encouraging. Future studies will be explored to improve PERs by, for example, adding more sensors and using a larger data set.

Figure 7. PERs using MagTrack and commercial EMA data from the same participant (Subject 4) in continuous silent speech recognition experiments using GMM–HMM and DNN–HMM speech recognizers. Error bars indicate standard errors. Significant differences ($p < .05$) are marked with “*”; nonsignificant difference is marked with *ns*. PER = phoneme error rate; GMM = Gaussian mixture model; HMM = hidden Markov model; DNN = deep neural network.

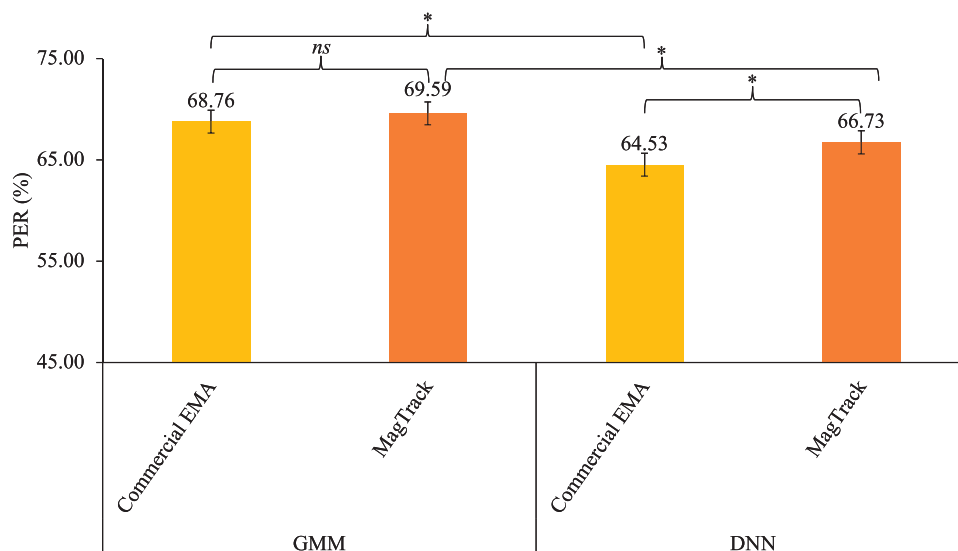


Table 2. The results of two-tailed *t* tests comparing MagTrack and commercial electromagnetic articulograph (EMA) data from Subject 4.

Comparisons	Subject 4
MagTrack vs. commercial EMA (GMM)	$t(14) = 1.33, p = .22$ $d = 0.44$
MagTrack vs. commercial EMA (DNN)	$t(14) = 7.30, p = .0002^*$ $d = 0.89$
GMM vs. DNN (MagTrack)	$t(14) = 6.06, p = .0005^*$ $d = 1.25$
GMM vs. DNN (commercial EMA)	$t(14) = 7.01, p = .0002^*$ $d = 2.04$

Note. Subject 4 has eight samples (cross-validations) from each of the two groups ($df = 14$). Significant results are marked with an asterisk. d = Cohen's d effect size; GMM = Gaussian mixture model; DNN = deep neural network.

In general, we think the results of continuous speech recognition experiment on the data from both of the subjects are promising, although Subject 4 performed better than Subject 3 (see Figure 6). This is possibly due to the smaller data set from Subject 3 (300 phrases) compared to that from Subject 4 (432 phrases). In addition, different stimuli were used. The 432-phrase list used by Subject 4 was chosen from daily used sentences, whereas the phrase list used by Subject 3 was more phoneme balanced, which led to more triphones (720 for Subject 3 compared to 128 for Subject 4; see the details in Table A1 in the Appendix).

When comparing the commercial EMA and MagTrack (with magnetic signal included), the commercial EMA outperformed MagTrack by 0.83% and 2.2% in GMM and DNN, respectively (see Figure 7). The performance differences (between commercial EMA and MagTrack) were not significant in GMM but significant in DNN (see Table 2). As introduced, MagTrack demonstrated a lower spatial tracking accuracy (1.6–2.4 mm) than the commercial EMA devices used (about 0.5–1.37 mm). Therefore, it is not surprising that MagTrack achieved lower SSR performance than the commercial EMA. Even with the improvement brought by adding the raw magnetic data of MagTrack, the performances were still not as good as the commercial EMA using DNN (see Figure 7). Overall, DNN obtained the best results with the commercial EMA. However, the results obtained from

MagTrack are not significantly lower than that with the commercial EMA using GMM, which is encouraging.

In summary, based on our preliminary results above and the advantages of wearability, we believe MagTrack has the potential of being used as the frontend of SSIs. In addition, we are actively improving the MagTrack. First, we have already added two additional sensors to support upper and lower lip tracking (unpublished). Second, the wireless connection between the glasses and the data recording computer (laptop) is under development. Third, the tracking accuracy of MagTrack is under ongoing improvement and can be improved in terms of tracking accuracy. The current localization (tracking) model that maps raw magnetic to positional signals is a DNN, which could be improved by replacing it with more advanced models such as recurrent neural network and convolutional neural network. Other ongoing improvements include further reducing the weight of the glasses and making the eyeglasses more comfortable.

Although this study focused on the validation of the potential of applying MagTrack in SSI, MagTrack has plenty of other potential applications as commercial EMA does and could be extended since it is wearable. For tongue-controlling rehabilitation applications such as tongue-controlling wheelchairs (Sebkhil et al., 2022), the wearable characteristic of MagTrack would be important. Other speech applications include basic and clinical speech kinematic studies (Gafos & van Lieshout, 2020; Recasens, 2002), visualization-based speech therapies (Katz et al., 2014), and second language learning (Li et al., 2019). Recasens (2002) collected EMA data during speech production and found that the tongue tip and tongue dorsum act more independently for more anterior consonantal productions. Katz et al. (2014) built a customized interface with EMA that allows users to view their current tongue position during speech training. Li et al. (2019) suggested that visual biofeedback can facilitate speech production training in clinical populations and second language learners in their study. In addition, our previous studies indicated that articulatory data could improve dysarthric speech recognition when articulation information was added on top of acoustic input (Hahm et al., 2015; Kim

Table 3. Percentages of different types of errors (substitution, deletion, and insertion) of the deep neural network–hidden Markov model experiment on Subject 4.

Experiment	PER (%)	Substitution (%)	Deletion (%)	Insertion (%)
EMA-xyz-RP	64.53	26.24	36.53	1.76
MagTrack-xyz-RP	68.84	28.50	38.23	2.11
MagTrack-xyz-RP-Mag	66.73	28.71	35.68	2.34
Average	66.70	27.82	36.81	2.07

Note. PER = phoneme error rate; EMA = electromagnetic articulograph; RP = roll and pitch.

Table 4. The 10 most frequent specific substitution, deletion, and insertion errors of the deep neural network–hidden Markov model experiment on Subject 4.

EMA-xyz-RP						MagTrack-xyz-RP						MagTrack-xyz-RP-Mag					
Sub	Num	Ins	Num	Del	Num	Sub	Num	Ins	Num	Del	Num	Sub	Num	Ins	Num	Del	Num
d → t	49	ih	13	sil	425	d → t	39	ah	16	sil	384	d → t	51	ih	17	sil	380
t → n	32	ah	12	ah	235	t → d	27	n	15	ah	237	t → d	34	iy	14	t	216
d → n	26	n	11	ih	200	n → d	18	ih	14	n	228	t → n	23	s	13	n	194
t → d	25	b	7	n	195	ah → ae	17	ae	12	t	224	ah → ih	22	ah	11	ah	193
ah → ih	21	dh	6	t	185	z → s	16	t	9	ih	170	s → z	21	uw	10	ih	155
ao → aa	19	f	5	d	147	ih → iy	14	d	8	d	150	z → s	19	m	9	d	136
r → er	17	w	4	iy	128	ah → uw	13	hh	7	m	122	ao → aa	18	ae	8	k	115
z → s	15	aa	3	m	119	t → dh	12	sil	5	k	111	ah → ow	17	z	7	m	108
ah → aa	14	ch	2	k	110	t → n	11	ey	4	ay	106	ih → ah	16	sh	6	s	104
iy → ih	13	g	1	eh	106	ah → ow	10	f	3	ae	100	n → iy	15	r	5	eh	102

Note. After top 10, the substitution and insertion errors happened mostly only once. EMA-xyz-RP means using the positional (x, y, z coordinates) and orientational data (roll and pitch) of EMA data. Mag = adding magnetic data to positional data and rotational data in the experiment; MagTrack = the data are from our device, MagTrack; Sub = substitution errors; Ins = insertion errors; Del = deletion (missing) errors; sil = silence; EMA = electromagnetic articulograph.

et al., 2019). With the wearable MagTrack, articulation information can be conveniently provided as a supplementary information source for dysarthric speech recognition.

Data Availability Statement

The data set is currently not publicly available, but it is planned to be ready for distribution in the future. Further enquiries can be directed to the last author. There is no licensed patent or start-up company associated with this technology or device at this moment.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders under Award R01DC016621 and its Supplemental Award R01DC016621-03S that were awarded to the last author. The authors would like to thank Ted Mau, Robin Samlan, and Kristin Teplansky for their support and assistance. Partial data and results in this submission were presented at the Motor Speech Conference in Charleston, South Carolina, February 2022.

References

- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.
- Bailey, B. J., Johnson, J. T., & Newlands, S. D. (2006). *Head & neck surgery—Otolaryngology* (4th ed.). Lippincott Williams & Wilkins.
- Beautemps, D., Badin, P., & Bailly, G. (2001). Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *The Journal of the Acoustical Society of America*, *109*(5), 2165–2180. <https://doi.org/10.1121/1.1361090>
- Benway, N. R., Hitchcock, E. R., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /s/ errors in American English: A single-case randomization design. *American Journal of Speech-Language Pathology*, *30*(4), 1819–1845. https://doi.org/10.1044/2021_AJSLP-20-00216
- Berry, J. J. (2011). Accuracy of the NDI wave speech research system. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1295–1301. [https://doi.org/10.1044/1092-4388\(2011\)10-0226](https://doi.org/10.1044/1092-4388(2011)10-0226)
- Beukelman, D. R., & Mirenda, P. (1998). *Augmentative and alternative communication: Management of severe communication disorders in children and adults* (2nd ed.). Brooks.
- Cao, B., Kim, M., Wang, J. R., van Santen, J., Mau, T., & Wang, J. (2018). Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information. *Proceedings of Interspeech*, *2018*, 3152–3156. <https://doi.org/10.21437/Interspeech.2018-2484>
- Cao, B., Sebki, N., Bhavsar, A., Inan, O. T., Samlan, R., Mau, T., & Wang, J. (2021). Investigating speech reconstruction for laryngectomees for silent speech interfaces. *Proceedings of Interspeech*, *2021*, 651–655. <https://doi.org/10.21437/Interspeech.2021-1842>
- Cao, B., Tsang, B. Y., & Wang, J. (2019). *Comparing the performance of individual articulatory flesh points for articulation-to-speech synthesis* [Paper presentation]. International Congress of Phonetic Sciences, Melbourne, Australia.
- Cao, B., Wisler, A., & Wang, J. (2022). Speaker adaptation on articulation and acoustics for articulation-to-speech synthesis. *Sensors*, *22*(16), 6056. <https://doi.org/10.3390/s22166056>
- Chen, H.-C., Tang, Y.-B., & Chang, M.-H. (2001). Reconstruction of the voice after laryngectomy. *Clinics in Plastic Surgery*, *28*(2), 389–402. [https://doi.org/10.1016/S0094-1298\(20\)32374-9](https://doi.org/10.1016/S0094-1298(20)32374-9)
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, *52*(4), 270–287. <https://doi.org/10.1016/j.specom.2009.08.002>
- Denny, M. (2020). *NDI company update—June 2020*. NDI. Retrieved September 5, 2022, from <https://www.ndigital.com/ndi-company-update-june-2020/>
- Diener, L., Umesh, T., & Schultz, T. (2019). Improving fundamental frequency generation in EMG-to-speech conversion using a quantization approach. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, *2019*, 682–689. <https://doi.org/10.1109/ASRU46091.2019.9003804>
- Eadie, T. L., Otero, D., Cox, S., Johnson, J., Baylor, C. R., Yorkston, K. M., & Doyle, P. C. (2016). The relationship between communicative participation and post-laryngectomy speech outcomes. *Head & Neck*, *38*(Suppl. 1), E1955–E1961. <https://doi.org/10.1002/hed.24353>
- Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., & Chapman, P. M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics*, *30*(4), 419–425. <https://doi.org/10.1016/j.medengphy.2007.05.003>
- Gafos, A., & van Lieshout, P. (2020). Editorial: Models and theories of speech production. *Frontiers in Psychology*, *11*, 1238. <https://doi.org/10.3389/fpsyg.2020.01238>
- Gonzalez, J. A., Cheah, L. A., Gomez, A. M., Green, P. D., Gilbert, J. M., Ell, S. R., Moore, R. K., & Holdsworth, E. (2017). Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2362–2374. <https://doi.org/10.1109/TASLP.2017.2757263>
- Gonzalez, J. A., & Green, P. D. (2018). A real-time silent speech system for voice restoration after total laryngectomy. *Revista de Logopedia, Foniatria y Audiología*, *38*(4), 148–154. <https://doi.org/10.1016/j.rlfa.2018.07.004>
- Gonzalez-Lopez, J. A., Gomez-Alanis, A., Martín Doñas, J. M., Pérez-Córdoba, J. L., & Gomez, A. M. (2020). Silent speech interfaces for speech restoration: A review. *IEEE Access*, *8*, 177995–178021. <https://doi.org/10.1109/ACCESS.2020.3026579>
- Gonzalez-Lopez, J. A., Gomez-Alanis, A., Pérez-Córdoba, J. L., & Green, P. D. (2022). Non-parallel articulatory-to-acoustic conversion using multiview-based time warping. *Applied Sciences*, *12*(3), 1167. <https://doi.org/10.3390/app12031167>
- Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorst, L., Zinman, L., & Berry, J. D. (2013). Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal*

- Degeneration*, 14(7–8), 494–500. <https://doi.org/10.3109/21678421.2013.817585>
- Hahm, S., Heitzman, D., & Wang, J.** (2015). Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 47–54. <https://doi.org/10.18653/v1/W15-5109>
- Hofmann, T., Schölkopf, B., & Smola, A. J.** (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3), 1171–1220. <https://doi.org/10.1214/009053607000000677>
- Juang, B. H., & Rabiner, L. R.** (1991). Hidden Markov models for speech recognition. *Technometrics*, 33(3), 251–272. <https://doi.org/10.1080/00401706.1991.10484833>
- Kalikow, D. N., Stevens, K. N., & Elliott, L. L.** (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5), 1337–1351. <https://doi.org/10.1121/1.381436>
- Katz, W. F., Campbell, T. F., Wang, J., Farrar, E., Eubanks, J. C., Balasubramanian, A., Prabhakaran, B., & Rennaker, R.** (2014). Opti-speech: A real-time, 3d visual feedback system for speech training. *Proceedings of Interspeech, 2014*, 1174–1178. <https://doi.org/10.21437/Interspeech.2014-298>
- Katz, W. F., & Mehta, S.** (2015). Visual feedback of tongue movement for novel speech sound learning. *Frontiers in Human Neuroscience*, 9. <https://doi.org/10.3389/fnhum.2015.00612>
- Kaye, R., Tang, C. G., & Sinclair, C. F.** (2017). The electrolarynx: Voice restoration after total laryngectomy. *Medical Devices: Evidence and Research*, 10, 133–140. <https://doi.org/10.2147/MDER.S133225>
- Kim, M., Cao, B., Mau, T., & Wang, J.** (2017). Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), 2323–2336. <https://doi.org/10.1109/TASLP.2017.2758999>
- Kim, M., Cao, B., & Wang, J.** (2019). Multi-view representation learning via canonical correlation analysis for Dysarthric speech recognition. In K. Deng, Z. Yu, S. Patnaik, & J. Wang (Eds.), *Recent developments in mechatronics and intelligent robotics* (pp. 1085–1095). Springer. https://doi.org/10.1007/978-3-030-00214-5_133
- Kimura, N., Su, Z., & Saeki, T.** (2020). End-to-end deep learning speech recognition model for silent speech challenge. *Proceedings of Interspeech, 2020*, 1025–1026.
- Li, J. J., Ayala, S., Harel, D., Shiller, D. M., & McAllister, T.** (2019). Individual predictors of response to biofeedback training for second-language production. *The Journal of the Acoustical Society of America*, 146(6), 4625–4643. <https://doi.org/10.1121/1.5139423>
- Meltzer, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., & Kline, J. C.** (2018). Development of sEMG sensors and algorithms for silent speech recognition. *Journal of Neural Engineering*, 15(4), 046031. <https://doi.org/10.1088/1741-2552/aac965>
- Nijdam, H. F., Annyas, A. A., Schutte, H. K., & Leever, H.** (1982). A new prosthesis for voice rehabilitation after laryngectomy. *Archives of Oto-Rhino-Laryngology*, 237(1), 27–33. <https://doi.org/10.1007/BF00453713>
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K.** (2011). The Kaldi speech recognition toolkit [Paper presentation]. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Waikoloa Village, Hawai'i. <http://infoscience.epfl.ch/record/192584>
- Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M.** (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, 12(1), 6. <https://doi.org/10.5334/labphon.237>
- Rebernik, T., Jacobi, J., Tiede, M., & Wieling, M.** (2021). Accuracy assessment of two electromagnetic articulographs: Northern Digital Inc. WAVE and Northern Digital Inc. VOX. *Journal of Speech, Language, and Hearing Research*, 64(7), 2637–2667. https://doi.org/10.1044/2021_JSLHR-20-00394
- Recasens, D.** (2002). An EMA study of VCV coarticulatory direction. *The Journal of the Acoustical Society of America*, 111(6), 2828–2841. <https://doi.org/10.1121/1.1479146>
- Savariaux, C., Badin, P., Samson, A., & Gerber, S.** (2017). A comparative study of the precision of Carstens and Northern Digital Instruments electromagnetic articulographs. *Journal of Speech, Language, and Hearing Research*, 60(2), 322–340. https://doi.org/10.1044/2016_JSLHR-S-15-0223
- Schultz, T., Wand, M., Hueber, T., Krusiński, D. J., Herff, C., & Brumberg, J. S.** (2017). Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), 2257–2271. <https://doi.org/10.1109/TASLP.2017.2752365>
- Sebki, 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. <https://doi.org/10.1109/JSEN.2020.3046469>
- Sebki, N., Bhavsar, A., Sahadat, M. N., Baldwin, J., Walling, E., Biniker, A., Hoefnagel, M., Tonuzi, G., Osborne, R., Anderson, D., & Inan, O.** (2022). Evaluation of a head-tongue controller for power wheelchair driving by people with quadriplegia. *IEEE Transactions on Biomedical Engineering*, 69(4), 1302–1309. <https://doi.org/10.1109/TBME.2021.3113291>
- Shandiz, A. H., Tóth, L., Gosztolya, G., Markó, A., Csapó, T. G.** (2021). Neural speaker embeddings for ultrasound-based silent speech interfaces. *Proceedings of Interspeech, 2021*, 1932–1936. <https://doi.org/10.21437/Interspeech.2021-1466>
- Taylor, P.** (2009). *Text-to-speech synthesis*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511816338>
- Teplansky, K. J., Wisler, A., Cao, B., Liang, W., Whited, C. W., Mau, T., & Wang, J.** (2020). Tongue and lip motion patterns in alaryngeal speech. *Proceedings of Interspeech, 2020*, 4576–4580.
- Wang, H., Roussel, P., & Denby, B.** (2021). Improving ultrasound-based multimodal speech recognition with predictive features from representation learning. *JASA Express Letters*, 1(1), 015205. <https://doi.org/10.1121/10.0003062>
- 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](https://doi.org/10.1044/1092-4388(2013)12-0030)
- Wang, J., Hahm, S., & Mau, T.** (2015). Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 79–85. <https://doi.org/10.18653/v1/W15-5114>
- Wang, J., Kothalkar, P. V., Cao, B., & Heitzman, D.** (2016). Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. *Proceedings of Interspeech, 2016*, 1195–1199. <https://doi.org/10.21437/Interspeech.2016-1542>
- Wang, J., Samal, A., & Green, J. R.** (2014). Preliminary test of a real-time, interactive silent speech interface based on

electromagnetic articulograph. *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies*, 38–45.

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(1), 15–26. https://doi.org/10.1044/2015_JSLHR-S-14-0112

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook*. University of Wisconsin.

Wisler, A., Goffman, L., Zhang, L., & Wang, J. (2022). Influences of methodological decisions on assessing the spatiotemporal stability of speech movement sequences. *Journal of Speech, Language, and Hearing Research*, 65(2), 538–554. https://doi.org/10.1044/2021_JSLHR-21-00298

Yunusova, Y., Green, J. R., & Mefferd, A. (2009). Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech, Language, and Hearing Research*, 52(2), 547–555. [https://doi.org/10.1044/1092-4388\(2008/07-0218\)](https://doi.org/10.1044/1092-4388(2008/07-0218))

Appendix

Technical Details of Continuous Silent Speech Recognition Experiment

Articulatory input

We performed phoneme-level speech recognition experiments on the MagTrack and electromagnetic articulograph (EMA) data collected. As introduced, the MagTrack returns 3D positional (xyz), 2D orientational (roll and pitch), and 3D magnetic signals. To add some contextual information, the first- and second-order derivatives were concatenated with the original data frames. The EMA data returns 3D positional (xyz) data and 3D positional quaternion data that represent roll and pitch orientational information. The EMA data were concatenated with the first- and second-order derivatives as well. In summary, the input MagTrack data have a maximum dimension of 24, and the EMA data were 18-dimensional. For MagTrack, we performed experiments with and without using the 3D magnetic signals, in which the input dimensions were 24 (with magnetic signal) and 15 (without magnetic signal).

Speech Recognition

We used two standard continuous speech recognizers, Gaussian mixture model (GMM)–hidden Markov model (HMM) and deep neural network (DNN)–HMM. GMM and DNN are used to modeling the probabilities of data for a given phoneme. HMM is to track the phoneme probability distribution with context. There are two types of representation of the phonemes to be recognized: monophone and triphone for HMM. For the monophone representation, we use three HMM states to represent each nonsilence phoneme (begin, middle, and end) and five states to represent the silence. For the triphone representation, we adopted decision tree–based clustering to find the optimal triphone combinations given the data set; each triphone was represented by a three-state HMM. As introduced, we used different stimuli for Subjects 3 and 4. The total number of triphones for Subject 3 is 720, since the stimuli used were highly phoneme-balanced. The triphone number for Subject 4 used 128 triphones, since the stimuli used were from daily communication. These numbers of triphones are also the output dimension of the speech recognizer (number of classes to classify). As we can see, Subject 3 has much more triphones, which means more possible classes and will be a more challenging classification task. We only report the performances of triphone recognitions since they were much better than the monophone recognitions.

The language model used in this study is the Bigram language model, which gives the probabilities of current phonemes given the previous phoneme. A summarized parameter setup in the speech recognition is shown in Table A1.

Table A1. Parameters of the models in the silent speech recognition.

Articulatory movement	
MagTrack	With magnetic signals: position (3-dim) + orientation (2-dim) + magnetic (3-dim) + Δ + $\Delta\Delta$ = 24-dim Without magnetic signals: position (3-dim) + orientation (2-dim) + Δ + $\Delta\Delta$ = 15-dim
EMA	Position (3-dim) + quaternion (3-dim) + Δ + $\Delta\Delta$ = (18-dim)
Sampling rate	MagTrack: 100 Hz (down-sampled from 250 Hz) EMA:100 Hz
GMM–HMM topology	
Monophone	122 states (39 phones \times 3 + 5 for silence) 1,000 Gaussians
Triphone	128 states (432-list) 720 states (400-list) 7,000 Gaussians
Training method	Maximum likelihood estimation (MLE)
DNN–HMM topology	
No. of nodes	512 nodes for each hidden layer
Depth	1–6 depth hidden layers
Training method	RBM pretraining, back-propagation
Input	9 frames at a time (4 previous + current +4 succeeding frames)
Input layer dimension	216 (9 \times 24-dim MagTrack with magnetic signal) 135 (9 \times 15-dim MagTrack w/o magnetic signal) 162 (9 \times 18-dim EMA)
Output layer dimension	Monophone: 122 Triphone:720 (Subject 3) Triphone:128 (Subject 4)
Language model	Bi-gram phone/word language model

Note. dim = dimensional; EMA = electromagnetic articulograph; GMM = Gaussian mixture model; HMM = hidden Markov model; DNN = deep neural network.