Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia

Arnav Kapur ARNAVK@MIT.EDU
Utkarsh Sarawgi UTKARSHS@MIT.EDU
Eric Wadkins EWADKINS@MIT.EDU
Matthew Wu MTWU@MIT.EDU
Media Lab
Massachusetts Institute of Technology

Nora Hollenstein NORAHO@INF.ETHZ.CH
ETH Zurich

Pattie Maes PATTIE@MIT.EDU
Media Lab
Massachusetts Institute of Technology

Editors: Adrian V. Dalca, Matthew Mcdermott, Emily Alsentzer, Sam Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

We present the first non-invasive real-time silent speech system that helps patients with speech disorders to communicate in natural language voicelessly, merely by articulating words or sentences in the mouth without producing any sounds. We collected neuromuscular recordings to build a dataset of 10 trials of 15 sentences from each of 3 multiple sclerosis (MS) patients with dysphonia, spanning a range of severity and subsequently affected speech attributes. We present a pipeline wherein we carefully preprocess the data, develop a convolutional neural architecture and employ personalized machine learning. In our experiments with multiple sclerosis patients, our system achieves a mean overall test accuracy of 0.81 at a mean information transfer rate of 203.73 bits per minute averaged over all patients. Our work demonstrates the potential of a reliable and promising human-computer interface that classifies intended sentences from silent speech and hence, paves the path for future work with further speech disorders in conditions such as amyotrophic lateral sclerosis (ALS), stroke, and oral cancer, among others.

1. Introduction

Multiple sclerosis (MS) is a chronic and progressive autoimmune neurodegenerative disease characterized by lesions in the central nervous system. As the disease progresses, the condition leads to severe physical and cognitive disabilities. The patients suffer from additional impairments with accompanying limitations in activities, restrictions to their participation in life, and compromised quality of life. These impairments include speech disorders such as dysphonia and dysarthria. About 50% of MS patients experience dysphonia or dysarthria (Brown, 2000). Dysphonia refers to disordered sound production at the level of the larynx or voice box. Speech in MS patients with dysphonia ranges from mild hoarseness to barely understandable and audible (Beukelman and Garrett, 1988). For many patients, it

© 2020 A. Kapur, U. Sarawgi, E. Wadkins, M. Wu, N. Hollenstein & P. Maes.

is difficult to speak with a loud and intelligible voice. These quality-of-life issues have been studied from several perspectives. Communication is one of the most fundamental human faculties. Thus, voice disorders can have a drastic impact on the quality of life of patients (Leigh, 2010; Ruotsalainen et al., 2007).

With our work we attempt to take a step towards improving the life quality of MS patients diagnosed with dysphonia by providing a method for them to communicate more easily. We recorded a dataset of silent speech from 3 patients with AlterEgo, a seamless peripheral neuromuscular-computer interface to record silent speech (Kapur et al., 2018; Kapur, 2018). In the scope of this paper, silent speech refers to the act of minimally articulating words without producing sounds. Producing silent speech is less fatiguing for the patients than trying to speak out loud. The signals of interest are the neuromuscular recordings produced as a result of subtle disruptions in the electric field surrounding internal speech articulators as well as surface electromyography (sEMG) signals. These are collected non-invasively from 8 electrodes attached to the user's face and neck regions (Section 3.1). We then preprocess the data and present a convolutional neural network (CNN) architecture to train subject-dependent models to classify the silently spoken sentences with a mean accuracy of 0.81 at a mean information transfer rate of 203.73 bits per minute averaged over all patients (Section 3 and 4). To the best of our knowledge, we are the first to present such neuromuscular recordings in a real-world clinical setting and personalized machine learning models to classify silently articulated sentences.

2. Related work

Existing silent speech interfaces (SSIs) are mostly trained and tested on audible speech (e.g. (Wand et al., 2013)). While there are some SSIs based on non-invasive technologies such as ultrasound (Tóth et al., 2018) or sensors on or around lips and tongue (Kim et al., 2017; Bedri et al.), these interfaces have been developed for healthy subjects and are either bulky or cumbersome and hence not yet practical in real-world clinical environments. Denby et al. (2010) showed the potential benefits of further developing EMG-based silent speech interfaces in terms of low cost, non-invasiveness and capability to use with speech pathologies such as laryngectomy patients, but which still required facial movements or mouthed speech.

AlterEgo specializes in recognizing voiceless silent speech with minimal movements, which makes it applicable in healthcare use cases, such as helping MS patients to communicate. Currently MS patients communicate using devices of augmentative and alternative communication, based on eye blinks or communication boards with letters or symbols, or voice amplifiers, dynamic touch screens, or head-tracking and eye-tracking technologies Cohen et al. (2009). While it has been shown that these communication methods improve the quality of life in early stages of neurodegenerative diseases Bramanti (2019), there are various drawbacks. For instance, these communication systems do not focus on transmitting full utterances in normal speaking speech, but merely slowly spoken single words or concepts, e.g. through word or picture boards Pino (2014). In contrast, AlterEgo focuses on enhancing the functionalities of the afflicted speech production system, by allowing patients to seamlessly communicate in real-time with high information transfer rate via natural language without having to learn a new communication method.

The few existing machine learning methods for recognizing speech from dyphonia or dysarthria patients focus mostly on recognizing isolated words (Asemi et al., 2019; Hawley et al., 2012). For instance, Green et al. (2003) used a Hidden Markov Model to recognize 10 individual words from 5 dysarthria speakers. However, these employ dyarthric speech recognition using voice. Due to the differences between normal and the disabled speech attributes, composing automatic speech recognition systems for the speech disabled is a challenging task (Selouani et al., 2009). In this work, we approach this challenge as a classification task of full sentences, allowing patients to communicate as fluently as possible. We develop a convolutional neural architecture for this task since it has been proven very efficient for physiological time series data. For instance, Yang et al. (2015) and Martinez et al. (2013) both successfully implemented deep convolutional neural networks for multi-channel human data. To the best of our knowledge, AlterEgo is the first silent speech interface to be trained and tested on voiceless speech from MS patients with dysphonia.

3. Methods and materials

3.1. Data collection and preprocessing

Data collection Our signals of interest are the neurological activations of internal speech articulators during voiceless and internal articulation of words and sentences. We allowed minimal movements during articulation as per the comfort of patients. This produces neuromuscular signals as a result of subtle disruptions in the electric field surrounding speech articulators and sEMG. We non-invasively collected such recordings from 3 Multiple Sclerosis (MS) patients with dysphonia using pre-gelled Ag/AgCl surface electrodes while they voicelessly and minimally articulated 15 different sentences 10 times each. The sentences were chosen after surveying the needs of the patients and are summarized in Table 1. This study was conducted with informed consent at The Boston Home under the approval of the Committee on the Use of Humans as Experimental Subjects at Massachusetts Institute of Technology (MIT). We worked with MS patients with dysphonia spanning a range of severity and subsequently affected speech attributes. An anonymized list of the same obtained from the nursing home is shared in Table 2. The table also shares the Kurtzke Expanded Disability Status Scale (EDSS) score of the patients which is a method of quantifying disability in multiple sclerosis. As shown in Figure 1(a)subfigure, the device uses 8 signal electrodes - 4 on the face and 4 on the neck - and a reference and bias electrodes on each earlobe, measuring the potential difference between each signal electrode and the reference electrode. Figure 1(b)subfigure shows a photo of one of the patients during the recordings. The electrode positions were initially chosen from Kapur et al. (2018), and further empirically tweaked after repeated testing for feature-dense signals. The data was digitized using 24-bit Analog-to-Digital converter and sampled at 250 Hz. We used a graphical user interface (GUI) to provide real-time visual feedback of the preprocessed data stream (see below) and annotated the desired data using a push-button. A recorded sentence, segmented between vertical dashed lines on the GUI, can be seen in Figure 1(c)subfigure.

[1] http://www.nationalmssociety.org/nationalmssociety/media/msnationalfiles/brochures/10-2-3-29-edss_form.pdf

Table 1: Sentences recorded with the MS patients.

Hello there good morning	Can you please help me	I have been doing good
How are you doing today	Going to the bathroom	Thank you I appreciate it
What is your name	I am very hungry	You are welcome
It was nice meeting you	Super tired already	I have been doing good
Goodbye see you later	I want to sleep now	I feel sorry for that

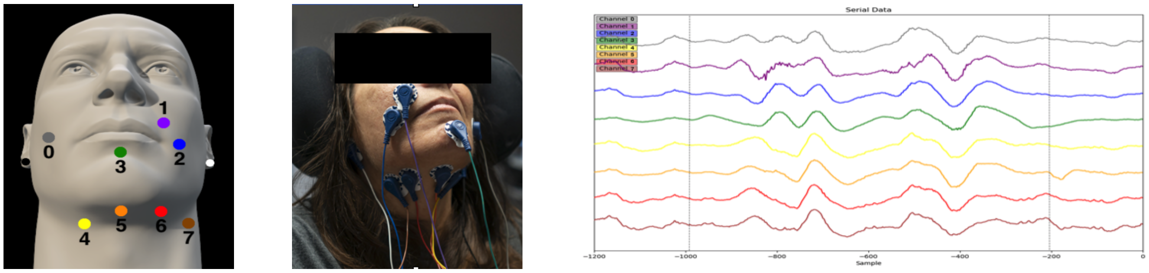
Table 2: MS and dysphonia severity and subsequent attributes of patients we worked with.

Patient ID	Dysphonia severity	Some qualities contributing to dysphonia severity	EDSS score[1]
P1	Moderate	Raspy, hoarse, reduced loudness	8.0
P2	Mild	Mildly strained, hoarse, drops at the end of phrases	8.0
P3	Mild	Reduced pitch control, elongated syllables	8.0

Correcting for DC offset, baseline drift and power line noise Each data channel encounters an arbitrary DC offset due to the variation in electrode placement relative to the reference electrode. Additionally, the data channels face a baseline drift caused by a very low frequency source of noise which is introduced by the device used to record potential differences between signal and reference electrodes. The baseline of each channel changes independently of the others and the direction and rate of drift is unpredictable. To remove these artifacts, the signals were normalized by their initial values and a 1st-order Butterworth high-pass filter was used to remove frequencies lower than 0.5 Hz. The resultant signals were centered by normalizing them to a mean amplitude of zero. 3rd-order Butterworth notch (band-stop) filters were applied at 60 Hz and its harmonics below the sampling rate to remove the power line noise. This technique is one of the many effective approaches documented by Mewett et al. (2001) for removing the power line noise.

Removing heartbeat artifacts Heartbeat artifacts manifest distinctly in the data regardless of electrode placement. A Ricker wavelet (also referred to as Mexican hat wavelet) convolution was applied to select heartbeat artifacts while preserving narrow features of interest. The convolution was then subtracted from the original (corrected for DC offset, baseline drift and power line noise) signal to remove the heartbeat artifacts without significantly distorting the original signal.

Removing high frequency noise The body introduces high frequency noise that lies beyond the frequency range of features of our interest. A 1st-order Butterworth band-pass



(a) Placement of surface electrodes.

(b) A patient recording data.

(c) Snapshot of the real-time GUI while recording data.

Figure 1: An instance of data collection.

filter was applied to allow for frequencies from 0.5 to 8 Hz, determined through analysis of the frequency domain.

Figure 2 (continued on the next page) illustrates all the preprocessing for a data segment of ≈3.5 seconds sampled at 1 kHz (≈3500 samples). The range of resultant signal values is under 100 microvolts.

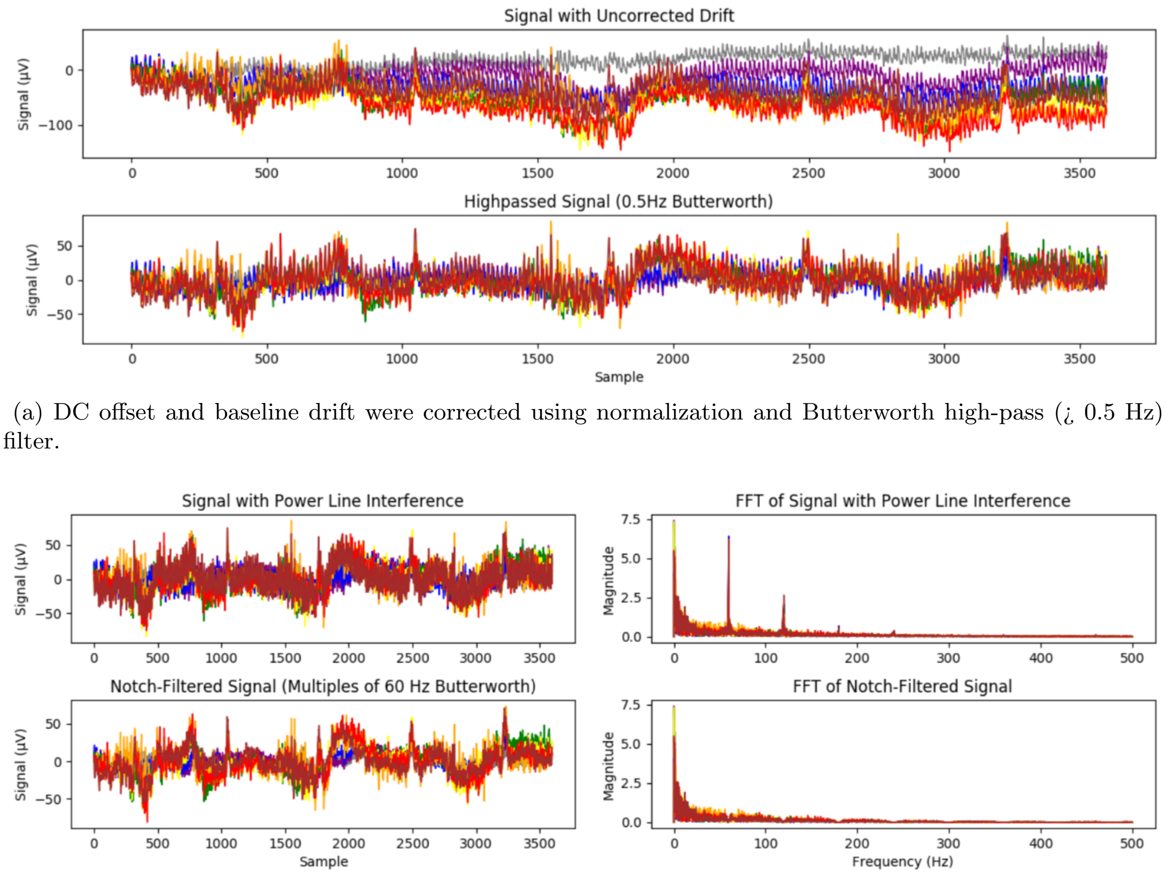
3.2. Dataset preparation

The desired data sequences annotated with their corresponding labels (sentences) were padded with respective surrounding values of preprocessed data stream. The padded sequences were then truncated to a fixed length of 900 samples, accepted by our CNN architecture as input (Section 3.3). This fixed length was empirically chosen based on the lengths of all annotated sequences. The complete dataset of each patient consists of 10 data points (i.e. sentences) of 15 classes each. All models were trained, validated and tested 5 times each with stratified 5-fold cross-validation (CV) with a 60-20-20 train-validation-test split, ensuring class balance in each split.

3.3. Model architecture and training

Figure 3 describes our convolutional neural architecture. It was trained end-to-end with data sequences (each data sequence being 900 x 8, for 8 channel data) and their corresponding labels (15 classes). The training used an Adam optimizer with a learning rate of e^{-4} to minimize cross-entropy loss. A batch size of 50 data points was used. The best model was saved according to validation accuracy and evaluated for test accuracy in each fold during the 5-fold repeated stratified cross-validation. Using an NVIDIA TITAN Xp, the CNN trained and validated each fold in ≈55 seconds, and evaluated each data instance in ≈2.3 milliseconds.

We employed personalized machine learning for the patients, i.e. the data of each patient was kept separate and not mixed with others. Hence, the architecture was trained with personal data to generate separate models for each of the 3 patients. The primary reasons are the attributes and features present in the data, which are unique to each patient due



(a) DC offset and baseline drift were corrected using normalization and Butterworth high-pass (¿ 0.5 Hz) filter.

(b) Power line noise was removed using Butterworth notch filters at 60 Hz and its harmonics - time-domain analysis (left) and frequency-domain analysis (right). FFT abbreviates for fast fourier transform.

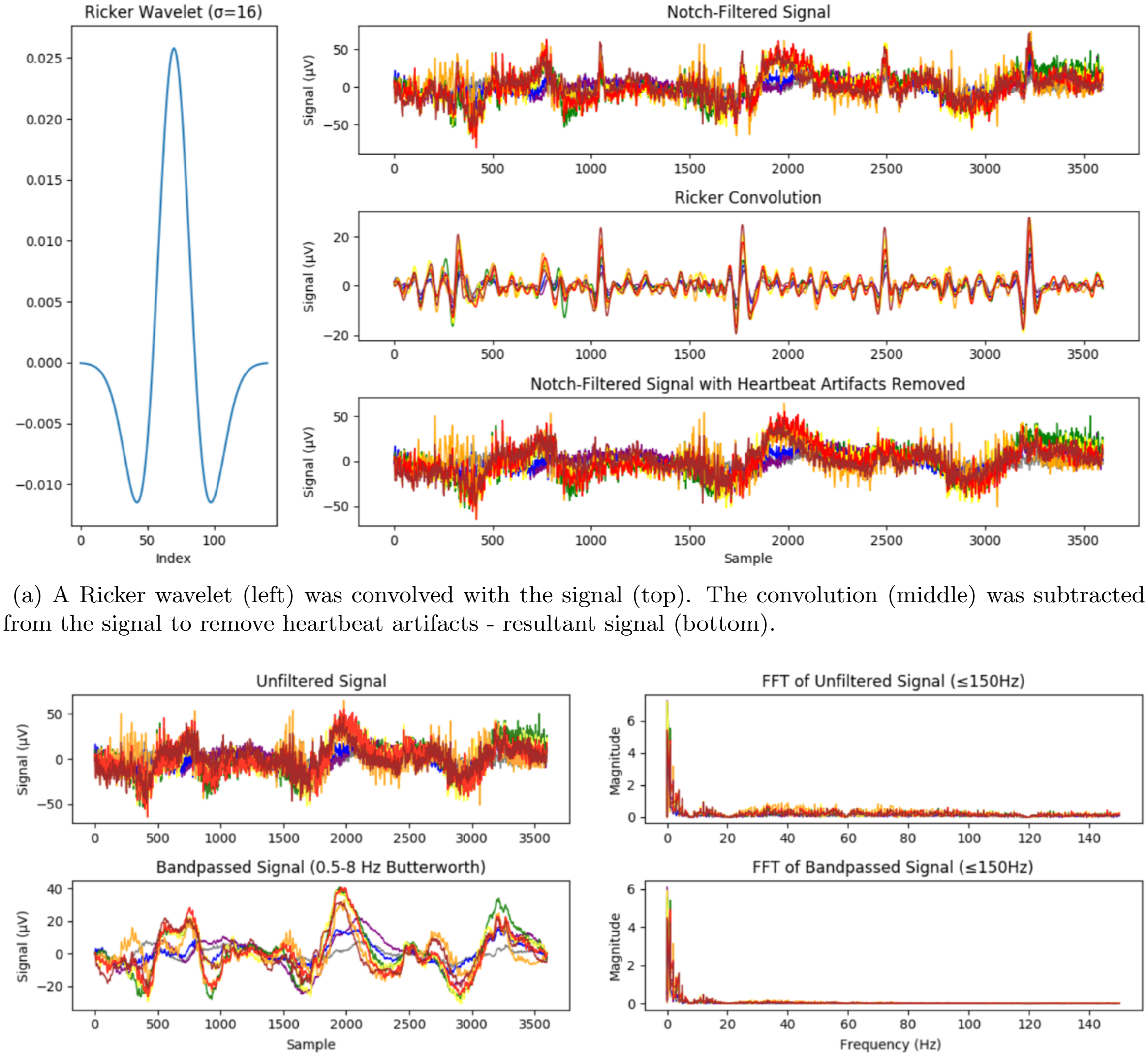
Figure 2: Preprocessing for a data segment of ≈3.5 seconds sampled at 1 kHz.

to their personal characteristics and stage of the disease and subsequent speech pathology, and the variance in the electrode positions.

Architecture design and tuning The model architecture was designed and tuned with data of patient P2 (chosen at random) divided into 80-20 train-validation split. All reported accuracy numbers in architecture design and tuning correspond to stratified 5-fold cross-validation results. Table 3 summarizes our progressive neural architecture design and corresponding CV accuracy while adding and removing layers. We then further tuned hyperparameters such as batch size, learning rate, rates of both dropout layers and number of units in fully connected layer among others.

4. Results

We used repeated (5 times) stratified 5-fold cross-validation to evaluate our model on the 15 class classification task. The models achieved overall test accuracy of 0.79 (±0.01), 0.87 (±0.01) and 0.77 (±0.02) for the three MS patients P1, P2 and P3 respectively. We



(a) A Ricker wavelet (left) was convolved with the signal (top). The convolution (middle) was subtracted from the signal to remove heartbeat artifacts - resultant signal (bottom).

(b) High frequency noise was removed using Butterworth band-pass filter to allow from 0.5 to 8 Hz - time-domain analysis (left) and frequency-domain analysis (right). FFT abbreviates for fast fourier transform.

Figure 2: Preprocessing for a data segment of ≈3.5 seconds sampled at 1 kHz. (continued)

calculated the Cohen's kappa coefficient κ to model the comparison between observed and chance agreement for each patient (Table 4). Table 5 shows the micro and macro average of precision, recall and F1-score for each patient. We calculated these by averaging the respective metrics over 5 runs of 5 folds each to account for little randomness in the model. Figure 4 shows the confusion matrices corresponding to the best and the worst of the 5 runs (according to their overall accuracy scores) for patient P2 (chosen at random). The same for other patients are shared in Appendix A.

While we evaluate the accuracy of our model, it is also important to evaluate the rate of speech (words per minute) and information transfer rate (Table 6). We calculated the rate of speech for each patient by averaging the number of words per minute of the sentences.

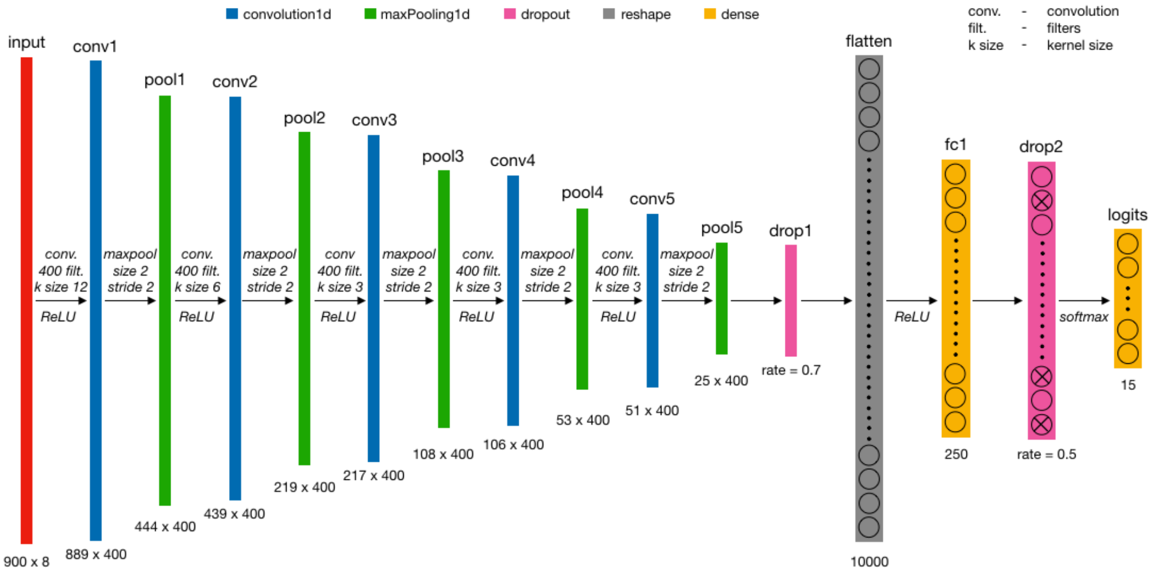


Figure 3: Convolutional neural architecture showing different layers, layer names and dimensions (above and below respective layers), activation functions and other related hyperparameters.

Table 3: Progressive neural architecture design using data of patient P2. Layer names are referenced with the architecture in figure 3. Layers 'input - conv1 - pool1 - conv2 - pool2 - conv3 - pool3' have been referred to as 'start1' together. Layers 'start1 - conv4 - pool4 - conv5 - pool5' have been referred to as 'start2' together. Layer fc1 consists of 250 units, and layers drop1 and drop2 use a dropout rate of 0.5 in each instance. Accuracy numbers correspond to stratified 5-fold CV results.

Layers	CV accuracy
start1 - flatten - fc1 - logits	0.21
start1 - conv4 - pool4 - flatten - fc1 - logits	0.37
start1 - conv4 - pool4 - conv5 - pool5 - flatten - fc1 - logits	0.87
start2 - conv6 - pool6 - flatten - fc1 - logits	0.85
start2 - conv6 - pool6 - conv7 - pool7 - flatten - fc1 - logits	0.83
start2 - drop1 - flatten - fc1 - logits	0.88
start2 - drop1 - flatten - fc1 - drop2 - logits	0.94

We also measured the information transfer rate (also referred to as bit-rate), as originally proposed by Wolpaw et al. (1998), to evaluate the performance of our human-computer interface. This metric accounts for the accuracy and size of the vocabulary in addition to the rate of speech. A less accurate system is less capable of transferring correct information due to errors. Larger vocabularies allow for the transfer of more varied information with a

Table 4: Observed and chance agreement, and κ value for each patient.

Patient ID	Observed agreement	Chance agreement	κ value
P1	0.79 (±0.01)	0.07	0.78
P2	0.87 (±0.01)	0.07	0.86
P3	0.77 (±0.02)	0.07	0.75

Table 5: 15-class classification report for each patient - micro and macro average of precision, recall and F1-score.

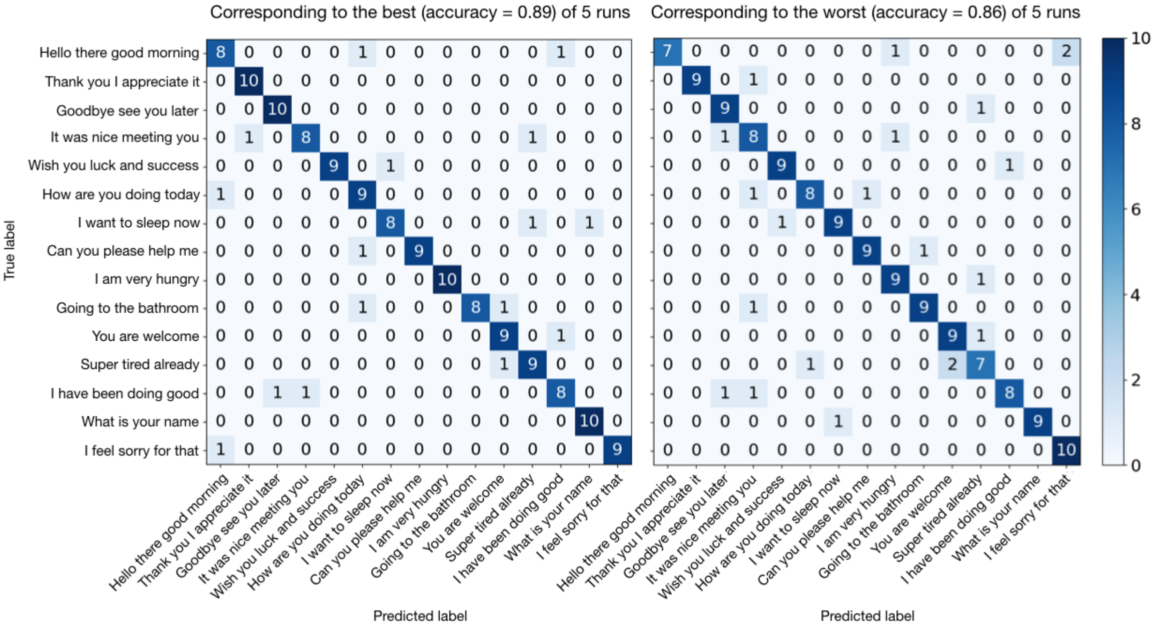
Patient ID	Average	Precision	Recall	F1-score
P1	Micro	0.79	0.79	0.79
	Macro	0.79	0.79	0.77
P2	Micro	0.87	0.87	0.87
	Macro	0.90	0.87	0.87
P3	Micro	0.77	0.77	0.77
	Macro	0.79	0.77	0.75

single symbol. And higher rates of speech allow faster transfer. The bit-rate of a system is calculated as shown in Equation 1 below, where B is the bit-rate corresponding to the amount of information reliably transferred over the system in bits per minute, V is the application speed (average words per minute in our case), P is the classification accuracy (overall test accuracy in our case) and N is the vocabulary size (15 in our case):

$$B = V\left[\log_2 N + P\log_2 P + (1-P)\log_2\left(\frac{1-P}{N-1}\right)\right] \tag{1}$$

Table 6: Performance report of the human-computer interface for each patient - classification accuracy, application speed and information transfer rate (bit-rate).

Patient ID	Classification accuracy	Application speed	Bit-rate
		(per minute)	(bits per minute)
P1	0.79	94.32	223.15
P2	0.87	79.66	227.39
P3	0.77	71.30	160.66



Corresponding to the best (accuracy = 0.89) of 5 runs

	Hello there good morning	Thank you I appreciate it	Goodbye see you later	It was nice meeting you	Wish you luck and success	How are you doing today	I want to sleep now	Can you please help me	I am very hungry	Going to the bathroom	You are welcome	Super tired already	I have been doing good	What is your name	I feel sorry for that
Hello there good morning	8	0	0	0	0	1	0	0	0	0	0	0	1	0	0
Thank you I appreciate it	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
Goodbye see you later	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
It was nice meeting you	0	1	0	8	0	0	0	0	0	0	0	1	0	0	0
Wish you luck and success	0	0	0	0	9	0	1	0	0	0	0	0	0	0	0
How are you doing today	1	0	0	0	0	9	0	0	0	0	0	0	0	0	0
I want to sleep now	0	0	0	0	0	0	8	0	0	0	0	1	0	1	0
Can you please help me	0	0	0	0	0	1	0	9	0	0	0	0	0	0	0
I am very hungry	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
Going to the bathroom	0	0	0	0	0	1	0	0	0	8	1	0	0	0	0
You are welcome	0	0	0	0	0	0	0	0	0	0	9	0	1	0	0
Super tired already	0	0	0	0	0	0	0	0	0	0	1	9	0	0	0
I have been doing good	0	0	1	1	0	0	0	0	0	0	0	0	8	0	0
What is your name	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
I feel sorry for that	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9

Corresponding to the worst (accuracy = 0.86) of 5 runs

	Hello there good morning	Thank you I appreciate it	Goodbye see you later	It was nice meeting you	Wish you luck and success	How are you doing today	I want to sleep now	Can you please help me	I am very hungry	Going to the bathroom	You are welcome	Super tired already	I have been doing good	What is your name	I feel sorry for that
Hello there good morning	7	0	0	0	0	0	0	0	1	0	0	0	0	0	2
Thank you I appreciate it	0	9	0	1	0	0	0	0	0	0	0	0	0	0	0
Goodbye see you later	0	0	9	0	0	0	0	0	0	0	0	1	0	0	0
It was nice meeting you	0	0	1	8	0	0	0	0	1	0	0	0	0	0	0
Wish you luck and success	0	0	0	0	9	0	0	0	0	0	0	0	1	0	0
How are you doing today	0	0	0	1	0	8	0	1	0	0	0	0	0	0	0
I want to sleep now	0	0	0	0	1	0	9	0	0	0	0	0	0	0	0
Can you please help me	0	0	0	0	0	0	0	9	0	1	0	0	0	0	0
I am very hungry	0	0	0	0	0	0	0	0	9	0	0	1	0	0	0
Going to the bathroom	0	0	0	1	0	0	0	0	0	9	0	0	0	0	0
You are welcome	0	0	0	0	0	0	0	0	0	0	9	1	0	0	0
Super tired already	0	0	0	0	0	1	0	0	0	0	2	7	0	0	0
I have been doing good	0	0	1	1	0	0	0	0	0	0	0	0	8	0	0
What is your name	0	0	0	0	0	0	1	0	0	0	0	0	0	9	0
I feel sorry for that	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10

Figure 4: Confusion matrices for patient P2 for the best (left) and the worst (right) of 5 runs.

5. Discussion and future work

To the best of our knowledge, our work is the first that presents a non-invasive wearable silent speech device to facilitate real-time natural communication using minimal speech engagement, and evaluates it with real-world clinical data. This allows the patients to communicate in natural language even with no voice at all. For instance, Patient P1 self-reported that he/she usually has to force his/her diaphragm in order to produce sounds, and looses his/her voice completely at times. Our device is able to decode his/her minimally articulated sentences with overall test accuracy of 0.79 at 223.25 bits per minute without any voice. We worked with MS patients with dysphonia spanning a range of severity and subsequently affected speech attributes (Table 2). We also collected patients' responses such as comfort, adaptability and usability of our technology to get a better perspective from people we are developing the technology for. The performance results of AlterEgo and responses of patients look promising to help improve the quality of life of such people.

The recorded silent speech data does contain unavoidable artifacts of irregular head, eye and eyebrow movements which are unpreventable due to their MS condition. However, even with the current levels of noise, our model seems robust enough to extract the desired features in a real-world setting as evidenced by the results. We do observe a slight decrease in accuracy with significant increase in dysphonia and MS severity. Patient P2, for example, has the mildest dysphonia and MS, and achieves the best overall results among the 3 patients.

Our next steps include deploying a trained model for extensive real-time evaluation to decode the sentences from silent speech and facilitate seamless conversation. We have tried the same with a patient using text-to-speech on the decoded sentences. This also provides an auditory feedback to the patient about the decoded text in real time.

Due to the condition of the patients, it was only possible to record a small dataset for each patient. The recordings lasted 30 to 40 minutes on average with each patient. Future work includes improving user experience and design, gathering and evaluating on more data points per sentence and extending to other various sentences. This includes collecting training data over many sessions. We also aim to test the system rigorously over long-term and analyze how the progression of the MS condition and the subsequent speech problems affect the results. While we have currently worked with MS patients with dysphonia, we also plan to test our system with ones with dysarthria and other speech disorders in conditions like ALS, stroke, and oral cancer, among others.

6. Conclusion

We present the first non-invasive and wearable silent speech interface that is evaluated on MS patients with dysphonia. Our system achieves a mean overall accuracy of 0.81 at a mean information transfer rate of 203.73 bits per minute averaged over 3 patients. This work demonstrates promising results and a novel step towards assisting people with speech disorders to communicate in natural language without having to produce or strain their voice. We aim to extend our work to people with other speech and voice disorders in conditions like amyotrophic lateral sclerosis (ALS), stroke, and oral cancer among others, and conduct longitudinal user studies going ahead.

Acknowledgments

We express our gratitude to the subjects and their families for their dedication to improving the lives of persons living with unusual speech pathologies. We thank A. Burnham for his help in conducting clinical experiments at the Boston Home. This work was supported by Microsoft's AI for Accessibility grant.

References

Adeleh Asemi, Siti Salwah Binti Salim, Seyed Reza Shahamiri, Asefeh Asemi, and Narjes Houshangi. Adaptive neuro-fuzzy inference system for evaluating dysarthric automatic speech recognition (asr) systems: a case study on mvml-based asr. *Soft Computing*, 23 (10):3529–3544, 2019.

Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. Toward silent-speech control of consumer wearables.

David Beukelman and Kathryn Garrett. Augmentative and alternative communication for adults with acquired severe communication disorders. *Augmentative and Alternative Communication*, 4(2):104–121, 1988.

Placido Bramanti. Augmentative and alternative communication improves quality of life in the early stages of amyotrophic lateral sclerosis. *Functional neurology*, 34(1):35–43, 2019.

Shirley A Brown. Swallowing and speaking challenges for the ms patient. *International Journal of MS Care*, 2(3):4–14, 2000.

Seth M Cohen, Alphi Elackattu, J Pieter Noordzij, Michael J Walsh, and Susan E Langmore. Palliative treatment of dysphonia and dysarthria. *Otolaryngologic Clinics of North America*, 42(1):107–121, 2009.

Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. Silent speech interfaces. *Speech Communication*, 52(4):270–287, 2010.

Phil Green, James Carmichael, Athanassios Hatzis, Pam Enderby, Mark Hawley, and Mark Parker. Automatic speech recognition with sparse training data for dysarthric speakers. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Mark S Hawley, Stuart P Cunningham, Phil D Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O'Neill. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(1):23–31, 2012.

Arnav Kapur. Human-machine cognitive coalescence through an internal duplex interface, 2018.

Arnav Kapur, Shreyas Kapur, and Pattie Maes. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*, pages 43–53. ACM, 2018.

Myungjong Kim, Beiming Cao, Ted Mau, and Jun Wang. Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2323–2336, 2017.

Angela Leigh. The experiences of intimacy for adults with acquired communication disorders using augmentative and alternative communication (aac). 2010.

Hector P Martinez, Yoshua Bengio, and Georgios N Yannakakis. Learning deep physiological models of affect. *IEEE Computational intelligence magazine*, 8(2):20–33, 2013.

David T Mewett, Homer Nazeran, and Karen J Reynolds. Removing power line noise from recorded emg. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2190–2193. IEEE, 2001.

Alexandros Pino. Augmentative and alternative communication systems for the motor disabled. In *Disability Informatics and Web Accessibility for Motor Limitations*, pages 105–152. IGI Global, 2014.

Jani H Ruotsalainen, Jaana Sellman, Laura Lehto, Leena K Isotalo, and Jos H Verbeek. Interventions for preventing voice disorders in adults. *Cochrane Database of Systematic Reviews*, (4), 2007.

Sid-Ahmed Selouani, Mohammed Sidi Yakoub, and Douglas O'Shaughnessy. Alternative speech communication system for persons with severe speech disorders. *EURASIP Journal on Advances in Signal Processing*, 2009(1):540409, 2009.

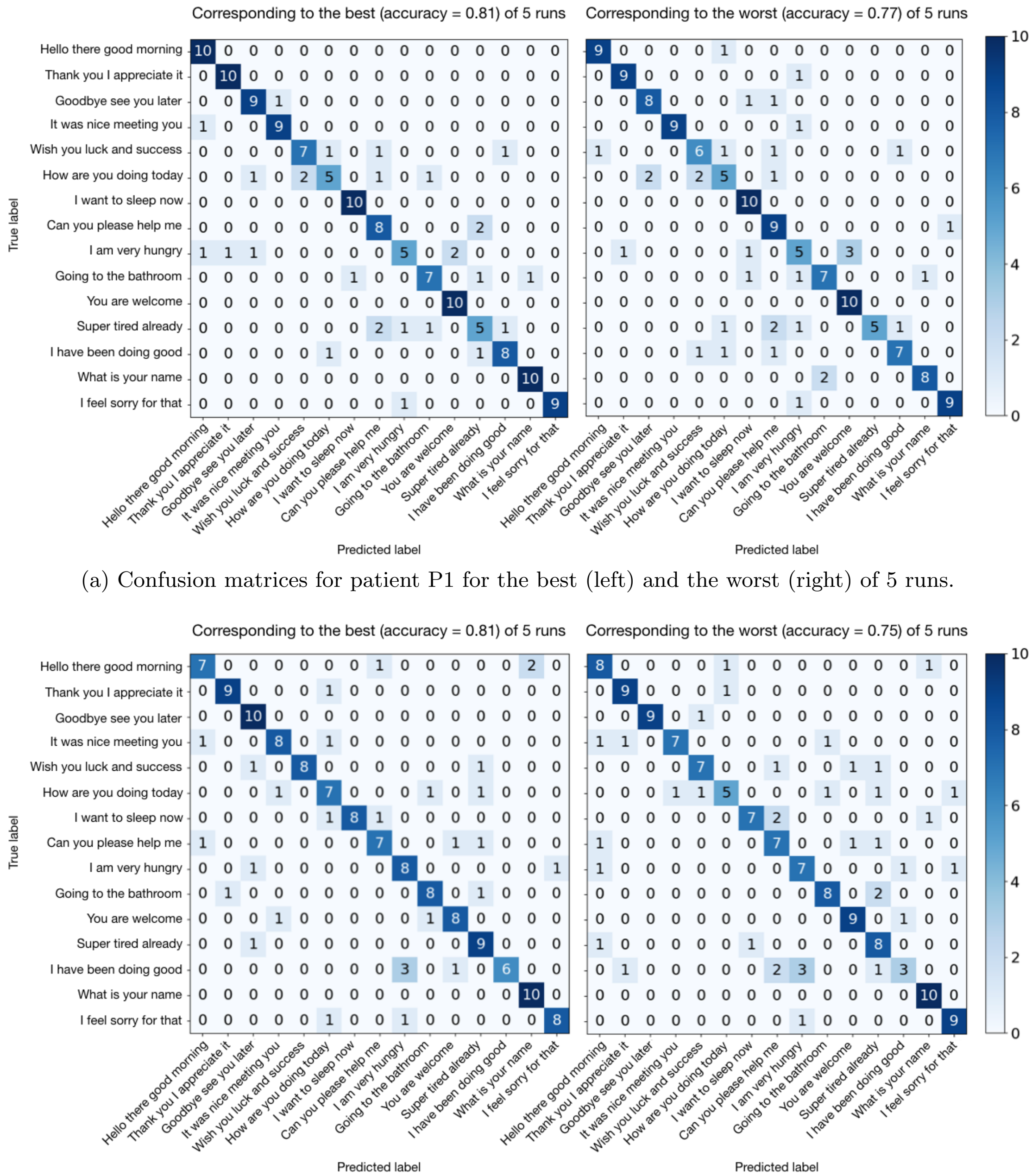
László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó. Multi-task learning of speech recognition and speech synthesis parameters for ultrasound-based silent speech interfaces. In *Interspeech*, pages 3172–3176, 2018.

Michael Wand, Christopher Schulte, Matthias Janke, and Tanja Schultz. Array-based electromyographic silent speech interface. In *BIOSIGNALS*, pages 89–96, 2013.

Jonathan R Wolpaw, Herbert Ramoser, Dennis J McFarland, and Gert Pfurtscheller. Eeg-based communication: improved accuracy by response verification. *IEEE transactions on Rehabilitation Engineering*, 6(3):326–333, 1998.

Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Appendix A.



Corresponding to the best (accuracy = 0.81) of 5 runs

	Hello there good morning	Thank you I appreciate it	Goodbye see you later	It was nice meeting you	Wish you luck and success	How are you doing today	I want to sleep now	Can you please help me	I am very hungry	Going to the bathroom	You are welcome	Super tired already	I have been doing good	What is your name	I feel sorry for that
Hello there good morning	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thank you I appreciate it	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
Goodbye see you later	0	0	9	1	0	0	0	0	0	0	0	0	0	0	0
It was nice meeting you	1	0	0	9	0	0	0	0	0	0	0	0	0	0	0
Wish you luck and success	0	0	0	0	7	1	0	1	0	0	0	0	1	0	0
How are you doing today	0	0	1	0	2	5	0	1	0	1	0	0	0	0	0
I want to sleep now	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
Can you please help me	0	0	0	0	0	0	0	8	0	0	0	2	0	0	0
I am very hungry	1	1	1	0	0	0	0	0	5	0	2	0	0	0	0
Going to the bathroom	0	0	0	0	0	0	1	0	0	7	0	1	0	1	0
You are welcome	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
Super tired already	0	0	0	0	0	0	0	2	1	1	0	5	1	0	0
I have been doing good	0	0	0	0	0	1	0	0	0	0	0	1	8	0	0
What is your name	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
I feel sorry for that	0	0	0	0	0	0	0	0	1	0	0	0	0	0	9

Corresponding to the worst (accuracy = 0.77) of 5 runs

	Hello there good morning	Thank you I appreciate it	Goodbye see you later	It was nice meeting you	Wish you luck and success	How are you doing today	I want to sleep now	Can you please help me	I am very hungry	Going to the bathroom	You are welcome	Super tired already	I have been doing good	What is your name	I feel sorry for that
Hello there good morning	9	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Thank you I appreciate it	0	9	0	0	0	0	0	0	1	0	0	0	0	0	0
Goodbye see you later	0	0	8	0	0	0	1	1	0	0	0	0	0	0	0
It was nice meeting you	0	0	0	9	0	0	0	0	1	0	0	0	0	0	0
Wish you luck and success	1	0	0	0	6	1	0	1	0	0	0	0	1	0	0
How are you doing today	0	0	2	0	2	5	0	1	0	0	0	0	0	0	0
I want to sleep now	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
Can you please help me	0	0	0	0	0	0	0	9	0	0	0	0	0	0	1
I am very hungry	0	1	0	0	0	0	1	0	5	0	3	0	0	0	0
Going to the bathroom	0	0	0	0	0	0	1	0	1	7	0	0	0	1	0
You are welcome	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
Super tired already	0	0	0	0	0	1	0	2	1	0	0	5	1	0	0
I have been doing good	0	0	0	0	1	1	0	1	0	0	0	0	7	0	0
What is your name	0	0	0	0	0	0	0	0	0	2	0	0	0	8	0
I feel sorry for that	0	0	0	0	0	0	0	0	1	0	0	0	0	0	9

(a) Confusion matrices for patient P1 for the best (left) and the worst (right) of 5 runs.

Corresponding to the best (accuracy = 0.81) of 5 runs

	Hello there good morning	Thank you I appreciate it	Goodbye see you later	It was nice meeting you	Wish you luck and success	How are you doing today	I want to sleep now	Can you please help me	I am very hungry	Going to the bathroom	You are welcome	Super tired already	I have been doing good	What is your name	I feel sorry for that
Hello there good morning	7	0	0	0	0	0	0	1	0	0	0	0	0	2	0
Thank you I appreciate it	0	9	0	0	0	1	0	0	0	0	0	0	0	0	0
Goodbye see you later	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
It was nice meeting you	1	0	0	8	0	1	0	0	0	0	0	0	0	0	0
Wish you luck and success	0	0	1	0	8	0	0	0	0	0	0	1	0	0	0
How are you doing today	0	0	0	1	0	7	0	0	0	1	0	1	0	0	0
I want to sleep now	0	0	0	0	0	1	8	1	0	0	0	0	0	0	0
Can you please help me	1	0	0	0	0	0	0	7	0	0	1	1	0	0	0
I am very hungry	0	0	1	0	0	0	0	0	8	0	0	0	0	0	1
Going to the bathroom	0	1	0	0	0	0	0	0	0	8	0	1	0	0	0
You are welcome	0	0	0	1	0	0	0	0	0	1	8	0	0	0	0
Super tired already	0	0	1	0	0	0	0	0	0	0	0	9	0	0	0
I have been doing good	0	0	0	0	0	0	0	0	3	0	1	0	6	0	0
What is your name	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
I feel sorry for that	0	0	0	0	0	1	0	0	1	0	0	0	0	0	8

Corresponding to the worst (accuracy = 0.75) of 5 runs

	Hello there good morning	Thank you I appreciate it	Goodbye see you later	It was nice meeting you	Wish you luck and success	How are you doing today	I want to sleep now	Can you please help me	I am very hungry	Going to the bathroom	You are welcome	Super tired already	I have been doing good	What is your name	I feel sorry for that
Hello there good morning	8	0	0	0	0	1	0	0	0	0	0	0	0	1	0
Thank you I appreciate it	0	9	0	0	0	1	0	0	0	0	0	0	0	0	0
Goodbye see you later	0	0	9	0	1	0	0	0	0	0	0	0	0	0	0
It was nice meeting you	1	1	0	7	0	0	0	0	0	1	0	0	0	0	0
Wish you luck and success	0	0	0	0	7	0	0	1	0	0	1	1	0	0	0
How are you doing today	0	0	0	1	1	5	0	0	0	1	0	1	0	0	1
I want to sleep now	0	0	0	0	0	0	7	2	0	0	0	0	0	1	0
Can you please help me	1	0	0	0	0	0	0	7	0	0	1	1	0	0	0
I am very hungry	1	0	0	0	0	0	0	0	7	0	0	0	1	0	1
Going to the bathroom	0	0	0	0	0	0	0	0	0	8	0	2	0	0	0
You are welcome	0	0	0	0	0	0	0	0	0	0	9	0	1	0	0
Super tired already	1	0	0	0	0	0	1	0	0	0	0	8	0	0	0
I have been doing good	0	1	0	0	0	0	0	2	3	0	0	1	3	0	0
What is your name	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
I feel sorry for that	0	0	0	0	0	0	0	0	1	0	0	0	0	0	9

(b) Confusion matrices for patient P3 for the best (left) and the worst (right) of 5 runs.

Figure 5: Confusion matrices for patients P1 and P3.