Pathology-specific settings for voice activity detection in a multimodal dialog agent for digital health monitoring

Jackson Liscombe, Hardik Kothare, Michael Neumann, David Pautler, and Vikram Ramanarayanan

Abstract Optimal voice activity detection (VAD) settings, used to automatically detect the end of speaker turns in an automated spoken dialog system, differ for pathological and non-pathological speech. However, VAD settings may be further conditioned on the cognitive or neurological condition of the user, especially where patients are severely affected. Here, we present preliminary investigations into optimal VAD parameter setting bounds, as measured with the NIST detection cost function (DCF), for pathological speech collected from severe presentations of four conditions: amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD), schizophrenia, and depression. We found, via simulation experiments, that the amount of non-speech time to wait before deciding that the participant has finished their speaking turn was especially discriminating. A 2.6 second wait time was found to be optimal for pathological conditions like ALS and PD that exhibit dysarthric/speech motor symptoms; whereas, 4.0-4.5 seconds was best for those with associated mental health symptoms. Our results suggest that optimizing voice activity detection systems for pathological cohorts in this manner can greatly enhance user experience by reducing interruptions while minimizing dialog system response time.

1 Introduction

Dialog systems offer the potential to improve availability, frequency, and quality of patient care for neurological and mental health conditions because such technology can be used to drive automated patient assessments that previously required face-to-face sessions with a clinician (1; 2; 3). As most dialog systems have been developed for non-disordered adult speech, their performance can decrease substantially when confronted with pathological speech, a symptom of conditions such as

Modality.AI, Inc.
e-mail: vikram.ramanarayanan@modality.ai

Parkinson's, Alzheimer's, multiple sclerosis (MS), or Amyotrophic Lateral Sclerosis (ALS) (4). One component that is particularly crucial for seamless dialog interaction is voice activity detection (VAD); however, the special characteristics of pathological speech, like poor articulation, disfluencies, variable intra-word pause lengths, atypical acoustic, motoric and emotional properties, etc., make VAD a much more challenging problem than when applied to healthy speech (5; 6).

Furthermore, to better personalize such dialog-based digital health monitoring solutions, VAD detection algorithms must account for the fact that patients suffering from different neurological and mental health conditions exhibit a variety of different pathological speech characteristics (7). For instance, people with ALS exhibit characteristic hoarseness in their voice (8), atypical spectral acoustic features (9), and longer and more variable pause durations (10). Atypical prosody and duration have also been demonstrated as markers of Parkinson's disease (11). On the other hand, patients suffering from mental health issues, such as those at a high risk of clinical depression and suicidality, display decreased verbal activity productivity, diminished prosodic variability and/or monotonous, "lifeless" sounding speech (12). Schizophrenic patients can have deficits in attention, memory, and executive functioning; confused and disordered thinking and speech; trouble with logical thinking; and sometimes bizarre behavior or abnormal movements (13).

To extend the utility of such VAD (and therefore dialog) systems to be truly generalizable across neurological and mental health conditions, and deal with speech patterns across multiple conditions—especially where patients are severely affected—we need a better understanding of optimal VAD parameters for these patients in particular. In previous work, we have shown that optimal VAD parameters are different for participants with ALS vs healthy controls, especially the amount of non-speech time to wait before considering a turn complete (14). We extend those results here with a preliminary investigation into how optimal VAD parameter settings vary across severe presentations of four different disorders: amyotrophic lateral sclerosis (ALS), Parkinson's disease, schizophrenia, and depression. To answer this question, we first simulate the performance of a VAD system that is optimized across severe presentations of four pathological speech cohorts: ALS, Parkinson's disease, schizophrenia, and depression. We then investigate the VAD performance within each cohort, particularly with respect to the interruption rate and NIST detection cost function (DCF), and examine how parameter tuning can ameliorate performance. To our knowledge, this is the first investigation of pathology-specific VAD settings for digital health dialog agents in the literature.

2 System

We use NEMSI (NEurological and Mental health Screening Instrument), a cloud-based multimodal dialog system that conducts on-demand automated screening interviews for the assessment or monitoring of various neurological and mental health conditions for the VAD experiments described in this paper (3). Dialog turn man-

agement in NEMSI is managed in part by voice activity detection (VAD) using the CMU Sphinx open source speech recognition toolkit.[1] The algorithm uses a two-step process to identify spans of speech and non-speech in a stream of audio. See (14) for more details.

There are six VAD parameters whose values can be configured to optimize performance. These are: (i) **`minSignal`**, the minimum required energy level (dB) for a speech frame; (ii) **`adjustment`**, the factor by which the background level estimation is increased with each successful speech frame; (iii) **`threshold`**, the energy level of the required difference between the background noise and average signal level estimations (dB); (iv) **`startSpeech`**, time in milliseconds of speech required to trigger the start of a speech event, (v) **`endSilence`**, time in milliseconds of non-speech required to designate the end of a speech turn, and (vi) **`initialSilence`**, time in milliseconds of non-speech to wait for detection of user speech before considering the user input devoid of speech.

3 Data

The experiments described in this paper analyze 697 utterances from 40 complete dialog interactions, each from unique participants, selected from 4 different medical conditions; 10 sessions from each domain. The number of turns in each session ranged from 9 to 36 (mean=17.4; SD=9.1) and turn duration ranged from 7.5s to 151.2s (mean=29.1s; SD=23.7s) depending on the task. In total, our corpus comprises 5.6 hours of user audio[2]. Conversations were selected based on severity of disease, as outlined below.

3.1 Parkinson's Disease

Demographic and diagnosis information was used to select conversations for the Parkinson's cohort. Parkinson's patients were recruited and consented through the Purdue Motor Speech Lab as part of an ongoing collaboration with Purdue University (15), approved by Purdue's Institutional Review Board. In addition to having an official diagnosis, each user completed a Communicative Participation Item Bank (CPIB) survey (16). The CPIB asks patients to rate how much their condition interferes with participation in ten everyday verbal communication situations such as talking to people they know, ordering a meal in a restaurant, having a conversation

[1] `https://cmusphinx.github.io/`

[2] Note that the sample size here is relatively smaller than typical studies that aim to generalize results to a population. However, we argue that our results are informative and useful nonetheless, given that we are looking at *extreme* cases, i.e., severe presentations of each disorder condition. Also, for various reasons including reduced patient ability and increased patient burden, it is often relatively challenging to collect data in large amounts from more severely progressed patients.

in a small group, etc. A summary score is obtained by assigning each answer a point value: “Not at all”=0, “A little”=1, “Quite a bit”=2, “Very much”=3. According to this system, a CPIB score of 30 indicates severely affected speech.

A complete session was chosen at random for five male and five female patients with the highest CPIB scores. Selected CPIB scores ranged from 15-30. The total number of patient turns in this cohort was 327. Each participant turn is a response to one of eight tasks: sustained vowel phonation; read short and long sentences, questions, and paragraphs; and short and long spontaneous speech.

3.2 Amyotrophic Lateral Sclerosis (ALS)

Data from ALS patients comes from a ongoing collaborative IRB-approved study with EverythingALS and the Peter Cohen Foundation[3] (17). Demographic and diagnosis information was used to select conversations for the ALS cohort. In addition to having an official diagnosis, each user completed an ALS-FRS-R survey (18). Though the ALS-FRS-R survey covers many topics related to possible ALS symptoms, the first question specifically asks the patient to rate their speech capabilities according to the following scale: “Normal speech processes”=4, “Detectable speech disturbance”=3, “Intelligible with repeating”=2, “Speech combined with nonvocal communication”=1, “Loss of useful speech”=0.

A complete session was chosen at random for the five participants of each sex with the lowest ALS-FRS-R Q1 scores. Selected ALS-FRS-R Q1 scores ranged from 0-2. The total number of patient turns in this cohort was 117. Each participant turn is a response to one of eight tasks: sustained vowel phonation, counting, and diadochokinetic syllables[4]; read short sentences and a longer paragraph; spontaneous speech; and description of a picture.

3.3 Schizophrenia

Demographic and diagnosis information was used to select conversations for the Schizophrenia cohort. Data collection was approved and done in collaboration with the Nathan S. Kline Institute for Psychiatric Research, and written informed consent was obtained from all participants at the time of screening after explaining details of the study. In addition to having an official diagnosis, each user also has a clinician-assessed Brief Negative Symptom Scale (BNSS) score (19). BNSS is an assessment of negative symptoms in patients with schizophrenia on the following sub-scales: anhedonia (0–18), distress (0-6), asociality (0–12), avolition (0–12), blunted affect

[3] `https://www.everythingals.org/research`

[4] The diadochokinetic task, usually abbreviated DDK, is one in which patients are asked to produce repeated syllables at a maximum rate of production.

(0-18), and alogia (0-12). The BNSS score can range from 0 (no negative symptoms) to 78 (maximum negative symptoms on all subscales).

As before, a session was chosen at random for six male and four female participants[5] with the highest total BNSS scores. Selected BNSS scores ranged from 33-61. The total number of patient turns in this cohort was 130. Each participant turn is a response to one of seven tasks: sustained vowel phonation and diadochokinetic syllables; read short sentences and a longer paragraph; short and long spontaneous speech; and description of a picture.

3.4 Depression

Data from patients in this cohort were obtained in partnership with Clarigent Health via a study in which all patients had a clinical diagnosis of depression. At the time of this writing, we did not have a sex-balanced cohort for the clinical depression cohort. In total, this cohort comprises 123 turns from 1 male and 9 female patients. Each participant turn is a response to one of 7 tasks: 1 read short sentence and 6 open-ended questions asking about how the user is feeling and their thoughts about anger, emotional pain, fear, hope, and secrets. Recruitment for the study was done via ResearchMatch, a national health volunteer registry that was created by several academic institutions and supported by the U.S. National Institutes of Health as part of the Clinical Translational Science Award program.

4 Methods

4.1 VAD Performance Measures

We employed the standard NIST Detection Cost Function (DCF) (20) to measure how well the CMU Sphinx VAD predictions were, given a set of values for the configurable parameters described in Section 2. The DCF score is a weighted penalty of the proportion of false positive and false negative time, when compared to a hand annotation of actual speech in an audio stream. Since ignoring true speech is usually most detrimental to a spoken dialog system, DCF traditionally penalizes false negatives more than false positives. Refer to Figure 1 for a visual aid in our discussion of the four possible outcomes of a VAD prediction when compared to a reference hand annotation. True negative time (TN) is the time when the VAD algorithm predicted no speech and the user was not speaking. True positive time (TP) is the time when the VAD algorithm predicted speech and the user was speaking. False negative time (FN) is the time when the VAD algorithm predicted no speech but the user was speaking. False positive time (FP) is the time when the VAD algorithm predicted

[5] There were not enough female participants to select five.

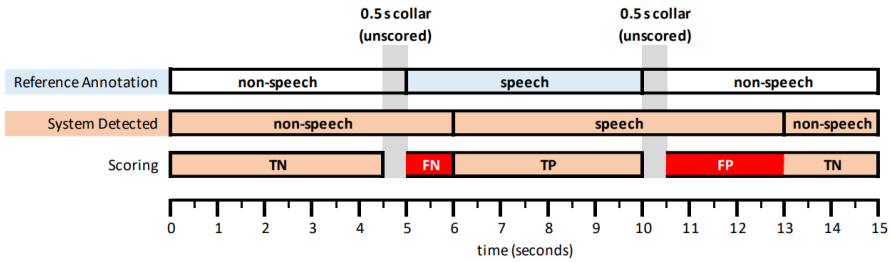

Fig. 1: Segmentation of hypothetical speech and VAD output of the same audio stream. The third tier shows the four possible outcomes used for scoring. This figure is reprinted from the original publication (20).

speech but the user was not speaking. Additionally, the calculation of DCF can take into account a "collar" of time both preceding and following the user speech which is not factored into the false negative or false positive times. Figure 1 shows a collar length of 0.5 seconds, though all the results presented herein use no collar. DCF is computed as follows:

$$P_{FP} = \frac{\text{total FP time}}{\text{annotated total non-speech time}}$$

$$P_{FN} = \frac{\text{total FN time}}{\text{annotated total speech time}}$$

$$DCF = 0.75 \times P_{FN} + 0.25 \times P_{FP}$$

Though we used DCF score to decide on optimal performance, in this paper we also report on additional metrics that are of interest due to the fact that they are discrete negative dialog events that are very detrimental to user satisfaction when conversing with an automated agent. We report on the interruption rate ($I\%$) as the number of turns in which the participant was interrupted by the automated agent / total participant turns $\times$ 100; turn false accept percentage ($FA\%$) as the number of turns in which the participant did not say anything but the automated agent thought they did / total participant turns $\times$ 100; and turn false reject percentage ($FR\%$) as the number of turns in which the participant did say something but the automated agent thought they did not / total participant turns $\times$ 100. For visual consistency, we also report on the the variables that comprise DCF, but expressed as percentages in the following manner: $FN\% = P_{FN} \times 100$ and $FP\% = P_{FP} \times 100$.

4.2 Simulation Experiments

We conducted offline simulated VAD experiments on annotated participant sessions with the aim of discovering the optimal configuration settings for the most accurate spoken turn detection. We chose a parameter space that amounted to 24,000

different VAD configurations. The bounds of this space were chosen empirically based on values that yielded successful past VAD performance and were sampled from within the following parameter values ranges: `endSilence`: 500-8000, `adjustment`: 0.0001, `threshold`: 30-40, `minSignal`: 0-20, `startSpeech`: 20-200, `initialSilence`: 2000-8000. For each offline simulation run, we chose a specific value set from within this feature space. We then split each session into discrete user turns and sent each turn through the VAD algorithm in order to obtain the VAD start and end time, if any. If more than one VAD event was detected, we only considered the first one since this event would end the turn in a deployed dialog system. We then computed DCF scores for each of these simulated runs and observed VAD configuration parameter values that optimized DCF.

5 Results

Table 1 shows the optimal VAD configurations for each cohort. We obtained two baseline VAD performances with which to compare the optimal cohort-only performances. The first results were obtained from optimal VAD settings for a control group, as published in (14). This corpus of 906 dialog turns was collected from individuals who suffered from no neurological or cognitive diseases. Table 2 shows the VAD performance per cohort when using VAD settings optimized on this control cohort, which was reported to have a DCF score of 0.043 and interruption percentage ($I\%$) of 5.91%. The last row in Table 1 lists the VAD settings used in this case.

To obtain another baseline of VAD performance, we ran a simulation experiment on all data combined across disease cohorts. The optimal VAD parameter values found can be seen in the penultimate row of Table 1 (*Combined*). Table 3 shows the VAD performance of each cohort using these baseline VAD parameters settings. The most drastic change in optimal VAD settings is a doubling of the `endSilence` time. We also see that DCF was reduced by an order of magnitude and that $I\%$ and percentage false negative time ($FN\%$) were significantly reduced across the board when moving from control-optimized settings to those of our disease cohorts.

Next, we ran a simulation experiment for each of the cohorts separately. The VAD performance using optimal parameter values for each cohort is shown in Table 4 and the optimal VAD configurations are shown in Table 1. We observe a further reduction in DCF, $I\%$, and $FN\%$, especially with respect to the depression cohort. We see, however, that the percentage of false positive speech ($FP\%$) increases. This is actually a desirable trade off since interrupting the user is worse that waiting a bit too long to end the VAD[6]; this relationship is encoded in DCF algorithm. Furthermore, `endSilence` once again emerged as the most variable VAD parameter setting. The optimal duration of non-speech time to ensure that the user turn was complete was 3,200ms for all cohorts combined. This is now shown to be a splitting of the difference between two pairings of cohorts: for both the Parkinson's and ALS

[6] Waiting too long to end VAD is what contributes to $FP\%$.

Table 1: Optimal VAD configuration per pathological cohort. Also shown, for reference, are configurations for all cohorts combined and healthy controls (as reported in (14), though here the sample size is larger that for the other cohorts).

Cohort	`initialSilence`	`endSilence`	`threshold`	`adjustment`	`minSignal`	`startSpeech`
Parkinson's	7500	2600	40	0.0001	0	100
ALS	7500	2600	32	0.0001	13	60
Depression	7500	4100	33	0.0001	19	60
Schizophrenia	8000	4500	32	0.0001	15	40
Combined	7500	3200	38	0.0001	5	20
Control	7500	1600	34	0.0001	6	140

Table 2: Baseline VAD performance per cohort using parameter settings optimized for control speakers on a different data set (14) using VAD parameter settings on last line of Table 1.

Cohort	*FA%*	*FR%*	*I%*	*FN%*	*FP%*	*DCF*
Parkinson's	0.00	0.61	5.81	14.23	2.00	0.1117
ALS	0.00	0.00	11.11	14.50	0.88	0.1110
Depression	0.00	0.00	39.02	29.80	0.32	0.2243
Schizophrenia	0.00	3.08	24.62	32.25	0.38	0.2428

Table 3: Baseline VAD performance per cohort using parameter settings optimized over all cohorts combined. Shaded cells indicate a better performance metric value compared to those in Table 2.

Cohort	*FA%*	*FR%*	*I%*	*FN%*	*FP%*	*DCF*
Parkinson's	0.00	0.31	1.22	2.30	7.63	0.0363
ALS	0.00	0.00	0.85	0.98	3.10	0.0151
Depression	0.81	0.00	10.57	5.07	0.63	0.0396
Schizophrenia	0.00	3.85	3.08	5.20	1.50	0.0427

Table 4: VAD performance per cohort using parameter settings optimized for the data in that cohort. Shaded cells indicate a better performance metric value compared to those in both Table 2 and Table 3.

Cohort	*FA%*	*FR%*	*I%*	*FN%*	*FP%*	*DCF*
Parkinson's	0.00	0.00	1.22	2.09	5.51	0.0295
ALS	0.00	0.00	0.00	0.67	2.21	0.0105
Depression	0.00	0.00	1.63	0.67	1.61	0.0090
Schizophrenia	0.00	3.08	0.00	2.95	6.12	0.0375

Table 5: Average optimal VAD configuration per cohort using leave-one-user-out cross validation.

Cohort	`initialSilence`	`endSilence`	`threshold`	`adjustment`	`minSignal`	`startSpeech`
Parkinson's	7500 ± 00.0	2560 ± 30.6	39.9 ± 0.1	0.0001	0.0 ± 0.0	92.0 ± 8.0
ALS	7500 ± 00.0	2630 ± 30.0	32.0 ± 0.0	0.0001	13.0 ± 0.0	38.0 ± 2.0
Depression	7500 ± 00.0	3790 ± 34.8	33.0 ± 0.0	0.0001	18.6 ± 0.4	78.0 ± 2.0
Schizophrenia	7900 ± 66.7	4410 ± 90.0	32.0 ± 0.0	0.0001	15.4 ± 0.3	40.0 ± 0.0

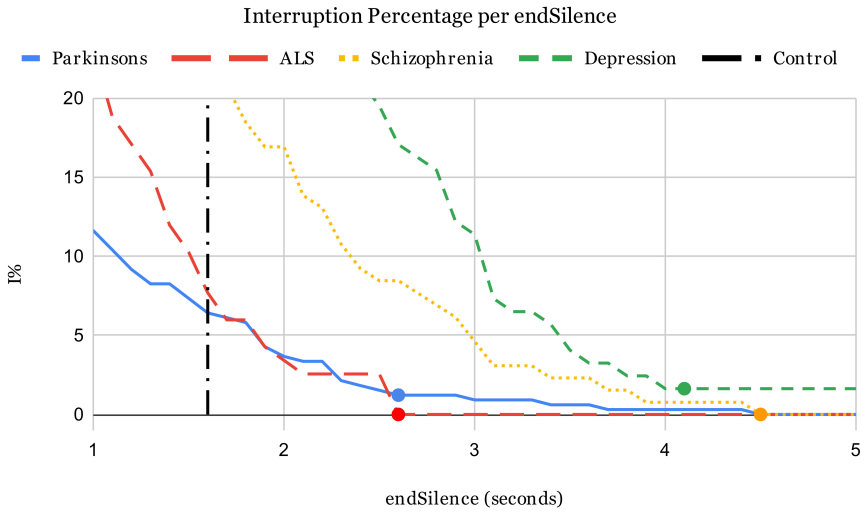

Fig. 2: Percent interrupted turns per domain plotted against `endSilence`. The circles on the graph indicate the lowest DCF score for each cohort and thus the optimal setting. Also shown is the optimal `endSilence` of 1.6 seconds computed for healthy controls (as seen in Table 1).

cohorts the optimal value was much shorter at 2,600ms; whereas, for the depression and schizophrenia cohorts the optimal values were much higher at 4,100ms and 4,500ms, respectively.

Figure 2 shows how *I%* changes as `endSilence` increases using data from the cohort-specific tuning simulations. A figure such as this could be done for all VAD parameters, but we only present this one because interruptions are usually of highest interest to users and developers of automated dialog systems and because `endSilence` shows the most striking differences across cohorts. The reason why the optimal *I%* is not 0% for all cohorts (see, Parkinson's) is because *DCF* starts to increase again at the point at which waiting too long for additional user speech (that never comes) begins to increase false positive time.

It is important to point out that *DCF* is a good metric for comparing VAD settings within a data set, but not for comparing across data sets. This is because *DCF* depends on the ratio of speech and non-speech lengths in a data set. These ratios will differ across data sets and this will be reflected in *DCF* values without actually representing a performance improvement or reduction. This is why we have also reported performance w.r.t. interruption rate; in addition to being a disruptive event in a conversation, it is a metric that can be used to compare performance across data sets.

In addition to the self-tuning simulation results already presented, we also ran cohort-specific 10-fold cross validation tests in which we found optimal VAD pa-

rameter settings for 9 users in each cohort and then tested VAD performance on the one held out user. Table 5 lists the average optimal settings and standard error over the 10 training sets and show little variation from the results we have already presented on self-tests in Table 1. This is a sign that the findings presented herein are robust.

6 Discussion

The aim of this research was to investigate optimal VAD parameter settings, as measured by the NIST detection cost function (DCF), for pathological speech collected by a multimodal dialog system from severe presentations of four conditions: ALS, Parkinson's disease, schizophrenia, and depression. This can allow us to extend the utility of such systems to be truly generalizable across neurological and mental health conditions, and deal with speech patterns across multiple conditions, especially where patients are severely affected. We observed clear evidence that one setting in particular—`endSilence`—differed based on pathology type. Across our cohort data, the value of this parameter should be set higher than for healthy speakers (1.6 seconds); specifically, it was better to wait for 2.6 seconds of non-speech before deciding the participant turn had ended in the cases of the ALS and Parkinson's patients, whereas for schizophrenic and depressed participants it was better to wait even substantially long longer than that: 4.5 and 4.1 seconds, respectively. Furthermore, we observed that these findings are robust via cross validation. Though all cohorts can suffer from both speech motor and cognitive impairment, ALS and Parkison's patients tend to exhibit more of the former symptoms and schizophrenic and depressed patients more of the latter. This was true for the data presented in this paper. We believe that for for schizophrenic and depressed patients, and possibly any condition that affects cognitive processing, the optimal time to wait before considering a patient turn as completed should take into account longer phrase-internal pause times symptomatic of cognitive processing impairment.

We did not see the same stratification in cohorts with respect to the `initialSilence` parameter, though the schizophrenia cohort stood out. In addition to the longest `endSilence` time, the optimal VAD performance for this cohort had the highest `initialSilence` value as well, at 8 seconds. As a reminder, `initialSilence` is the maximum amount of time to give the participant to speak at the beginning of an utterance. If it is the case that these parameter values are tuned to pausing behavior indicative of cognitive impairment, then this would suggest that schizophrenic patients have higher impairment than depressed patients and VAD settings should account for this.

The cohort-specific optimal values of the remaining CMU Sphinx VAD parameters—`threshold, adjustment, minSignal, startSpeech`—do not seem to generalize well. We believe this is because they are parameters that address acoustic environmental conditions. Such environmental conditions are difficult to replicate across studies and applications; and indeed, our data did not have the exact same

recording conditions across cohorts. Additionally, these parameters are particular to CMU Sphinx's VAD algorithm and these settings would therefor not be useful for researchers using a different VAD algorithm. However, we believe that the general paradigm of tuning VAD configuration parameters using extensive simulations should be conducted and does generalize to any application and domain. While the exact values of this type of VAD parameter will differ, the optimal values should be discovered for the recording and environmental conditions of the particular application and pathological cohort.

In sum, this paper has presented the first investigation of pathology-specific VAD settings for digital health dialog agents in the literature, to our knowledge. We simulated the performance of a VAD system optimized across severe presentations of four pathological speech cohorts—ALS, Parkinson's disease, schizophrenia, and depression—and found two optimal modes of operation based on optimal `endSilence` values for speech motor conditions and mental health conditions, respectively. This paper has argued that the estimation of such optimal VAD settings are essential for remote monitoring systems, not just to improve accessibility and user experience (UX) of such systems for disordered populations (and therefore continued buy-in), but also because they are the gateway into collecting proper data necessary for all other aspects of later analysis. For example, if an utterance-internal pause is detected as an utterance-final pause then the system interrupts the user (poor user experience), while also adversely impacting all pause-related metrics. On the other hand, if the VAD waits too long to end, the participant may think that the system did not hear them, leading to repetitions. This in turn could negatively impact all metrics that rely on expected utterance durations and word counts. Going forward, this analysis paves the road for several planned future directions: one, towards building a truly generalizable virtual health agent that can adapt to the speech patterns of participants across a range of neurological and mental health conditions; two, towards informing multimodal user interface and UI/UX designs to better adapt to severely pathological speakers; and three, towards investigating the utility and robustness of multimodal VAD (integrating both face and speech information) to improve the robustness of such systems even further.

7 Acknowledgements

We would like to thank the following people and organizations for all their help and collaborative effort in collecting the various datasets described in this paper: Jessica Huber, Sandy Snyder, and Andrew Exner at Purdue University for the Parkinson's data; EverythingALS and the Peter Cohen Foundation for the ALS data; Anzalee Khan, Christian Yavorsky, Sebastian Prokop, and Jean-Pierre Lindenmayer at the Nathan Kline Institute for the schizophrenia data; and Josh Cohen and David Black at Clarigent Health for the depression data.

References

[1] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, Paris, France, May 2014.

[2] C. Lisetti, R. Amini, and U. Yasavu, "Now all together: Overview of virtual health assistants emulating face-to-face health interview experience," *KI-Künstliche Intelligenz*, vol. 29, pp. 161–172, March 2015.

[3] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill *et al.*, "Nemsi: A multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 245–247.

[4] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[5] R. D. Kent and Y.-J. Kim, "Toward an acoustic typology of motor speech disorders," *Clinical linguistics & phonetics*, vol. 17, no. 6, pp. 427–445, 2003.

[6] I. Kodrasi and H. Bourlard, "Spectro-temporal sparsity characterization for dysarthric speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1210–1222, 2020.

[7] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a biomarker: Opportunities, interpretability, and challenges," *Perspectives of the ASHA Special Interest Groups*, pp. 1–8, 2022.

[8] D. Robert, J. Pouget, A. Giovanni, J.-P. Azulay, and J.-M. Triglia, "Quantitative voice analysis in the assessment of bulbar involvement in amyotrophic lateral sclerosis," *Acta oto-laryngologica*, vol. 119, no. 6, pp. 724–731, 1999.

[9] J. Lee, E. Dickey, and Z. Simmons, "Vowel-specific intelligibility and acoustic patterns in individuals with dysarthria secondary to amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 1, pp. 34–59, 2019.

[10] J. R. Green, D. R. Beukelman, and L. J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech," *Journal of medical speech-language pathology*, vol. 12, no. 4, p. 149, 2004.

[11] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Sci Rep*, vol. 7, no. 1, p. 12, Feb. 2017.

[12] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[13] N. C. Andreasen and S. Olsen, "Negative v positive schizophrenia: Definition and validation," *Archives of general psychiatry*, vol. 39, no. 7, pp. 789–794, 1982.

[14] J. Liscombe, H. Kothare, M. Neumann, A. Ocampo, O. Roesler, D. Habberstad, A. Cornish, D. Pautler, D. Suendermann-Oeft, and V. Ramanarayanan, "Voice activity detection considerations in a dialog agent for dysarthric speakers," in *Proceedings of the 12th International Workshop on Spoken Dialog System Technology*, ser. IWSDS'21. Berlin, Heidelberg: Springer-Verlag, 2021.

[15] H. Kothare, V. Ramanarayanan, O. Roesler, M. Neumann, J. Liscombe, W. Burke, A. Cornish, D. Habberstad, A. Sakallah, S. Markuson, S. Kansara, A. Faerman, Y. Bensidi-Slimane, L. Fry, S. Portera, D. Suendermann-Oeft, D. Pautler, and C. Demopoulos, "Investigating the interplay between affective, phonatory and motoric subsystems in autism spectrum disorder using a multimodal dialogue agent," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (Interspeech)*, Brno, Czech Republic, August-September 2021.

[16] C. Baylor, K. Yorkston, T. Eadie, J. Kim, H. Chung, and D. Amtmann, "The communicative participation item bank (CPIB): item bank calibration and development of a disorder-generic short form," *J Speech Lang Hear Res*, vol. 56, no. 4, pp. 1190–1208, Jul. 2013.

[17] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, D. Pautler, I. Navar, A. Anvar, J. Kumm, R. Norel, E. Fraenkel, A. V. Sherman, J. D. Berry, G. L. Pattee, J. Wang, J. R. Green, and V. Ramanarayanan, "Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (Interspeech)*, Brno, Czech Republic, August-September 2021.

[18] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. complete listing of the BDNF Study Group *et al.*, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[19] B. Kirkpatrick, G. P. Strauss, L. Nguyen, B. A. Fischer, D. G. Daniel, A. Cienfuegos, and S. R. Marder, "The brief negative symptom scale: psychometric properties," *Schizophrenia bulletin*, vol. 37, no. 2, pp. 300–305, Mar 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20558531

[20] F. R. Byers, "NIST Open Speech Analytic Technologies Evaluation OpenSAT 2019," 2019.