Proceedings of the RaPID-4 @LREC 2022, pages 56–61
Marseille, 25 June 2022

Post-Stroke Speech Transcription Challenge (Task B): Correctness Detection in Anomia Diagnosis with Imperfect Transcripts

Trang Tran
University of Southern California, Institute for Creative Technologies
Los Angeles, CA, USA
ttran@ict.usc.edu

Abstract

Aphasia is a language disorder that affects millions of adults worldwide annually; it is most commonly caused by strokes or neurodegenerative diseases. Anomia, or word finding difficulty, is a prominent symptom of aphasia, which is often diagnosed through confrontation naming tasks. In the clinical setting, identification of correctness in responses to these naming tasks is useful for diagnosis, but currently is a labor-intensive process. This year's Post-Stroke Speech Transcription Challenge provides an opportunity to explore ways of automating this process. In this work, we focus on Task B of the challenge, i.e. identification of response correctness. We study whether a simple aggregation of using the 1-best automatic speech recognition (ASR) output and acoustic features could help predict response correctness. This was motivated by the hypothesis that acoustic features could provide complementary information to the (imperfect) ASR transcripts. We trained several classifiers using various sets of acoustic features standard in speech processing literature in an attempt to improve over the 1-best ASR baseline. Results indicated that our approach to using the acoustic features did not beat the simple baseline, at least on this challenge dataset. This suggests that ASR robustness still plays a significant role in the correctness detection task, which has yet to benefit from acoustic features.

Keywords: anomia, psst challenge, stroke, aphasia, automatic speech recognition

1. Introduction

Aphasia is a language disorder that affects 2–4 million people annually just in the US alone.[1] Aphasia most commonly occurs after a stroke or head injury, or can be acquired slowly from growing brain tumors or neurological diseases.[2] Patients with aphasia suffer difficulty in communication, which can manifest as various forms of language impairments, including both comprehension and expression.

One of the most prominent symptoms of aphasia is *anomia*, or word finding difficulty. Specifically, aphasia patients with anomia might make word production errors that are semantic (e.g. “dog” for the target “cat”), phonological (e.g. “tat” for the target “cat”), both, or even unrelated (e.g. “chair” for the target “cat”). These errors are typically diagnosed in the clinical setting through confrontation naming tasks, where the patient is presented with hundreds of items to identify/name. The resulting error profiles are then analyzed by professionals to provide overall assessment. Understanding these errors is therefore critical in diagnosis as well development of treatment plans.

However, current approaches for anomia test assessments are labor intensive for clinicians, especially with a large number of patients, each completing a large set of tests. Further, speech recognition for atypical speech, such as that produced by aphasia patients, is especially challenging, since most state-of-the-art automatic speech recognizers (ASR) were trained on clean (and often read) speech in controlled environments.

Recently, self-supervised speech representation approaches (Liu et al., 2020a; Liu et al., 2020b; Baevski et al., 2020), commonly learned from raw audio, have shown promising results on multiple tasks. Their utility has been evaluated on a range of spoken language processing tasks, from word/phoneme recognition to emotion and sentiment analysis (Yang et al., 2021; Shon et al., 2021). The natural question is then whether these systems can be adapted to aphasic speech, especially when the aphasia data is recorded in conditions often much different from the pretrained ASR data. In this work, however, we take a more incremental approach in assessing the possibility of detecting anomia with a simple combination of pretrained ASR output and acoustic features. This approach is inspired by the earlier works showing the utility of prosody (i.e. *how* something is said vs. *what* is said) in aiding spoken language understanding systems, both when applied to hand transcripts and ASR transcripts (Kahn and Ostendorf, 2012; Marin and Ostendorf, 2014; Tran et al., 2019; Tran and Ostendorf, 2021). In particular, we focus on Task B: correctness prediction of naming responses in the Post-Stroke Speech Transcription Challenge (PSST) 2022. We aim to answer the following questions:

- Using a pretrained ASR system, can correctness prediction be improved using acoustic features?
- Are there salient differences in the acoustic patterns of correct vs. incorrect naming responses?

In answering these questions, we hope to understand whether such a simple and low-cost system (i.e. not

[1] `https://www.aphasiaaccess.org/white-papers/`

[2] `https://www.nidcd.nih.gov/health/aphasia`

requiring additional aphasia-specific data and annotations) helps predict response correctness, or whether it is worth investing more effort in improving ASR for the domain of aphasic speech and language disorders in general.

2. Related Work

Many researchers have explored the potential of using speech for the diagnosis of language disorders. For example, Roark et al. (2011) showed that both lexical and acoustic signals can help detect mild cognitive impairment (MCI). In particular, noun and verb counts, syntactic complexity (as measured by Yngve score (Yngve, 1960)), pause durations and pause rates seemed to be most useful. For Primary Progressive Aphasia (PPA) detection and subtype classification, Fraser et al. (2013; Fraser et al. (2014) also found that syntactic complexity features were among the most useful. In addition, while acoustic features were not as useful in differentiating PPA from control, they were important in classification of PPA's subtypes.

In an investigation to push towards a fully automated diagnosis pipeline, Zhou et al. (2016) compared using hand-transcribed speech conversations vs. ASR outputs to detect Alzheimer's disease in participants. Not surprisingly, they found that accuracy is higher using perfect transcripts, but also identified key features that have distinguishing power in both gold and ASR transcripts, such as word length and frequency. In addition, the authors observed that accuracies can vary within a narrow band of word error rates (WER), i.e. ASR transcripts with the same low WER can contain drastically different information. For predicting aphasia quotient (AQ), Le et al. (2018) trained a speech recognition system on AphasiaBank (MacWhinney et al., 2011) and achieved a new recognition benchmark for ASR in aphasic speech, in addition to obtaining higher accuracy on AQ prediction.

The research so far has largely been limited by ASR quality, as aphasic speech proves to be a challenge. However, to the best of our knowledge, little has been explored on whether acoustic features are informative in aiding correctness prediction on top of ASR transcripts.

3. Data and Metrics

The dataset we use is provided by the PSST Challenge 2022 organizers (Gale et al., 2022). In particular, the dataset is a subset of AphasiaBank (MacWhinney et al., 2011), a database of multimedia interactions in clinical settings for the study of aphasia. For the PSST challenge, the subset includes responses from the Boston Naming Test – Short Form (BNT) and the Verb Naming Test (VNT). In addition to the audio and metadata from AphasiaBank, Gale et al. (2022) provided human phone-level annotations, as well as the correctness label for the naming responses (i.e. whether the utterance was considered correct by clinicians). The dataset is well-balanced with approximately 50%:50% split of correct vs. incorrect labels (binary classes), both in the training and validation set. The train/validation/test splits were predefined by the challenge organizers. Overall dataset statistics is shown in Table 1.

Split	# Utterances	PER	FER
Train	2298	4.0%	2.4%
Validation	341	22.6%	10.6%
Test	652	n/a	n/a

Table 1: Dataset Statistics for the PSST Challenge

For ASR, we use the pretrained system provided by Gale et al. (2022), and obtained phone transcripts from this off-the-shelf ASR. The phone error rate (PER) and feature error rate (FER) are also reported for each set. PER is a standard metric in ASR research (i.e. % phone recognition errors out of reference phones); FER is a metric provided by the challenge organizers that emphasizes evaluation of errors regarding *distinctive phone features* (i.e. putting more value on transcripts that *sound* correct as opposed to strict comparison with phone representations).

For correctness prediction, we use standard evaluation metrics for binary classification, i.e. F1 score (in addition to reporting precision and recall), as instructed by the organizers (Gale et al., 2022).

4. Methods

4.1. Acoustic Features

Inspired by previous works exploring acoustic features for aphasia classification, we extracted several feature sets reported in literature to be generally useful in speech analysis.

- Librosa (McFee et al., 2015) feature set: we extract the pitch contour for each utterance using librosa's implementation of the pYIN algorithm (Mauch and Dixon, 2014; de Cheveigné and Kawahara, 2002). This gives us the estimated pitch contour, as well as voice activity detection per frame. To summarize the pitch contour and voicing characteristics for the whole utterance, we compute the voice activity rate (active_rate) for each utterance, which we consider the proxy for pause characteristics of the utterance. Pause features have been shown to be useful in acoustic analysis of speech disorders, e.g. as in (Roark et al., 2011; Le et al., 2018). Additionally, we hypothesize that pauses are important indicators of speech fluency, i.e. aphasic speech might be less fluent than healthy speech due communication difficulties reflected by hesitations and self-corrections.

  To potentially alleviate the loss of acoustic information in summarizing features for the whole utterance, we also estimate the polynomial fit coefficients of the pitch contour. We used a 5th order polynomial fit, resulting in a six dimensional feature vector for each utterance, i.e. the coefficients $[a_5, a_4, a_3, a_2, a_1, a_0]$.

- The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing (Eyben et al., 2010): We extract the low-level descriptors as recommended in (Eyben et al., 2010); this gave us 18 features that cover different pitch, energy, and spectral balance characteristics of the speech utterances. Detailed descriptions for each feature can be found in Eyben et al. (2010). For each feature in this set, we compute the mean and standard deviation of each utterance.

In addition to acoustic features, we explore potentially using ASR scores for the utterances as a proxy of how confident the speech recognizer was. We hypothesize that lower confidences could potentially indicate anomalies in the speech patterns and thus could inform correctness in the naming task. For this, we use the min, max, mean, median, and standard deviation of the softmax normalized logit scores generated by the pretrained ASR system. Specifically, the logit scores were first normalized to sum up to 1 before the sufficient statistics calculations.[3] We did not excluded silent or pad tokens in this work (a possible future tweak), and this was only a simple way to assess the *global* ASR confidence for each utterance.

To select the potentially most useful features for discriminating between correct and incorrect responses, we perform a t-test for each feature between the correct and incorrect samples in the training data. Features that have statistically significant differences ($p < 0.001$, using Bonferroni correction) in correct vs. incorrect samples are the following (henceforth referred to as CoreFeats):

- max_logit: max value of the (normalized) logit scores in each utterance

- mean_logit: mean value of the (normalized) logit scores in each utterance

- mean_Loudness_sma3 (GeMAPS feature): mean value of loudness in each utterance, i.e. mean estimate of perceived signal intensity from an auditory spectrum

- sd_Loudness_sma3 (GeMAPS feature): standard deviation of loudness in each utterance

- mean_spectralFlux_sma3 (GeMAPS feature): mean value of spectral flux in each utterance, i.e. the mean difference of the spectra of two consecutive frames

- sd_spectralFlux_sma3 (GeMAPS feature): standard deviation of spectral flux in each utterance

Interestingly, none of the librosa features were significantly different between correct and incorrect samples. This is surprising since previous work has shown pauses are a useful indicator, but the feature active_rate is not among CoreFeats according to our selection heuristics.

4.2. Classifiers

Our baseline model is a simple string matching procedure as implemented by Gale et al. (2022), i.e. we use the 1-best ASR output and run the program to evaluate whether the transcript is found among acceptable pronunciations. This baseline output also is chosen to be our "base" feature, i.e. a binary feature indicating whether a correct pronunciation is found in the ASR transcripts.

We experimented with all acoustic features listed in Section 4.1. In particular, our classifiers were trained on all the subsets of features listed, as well as those selected through the statistical significance test above, i.e. CoreFeats.

We explored two types of standard classifiers, since the dataset is relatively small: logistic regression (LR) and support vector machine (SVM). Hyperparameter search included the regularization coefficient $C \in [10^{-4}, 10^{-3}, ..., 10^{4}]$ for both LR and SVM, and we additionally experimented with both linear and RBF kernels for the SVM. We use cross validation with 5 folds in the training set to select the hyperparameters. Our models were implemented using the Scikit-learn toolkit (Pedregosa et al., 2011).

5. Results and Discussion

The baseline model (using string match on 1-best ASR output) turned out to be a very strong baseline. All our configurations without using this baseline (i.e. using only acoustic features) yielded very poor results, often comparable to random guessing (F1 $\approx$ 0.5). Results from experiments with all different combinations of $\{$Librosa, GeMAPS, logit$\}$ features as described in Section 4.1 all showed similarly poor performances.

Using only CoreFeats did slightly better than random, but combining CoreFeats with the baseline indicator does not beat simply using the baseline. In fact, the predicted outputs from Baseline and Baseline+CoreFeats were identical.

Table 2 shows the best results with SVM (linear kernel, $C = 0.01$).

On the final test set, our best-performing classifier (Baseline+CoreFeats) obtained F1 score = 0.89 (precision = 0.93, recall = 0.86) and accuracy = 0.90. This result is similar to those on the validation set, likely thanks to similarly balanced data distributions.

[3] Raw "logit scores" are a bit of a misnomer since they are usually not normalized to sum up to 1 for general purposes, e.g. in inference.

Model	Precision	Recall	F1
Baseline	0.92	0.81	0.86
CoreFeats only	0.64	0.59	0.61
Baseline+CoreFeats	0.92	0.81	0.86

Table 2: Results of Classification on the Validation Set

To diagnose our results, we looked specifically at the set of samples where the results from our CoreFeats-only classifier differ from those using Baseline. Our motivation is to see whether particularly difficult samples, i.e. those Baseline got wrong, had any indicators that the acoustic features might have identified.

In both training and validation sets, using only CoreFeats (without baseline) performed better on the VNT set compared to BNT. Specifically, out of 1467 utterances in the training set where CoreFeats obtained correct predictions, 1023 are from VNT while 444 are from BNT. Similarly for the validation set, out of 214 correct predictions by CoreFeats, twice as many are from VNT than BNT (143 vs. 71). This pattern persists even when looking into the subset where CoreFeats managed to predict correctly those Baseline predicted incorrectly. In the validation set, while CoreFeats performed better than Baseline for only 15 utterances, only 2 are from BNT while the rest are from VNT. Anecdotally (from listening to a few samples), we observed that the BNT task involves isolated word naming while VNT elicits potentially longer, more sentence-like speech to include the verb being tested. We hypothesize that this is where acoustic features are likely more useful, as these longer speech samples exhibit more diverse prosodic phenomena easier to model by acoustic features (Tran, 2020).

Figures 1 and 2 show the histograms of subset of samples where the outputs of Baseline and CoreFeats classifiers differ. In the training set, it appears that acoustic features could potentially help identify additional true positives (correct naming responses). However, the majority of instances are correctly classified by Baseline, so it is not obvious that acoustic features could help in a significant way.

The similar analysis on the validation set shows a slightly different trend: here Baseline misses more incorrect responses, i.e. it failed to identify utterances with incorrect pronunciations/reading. Arguably this is the more interesting case where acoustic features should help: for example, while there might be a good string match between the ASR transcript and the true transcript, the acoustic characteristics of the utterance might help flag these as incorrect responses to help reduce misdiagnosis. However, again, the majority of cases are still correctly classified using Baseline.

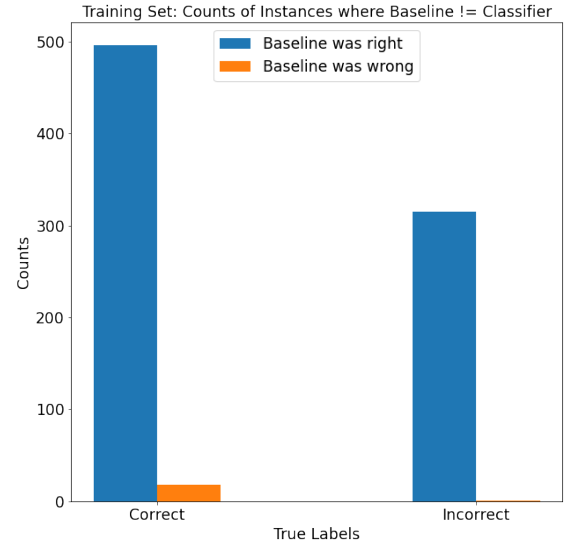

Figure 1: Distribution of samples where Baseline predictions are different from CoreFeats predictions; Training set.

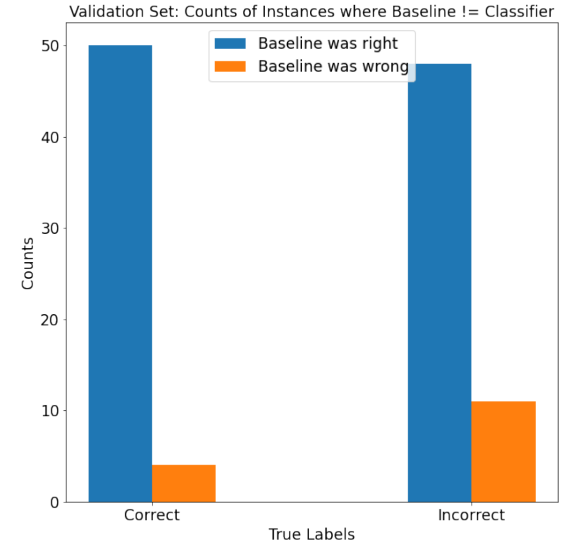

Figure 2: Distribution of samples where Baseline predictions are different from CoreFeats predictions; Validation set.

This difference in behavior between the training and validations sets, coupled with the large difference in PER and FER as shown in Table 1, suggests that the pretrained ASR system might have overfitted on the training set.

6. Potential Next Directions

Our first attempt at a simple system to classify correctness of naming responses in anomia diagnosis has yielded negative results so far. Specifically, the challenge seems to be two-fold: (1) acoustic feature selection and (2) over-reliance on robust ASR.

Regarding acoustic feature selection, it is largely unclear how to select the best set of features, despite a large amount of study dedicated to this area. Using acoustic features in this setting is also difficult both from the modeling (how to aggregate frame-level features to the utterance level representation) and the data quality (which features are robust to recording noise, dialects, age, etc.) perspectives. The dataset in this challenge is quite small, and the acoustic feature space is large. Perhaps redoing this feature analysis on a larger aphasia dataset might yield a different result.

Regarding ASR systems, the difference in both classification results and FER/PER between the training and validation sets highlights the difficulty in domain adaptation. One experiment we would have liked to try is to use several off-the-shelf pretrained ASR systems and devise heuristics for ensembling the results. For example, in addition to a Baseline as in this work, we could look at the differences in prediction and confidences of various ASR systems, and use these differences as another proxy the transcription quality.

Overall, from this small study, it appears that the robustness of ASR plays a more important role than acoustic feature exploration.

7. Conclusion

In this work, we focus on Task B: Correctness Evaluation of the PSST Challenge 2022. Our goal was to investigate whether using acoustic features in addition to ASR transcripts would improve correctness prediction. The motivation was that if acoustic features helped, this augmentation approach would only need a relatively good pretrained ASR system without further collecting costly annotations or additional data for fine-tuning ASR. Unfortunately, this was not the case, as our approach to using acoustic features could not improve over a simple baseline (string match between 1-best ASR output and acceptable pronunciations). However, we did find potential indicators of acoustic feature usefulness in tasks eliciting longer speech. Specifically, using acoustic features obtained better results in the verb naming test (VNT) than in the isolated noun naming test (BNT), likely because the former elicits longer, more sentence-like utterances.

Our results suggest that ASR robustness still plays critical role in this task, and that it is worth investing more effort in improving ASR for the domain of aphasic speech and language disorders in general.

8. Acknowledgements

We would like to thank the organizers of the PSST Challenge for initiating this effort and help raise awareness to this important problem in the speech and NLP community. We also thank the organizers for providing the data and baseline codes. We thank the anonymous reviewers for their helpful feedback.

9. Bibliographical References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

de Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111 4:1917–30.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia*, MM '10, pages 1459–1462, New York, NY, USA. Association for Computing Machinery.

Fraser, K. C., Rudzicz, F., and Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. In FrÃ©dÃ©ric Bimbot, et al., editors, *INTERSPEECH*, pages 2177–2181. ISCA.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43 – 60. Language, Computers and Cognitive Neuroscience.

Gale, R., Fleegle, M., Bedrick, S., and Fergadiotis, G. (2022). Dataset and tools for the PSST Challenge on Post-Stroke Speech Transcription, March. Project funded by the National Institute on Deafness and Other Communication Disorders grant number R01DC015999-04S1.

Kahn, J. G. and Ostendorf, M. (2012). Joint reranking of parsing and word recognition with automatic segmentation. *Computer Speech & Language*, 26(1):1–51.

Le, D., Licata, K., and Mower Provost, E. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100:1–12.

Liu, A. T., Li, S.-W., and yi Lee, H. (2020a). TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech.

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020b). Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology*, 25(11):1286–1307.

Marin, A. and Ostendorf, M. (2014). Domain adaptation for parsing in automatic speech recognition. In *Proc. ICASSP*, pages 6379–6383.

Mauch, M. and Dixon, S. (2014). Pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). "librosa:

Audio and music signal analysis in python". In *Proceedings of the 14th python in science conference*, pages 18–25.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Roark, B., Mitchell, M., Hosom, J., Hollingshead, K., and Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2081–2090, Sept.

Shon, S., Pasad, A., Wu, F., Brusco, P., Artzi, Y., Livescu, K., and Han, K. J. (2021). SLUE: new benchmark tasks for spoken language understanding evaluation on natural speech. *CoRR*, abs/2111.10367.

Tran, T. and Ostendorf, M. (2021). Assessing the use of prosody in constituency parsing of imperfect transcripts. In *Proc. Interspeech*, pages 2626–2630.

Tran, T., Yuan, J., Liu, Y., and Ostendorf, M. (2019). On the Role of Style in Parsing Speech with Neural Models. In *Proc. Interspeech*, pages 4190–4194.

Tran, T. (2020). *Neural Models for Integrating Prosody in Spoken Language Understanding*. Ph.D. thesis, University of Washington.

Yang, S.-W., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-T., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and Lee, H.-Y. (2021). SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

Zhou, L., Fraser, K. C., and Rudzicz, F. (2016). Speech recognition in Alzheimer's disease and in its assessment. *Interspeech 2016*, pages 1948–1952.