INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

An Utterance Verification System for Word Naming Therapy in Aphasia

David S. Barbera[1], Mark Huckvale[2], Victoria Fleming[1], Emily Upton[1], Henry Coley-Fisher[1], Ian Shaw[3], William Latham[4], Alexander P. Leff[1], Jenny Crinion[1]

[1]Institute of Cognitive Neuroscience, University College London, U.K.
[2]Speech, Hearing & Phonetic Sciences, University College London, U.K.
[3]Technical Consultant at SoftV, U.K.
[4]Goldsmiths College, University of London, U.K.

david.barbera.16@ucl.ac.uk

Abstract

Anomia (word finding difficulties) is the hallmark of aphasia an acquired language disorder, most commonly caused by stroke. Assessment of speech performance using pijcture naming tasks is therefore a key method for identification of the disorder and monitoring patient's response to treatment interventions. Currently, this assessment is conducted manually by speech and language therapists (SLT). Surprisingly, despite advancements in ASR and artificial intelligence with technologies like deep learning, research on developing automated systems for this task has been scarce. Here we present an utterance verification system incorporating a deep learning element that classifies 'correct'/'incorrect' naming attempts from aphasic stroke patients. When tested on 8 native British-English speaking aphasics the system's performance accuracy ranged between 83.6% to 93.6%, with a 10 fold cross validation mean of 89.5%. This performance was not only significantly better than one of the leading commercially available ASRs (Google speech-to-text service) but also comparable in some instances with two independent SLT ratings for the same dataset.

Index Terms: speech disorders, word naming, aphasia

1. Introduction

Word retrieval difficulties, or anomia, is the most pervasive symptom of post-stroke aphasia [1]. Recent data suggests there are around 350,000 people in the UK alone who have chronic aphasia post-stroke [2]. Despite the prevalence of aphasia, few patients receive a sufficient dose of speech and language therapy to recover maximally. For example, in the UK through the National Health Service patients receive on average 8-12 hours when the recommended dose to see a significant change is in the order of 100 hours [3]. Assessment of patients' spoken picture naming abilities and then practising repetitively over time a range of vocabulary using spoken picture naming tasks is an integral part of anomia treatment [4]. The intervention is primarily administered by a speech and language therapist (SLT), and the patient is confronted with a picture or drawing of an object to name. An Automated Speech Recognition system (ASR) that could reliably assess patient's speech performance on these picture naming tests would not only offer increased consistency and sensitivity to changes in patient's speech abilities but also enable patients to perform these tests independent of SLTs, potentially remotely away from the clinic in the comfort of their own home. This would not only 'free-up' clinicians to deliver more complex interventions in their 'face-to-face' time but also support more patients who are unable to travel into the clinic, a need which has become more pressing in light of recent COVID-19 travel restrictions.

1.1. ASR for aphasic's single word naming performance

Different to single and isolated spoken word recognition, assessing spoken picture naming performance has the advantage that the target word is known. Therefore, the challenge for ASR in this context is actually to verify that a particular target word is uttered in a given segment of speech [5]. Furthermore, an ASR-based system, or utterance verifier system, within a therapy app must immediately provide a binary response *'correct'/'incorrect'* feedback to the patient for each spoken naming attempt, often 1000s of trials repeatedly over time.

To the best of our knowledge, only two groups have used and assessed an ASR-based system of such type in aphasic's single word picture naming performance. In the project Vithea [6], researchers developed an aphasia treatment app for Portuguese speakers. Their in-house ASR-engine called AUDIMUS [7] using a keyword spotting technique to score spoken naming attempts as 'correct'/'incorrect' reported an average accuracy of 82%, with ranges between 69% and 93% across patients [5]. The second group [8] evaluated a digitally delivered picture naming intervention in native Australian English speaking people with apraxia plus aphasia. They used the open-source ASR engine CMU PocketSphinx [9] to provide patients with 'correct'/'incorrect' feedback. For 124 words, which were phonetically different, they reported an overall ASR accuracy of 80% and a range of scores between 65.1% and 82.8% across patients, depending on impairment severity. Both these systems provide useful 'proof-of-concept' data that ASR systems for anomia assessment are feasible. Still, the high error rate and variable performance across aphasic patients meant its clinical utility remained low.

This project aims to present and assess the feasibility of a tailor-made system incorporating a deep learning element to assess word naming attempts in people with aphasia. We will provide an open-access implementation of our system and trained models online for researchers, therapists and clinicians interested in adopting this approach.

2. An utterance verifier for word naming

Given the scarcity of speech corpora in aphasia, we used a template-based system for picture naming verification. We built on the framework developed by Ann Lee and James Glass in "A comparison-based approach to mispronunciation detection" [10]. Their ASR system was developed to detect

http://dx.doi.org/10.21437/Interspeech.2020-2265

word-level mispronunciations in non-native speech. It was initially designed to be language agnostic. It works by comparing a word uttered by a native speaker, or teacher, with the same word uttered by a non-native speaker or student. It relies on posteriorgram based pattern matching via a dynamic time warping (DTW) algorithm to compare the utterances. Our system replaced their Gaussian Mixture Model trained on unlabeled corpora with an acoustic model based on a deep neural architecture trained on English corpora from healthy speakers to generate phone-based posteriors. Then, similar to Lee's teacher-versus-student framework, we compare healthy-versus-aphasic utterances. We defined a posteriorgram as a vector of posterior probabilities over phoneme classes in the English language for which we employed the ARPAbet system as used in the BEEP dictionary [11] consisting of 45 symbols: 44 ARPAbet symbols plus silence. To enable future clinical utility of our system, we developed it to run embedded on mobile devices without sophisticated model compression techniques.

2.1. Signal pre-processing and acoustic modelling

Speech recordings were pre-processed in overlapping frames of 30 milliseconds every 10 milliseconds, and a fast Fourier transform size of 64 milliseconds after a pre-emphasis filter of 0.95 to obtain a vector of 26 acoustic features per frame: 12 Mel-frequency cepstral coefficients (with a final liftering step with a coefficient of 23 applied to them), energy and 13 deltas. See step 1 and 2 in Figure 1.

To train our acoustic model, we used a corpus of healthy British speakers WSJCAM0 [12] to match the native spoken language of our patients. WSJCAM offers phone-level transcriptions using the ARPAbet phone set for British English. We then used Keras deep learning framework [13] with TensorFlow [14] as the back-end. All our models used batch normalisation, dropout rate of 0.5 and a categorical cross-entropy over 45 classes as the loss function. Training lasted until there was no improvement in accuracy for 50 epochs. We explored several types and configurations of recurrent neural networks and choose our final model as the one with the lowest Phone Error Rate (PER) on the WSJCAM0 test set. Our winning model was a Bidirectional GRU [15] of 128 units and 7 layers of depth trained with the Adam optimiser [15] resulting in around 2 million parameters and achieving a segment-based phone error rate (PER) of 15.85%. See step 3 in Figure 1.

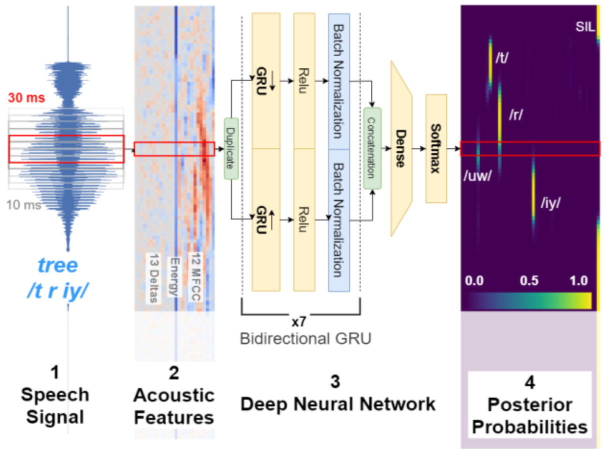

Figure 1. *From signal to posterior probabilities. Left to right: speech signal is fragmented into frames every 10 milliseconds of a window size of 30 milliseconds (1), from each frame a vector of acoustic features is extracted (2) then each vector is fed to a Deep Neural Network (3) which outputs a vector of posterior probabilities or posteriorgram (4).*

2.2. Comparison of utterances

Our system uses two recordings from native healthy speakers for each target word, which are transformed into posteriorgrams offline via our DNN, as shown in Figure 1 (steps 1-4). Each naming attempt by an aphasic speaker is transformed into posteriorgrams using our DNN and then compared to each of the posteriorgrams from the two healthy speakers via the DTW algorithm as in [10], see Figure 2. Adapting Lee's notation, given a sequence of posteriorgrams for the healthy speaker $H = (p_{h_1}, p_{h_2}, \dots, p_{h_n})$ and the aphasic speaker $A = (p_{a_1}, p_{a_2}, \dots, p_{a_m})$, a $n \times m$ distance matrix can be defined using the following inner product:

$$\varphi_{ha}(i,j) = -\log(p_{h_i} * p_{a_j}) \qquad (1)$$

For such a distance matrix, DTW will search for the path from $(1{,}1)$ to (n, m) that minimises the accumulated distance. Different from Lee's work, we used the minimum of the DTW accumulated distances for all comparisons with the two healthy speakers to make a final decision.

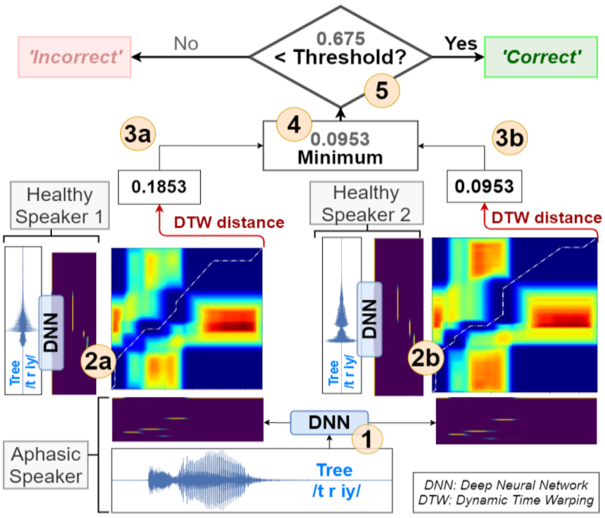

Figure 2. *An utterance verification system for word naming. Given a naming attempt, e.g. target word tree, the voice of an aphasic patient is recorded and processed through our DNN to generate posteriorgrams (1). The system keeps posteriorgrams of previously recorded healthy speakers' utterances for each target word, (2a and 2b). Posteriorgrams are compared using the DTW algorithm yielding a distance number between 0 and +∞ (3a and 3b). The minimum of both distances is selected (4) and compared to a set threshold (5) calibrated per speaker, in this example 0.675. If the distance is less than the threshold then the decision is that the aphasic speaker has uttered the target word correctly, otherwise it is classified as incorrect.*

3. Experiment and data

3.1. Participants

Eight native English speakers, 6 male, with chronic anomia post aphasic stroke were recruited. Demographics are shown in Table 1 below. Inclusion criteria were chronic aphasia in the absence of speech apraxia (severe motor speech impairment) as evidenced by: (i) impaired naming ability on the object naming subtest of the Comprehensive Aphasia Test [16]; scores below < 38 are classified as impaired ; (ii) good single word repetition from the same test; normative cut-off>12. All patients gave written consent, and data were processed in accordance with current GDPR guidelines. Ethical approval was granted by NRES Committee East of England– Cambridge, 18/EE/228.

Table 1. *Demographic and clinical data of the patients*

Patient ID	Sex	Age	Months post-stroke	CAT Object naming	CAT Repetition
P1	M	65	108	32	19
P2	M	58	90	19	22
P3	M	70	91	10	28
P4	F	62	21	28	24
P5	M	64	14	6	25
P6	M	59	98	30	31
P7	M	57	109	27	24
P8	F	82	38	29	23
Mean (SD)		65 (8)	71 (40)	23 (10)	25 (4)

3.2. Stimuli

Picture naming stimuli consisted of 220 coloured drawings. They were selected from the top 2000 most frequent words using the *Zipf* index of the SUBTLEX-UK corpus [17] keeping the same distribution of parts of speech for nouns, verbs and adjectives.

3.3. Dataset Collection

We used a tailor-made gamified picture naming treatment app developed in Unity on an Android tablet Samsung SM-T820 to deliver the picture stimuli and record the patients' speech responses. Patients wore a Sennheiser headset SC 665 USB to obtain the speech recordings at 16 kHz which were then stored in a compliant WAVE-formatted file using a 16 bit PCM encoding.

Patients were instructed to name each item presented on screen as quickly and accurately as possible using a single word response. They were given up to 6 seconds to complete each picture naming attempt. The item presentation order was randomised across patients. A SLT was present throughout the assessment and scored the naming responses online in a separate file without giving the patient any performance feedback. A total of 1760 speech recordings (220 words x 8 patients) were acquired.

3.4. Procedure

The SLT classified all naming attempts into one of the following categories: "Correct", "No Response", "Filler", "Phonological Error", "Circumlocution" and "Other". When patients produced multiple speech responses, only the most representative response was selected. For example, when a patient response was scored as 'Filler', and the corresponding recording comprised of multiple 'um', 'ah', 'eh', only one of those attempts was selected to create a single-utterance naming attempt per item. These single-utterance recordings were the data used to evaluate our spoken word verification system and the baseline. Each naming attempt was then re-labelled as 'correct' or 'incorrect', and this last classification was used as the ground truth to evaluate our system's performance and baseline. Figure 3 describes the dataset and each of the patient's naming performance.

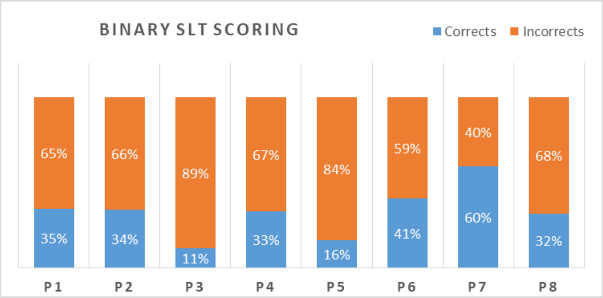

Figure 3. *Each patient's picture naming performance on the 220 item test, as classified by a speech and language therapist (SLT).*

3.4.1. *Inter-SLT-rater Agreement*

A second SLT independently rated all patients' naming attempts to obtain a SLT 'gold-standard' performance metric. Inter-SLT-rater reliability was high overall, with an overall Cohen's kappa of 0.92 ranging between 0.84 and 0.99 across patients. To compare our system to the gold-standard, the performance between SLT raters was calculated across all reported metrics (accuracy, F1-score, Pearson's r).

3.4.2. *ASR Baseline*

We used to a commercially available ASR-engine, Google standard speech-to-text service configured with British English (date used: 24/3/20) to create a baseline with which to compare the performance of our utterance verification system. For each aphasic patient's naming attempt, the same recording to test our system was send to Google's server and a transcription obtained, if the target word was found in the transcript, then the attempt was classified as 'correct', otherwise 'incorrect'.

4. Results

4.1. System Performance

As indicated in section 2, our system utilised a set threshold to make a final decision on marking a patient's naming attempt either 'correct' or 'incorrect'. Two ways of calculating the best threshold were evaluated offline: one that was fixed after optimising it across all patients, and one that was adapted per patient after optimising for each patient separately. Performance results are shown in Table 2. Where significant a pairwise McNemar post-hoc test with Bonferroni correction was calculated. Fixed and adapted versions of our system were significantly better than the baseline with $p<0.05$ and $p<0.005$, respectively.

Table 2. *Overall performance of our system (fixed and adapted versions) and the commercial baseline. A second SLT scoring (SLT2) is also shown.*

System	Accuracy	F1-Score	Pearson's r
baseline	0.882	0.795	0.727
fixed	0.905	0.855	0.784
adapted	0.913	0.871	0.807
SLT2	0.965	0.947	0.921

Performance per patient is illustrated in Figure 4, and the significance of these results is shown in Table 3.

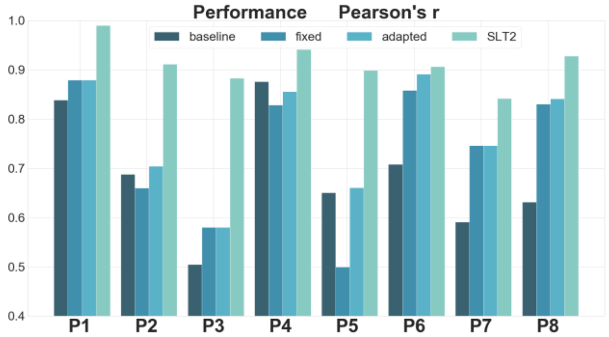

Figure 4. *Comparison of performance between (i) a commercial baseline,(ii) the 'fixed' version of our system, (iii) the 'adapted' version, and (iv) a second independent SLT. The higher the score, the better the performance.*

Table 3. *Post hoc significance testing per patient; pairwise McNemar test with Bonferroni correction. *** p<0.0005, ** p<0.005, * p<0.05 and NS, non-significant.*

Pair	P1	P2	P3	P4	P5	P6	P7	P8
fixed-baseline	NS	NS	NS	NS	NS	**	*	**
adapted-baseline	NS	NS	NS	NS	NS	***	*	***
fixed-adapted	NS	NS	NS	NS	NS	NS	NS	NS
baseline-SLT2	***	**	*	*NS*	*NS*	***	***	***
fixed-SLT2	*	**	**	*	***	*NS*	*NS*	*NS*
adapted-SLT2	*	**	**	*NS*	**	*NS*	*NS*	*NS*

The fixed and adapted versions performed significantly better than the baseline and comparable to the second SLT rater for patients 6, 7 and 8. For the rest of the patients, there are no significant differences in performance. The fixed and adapted versions were not significantly different from each other.

4.2. System Cross-validation

Generalisation of the adapted version of our system to unseen data using offline data was assessed using cross-validation. The assumption, in this case, was that previously collected speech samples from patients could be used to optimise the system's deciding threshold. For each patient, a 10-fold cross-validation procedure was applied, and the average performance across folds is reported, see Table 4. Accuracies for all patients was high, above 84% with a range of 10% and a group average of 89.5%

Table 4. *Results for a 10-fold cross-validation for each patient of the adapted system. For each patient the average across all folds is reported as Mean (±SD).*

Patient	Accuracy	F1-Score	Pearson's r
P1	0.93(±0.068)	0.89(±0.106)	0.85(±0.149)
P2	0.84(±0.082)	0.78(±0.116)	0.67(±0.162)
P3	0.88(±0.055)	0.51(±0.247)	0.46(±0.278)
P4	0.94(±0.055)	0.89(±0.088)	0.85(±0.123)
P5	0.87(±0.060)	0.61(±0.247)	0.56(±0.261)
P6	0.93(±0.071)	0.91(±0.104)	0.85(±0.150)
P7	0.87(±0.081)	0.90(±0.065)	0.72(±0.183)
P8	0.90(±0.038)	0.85(±0.067)	0.79(±0.087)
Mean(SD)	0.895(0.03)	0.790(0.14)	0.718(0.14)
Min	0.836	0.506	0.462
Max	0.936	0.905	0.852
Range	0.1	0.399	0.389

5. Conclusion

We present here a tailor-made system based on a deep learning architecture to automatically assess word naming attempts for people with aphasia. In a sample of eight patients' 1760 naming attempts, our system performed significantly better than the commercial baseline (Google STT service) and, in some instances comparable to the gold-standard SLT scoring. Given the scarcity of aphasic speech corpora, this represents a significant step towards creating a reliable and automatic spoken word assessment system for aphasic speakers and offers clinical practice a deployable preliminary solution for further research and optimisation of similar systems.

Future work will focus on analysing the effects of live feedback on digitally delivered naming interventions. We will adapt our current system to parse large volumes of aphasic speech recordings of word naming attempts offline. Also, given the language-agnostic framework our system is based upon, it will be interesting to see if our system can be used in other languages despite being initially trained in English. This would offer an invaluable tool for aphasic speakers of under-researched languages.

Our system is available open-source to encourage reproducibility and further development in this field; we welcome further insights and collaborations[1].

6. Acknowledgements

DB and the ASR technical development was funded by a Medical Research Council iCASE PhD studentship, award number 1803748. JC is funded a Wellcome Senior Clinical Fellowship.

7. References

[1] Matti. Laine, *Anomia: theoretical and clinical aspects*. Hove: Psychology, 2006.

[2] 'Stroke Association', *Stroke Association*, 2018. https://www.stroke.org.uk/ (accessed Nov. 28, 2018).

[1] https://github.com/DavidBarbera/WNUVforPWA

[3] S. K. Bhogal, R. W. Teasell, N. C. Foley, and M. R. Speechley, ‘Rehabilitation of Aphasia: More Is Better’, *Topics in Stroke Rehabilitation*, vol. 10, no. 2, pp. 66–76, Jul. 2003, doi: 10.1310/RCM8-5TUL-NC5D-BX58.

[4] A. Whitworth, J. Webster, and D. Howard, *A cognitive neuropsychological approach to assessment and intervention in aphasia: a clinician’s guide*, Second edition. London ; New York: Psychology Press, 2014.

[5] A. Abad *et al.*, ‘Automatic word naming recognition for an online aphasia treatment system’, *Computer Speech & Language*, vol. 27, no. 6, pp. 1235–1248, Sep. 2013, doi: 10.1016/j.csl.2012.10.003.

[6] A. Pompili *et al.*, ‘An online system for remote treatment of aphasia’, in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 2011, pp. 1–10, Accessed: Sep. 26, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?id=2140501.

[7] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, ‘AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language’, in *Computational Processing of the Portuguese Language*, vol. 2721, N. J. Mamede, I. Trancoso, J. Baptista, and M. das Graças Volpe Nunes, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 9–17.

[8] K. J. Ballard, N. M. Etter, S. Shen, P. Monroe, and C. T. Tan, ‘Feasibility of Automatic Speech Recognition for Providing Feedback During Tablet-Based Treatment for Apraxia of Speech Plus Aphasia’, *American Journal of Speech - Language Pathology (Online); Rockville*, vol. 28, no. 2S, pp. 818–834, Jul. 2019, doi: http://dx.doi.org.libproxy.ucl.ac.uk/10.m44/2018_AJSLP-MSC18-18-0m9.

[9] *cmusphinx/pocketsphinx*. cmusphinx, 2020.

[10] A. Lee and J. Glass, ‘A comparison-based approach to mispronunciation detection’, in *2012 IEEE Spoken Language Technology Workshop (SLT)*, Miami, FL, USA, Dec. 2012, pp. 382–387, doi: 10.1109/SLT.2012.6424254.

[11] T. Robinson, ‘BEEP dictionary’, *BEEP dictionary*, 1996. http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html (accessed Nov. 09, 2018).

[12] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, ‘Wsjcam0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition’, in *In Proc. ICASSP 95*, 1995, pp. 81–84.

[13] F. Chollet and others, *Keras*. 2015.

[14] Martín Abadi *et al.*, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, ‘Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling’, *arXiv:1412.3555 [cs]*, Dec. 2014, Accessed: Jan. 21, 2019. [Online]. Available: http://arxiv.org/abs/1412.3555.

[16] D. P. Kingma and J. Ba, ‘Adam: A Method for Stochastic Optimisation’, *arXiv:1412.6980 [cs]*, Dec. 2014, Accessed: Sep. 16, 2017. [Online]. Available: http://arxiv.org/abs/1412.6980.

[17] Kate, Swinburn, *Comprehensive aphasia test: CAT / Kate Swinburn, Gillian Porter and David Howard.* Hove: Psychology Press, 2004.

[18] W. J. B. van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, ‘Subtlex-UK: A New and Improved Word Frequency Database for British English’, *Quarterly Journal of Experimental Psychology*, vol. 67, no. 6, pp. 1176–1190, Jun. 2014, doi: 10.1080/17470218.2013.850521.