*Proceedings of the RaPID-4 @LREC 2022*, pages 41–55
Marseille, 25 June 2022


# The Post-Stroke Speech Transcription (PSST) Challenge

**Robert C. Gale,[†] Mikala Fleegle,[‡] Gerasimos Fergadiotis,[‡] Steven Bedrick[†]**
[†]Oregon Health and Science University, Portland, Oregon, USA
[‡]Portland State University, Portland, Oregon, USA
galer@ohsu.edu, soroka@pdx.edu, gf3@pdx.edu, bedricks@ohsu.edu

**Abstract**

We present the outcome of the Post-Stroke Speech Transcription (PSST) challenge. For the challenge, we prepared a new data resource of responses to two confrontation naming tests found in AphasiaBank, extracting audio and adding new phonemic transcripts for each response. The challenge consisted of two tasks. Task A asked challengers to build an automatic speech recognizer (ASR) for phonemic transcription of the PSST samples, evaluated in terms of phoneme error rate (PER) as well as a finer-grained metric derived from phonological feature theory, feature error rate (FER). The best model had a 9.9% FER / 20.0% PER, improving on our baseline by a relative 18% and 24%, respectively. Task B approximated a downstream assessment task, asking challengers to identify whether each recording contained a correctly pronounced target word. Challengers were unable to improve on the baseline algorithm; however, using this algorithm with the improved transcripts from Task A resulted in 92.8% accuracy / 0.921 F1, a relative improvement of 2.8% and 3.3%, respectively.
**Keywords:** anomia, aphasia, speech language pathology assessment, automatic speech recognition

## 1 Introduction

Anomia, or word-finding difficulty, is the primary feature of aphasia (Goodglass and Wingfield, 1997; Raymer and Rothi, 2001), an acquired neurogenic language disorder that affects 2.5-4 million people in the US (Simmons-Mackie, 2018). The primary cause of aphasia is stroke, and 21%–40% of acute stroke patients are diagnosed with anomia by the time they are discharged. Anomia is believed to be indicative of disruption in accessing a semantic description of the target concept, and/or retrieving a fully phonologically specified representation (Dell, 1986; Dell et al., 1997).

Specifically, paraphasias, which are unintended word production errors, typically result from reduced or insufficiently persistent activation of target representations relative to competing non-target representations and/or noise in the system (Dell et al., 1999; Dell et al., 1997). In some cases, people with aphasia produce real word errors. For example, reduced activation of lexical-semantic representations may result in semantic errors (e.g., "dog" for the target "cat") or unrelated errors, sharing no obvious semantic or phonological features with the target word ("chair" for "cat"). Activation of inappropriate phoneme representations may sometimes result in real word errors (e.g., "dog" for the target "log"). However, breakdowns in phonological processing may also lead to non-word productions known as neologisms that may or may not be phonologically related to the target (e.g., "tat" for the target "cat" and "blat" for the target "dog", respectively).

Given the prevalence of anomia in the aphasic population and its tendency to persist even when other symptoms of aphasia remit (Goodglass and Wingfield, 1997), professionals typically assess anomia using confrontation naming tests (Cho-Reyes and Thompson, 2012; Roach et al., 1996; Kaplan et al., 2001), during which a patient is presented with pictures of simple objects and they are asked to name them. The overall accuracy on such tests is an important clinical metric that has been found to be a good indicator of overall aphasia severity (Schuell et al., 1964; Walker and Schwartz, 2012) and is predictive of the ability to convey information during discourse production (Fergadiotis et al., 2019). Furthermore, improvement in naming accuracy has been linked to improvement in overall communicative skills (Carragher et al., 2012; Herbert et al., 2008).

Further, in research settings, professionals develop individualized profiles based on the different types of errors elicited through confrontation naming tests (e.g., phonological, semantic, non-word errors, etc.) and then use these profiles to characterize patients' cognitive-linguistic deficits. Such individualized error profiles have informed theoretical accounts of the cognitive machinery underlying word production (Dell, 1986; Dell et al., 1997; Dell and O'Seaghdha, 1992); lesion-symptom mapping (Schwartz et al., 2009; Schwartz et al., 2012; Walker et al., 2011); personalization of treatments (Best et al., 2013); treatment efficacy studies (Brookshire et al., 2014; Kendall et al., 2003; Kendall et al., 2006; Kendall et al., 2008); the understanding of cross-linguistic treatment generalization (Edmonds and Kiran, 2006); and cortical reorganization investigations after a stroke (Fridriksson et al., 2012)

Error profiles also have the potential to be highly informative in clinical settings for developing individualized intervention plans (Abel et al., 2007). However, currently, developing a patient's profile is prohibitively time- and labor-intensive because it requires phonemic transcriptions for determining response accuracy and the nature of the errors. For naming tests with dozens or hundreds of items, this is rarely feasible in fast-paced clinical settings. As such, there is much interest in the clinical community in automating this process.

To this end, we introduce the Post-Stroke Speech Transcription Challenge (PSST). The Challenge is a shared task consisting of two sub-tasks, one for phonemic transcription (Task A), one for binary classification (Task B). The goals of the PSST Challenge are threefold: first, to produce an accessible dataset relating to these clinical tasks for use by the machine learning community; second, to establish benchmarks for these tasks; and third, to lay the groundwork for a community of practice for machine learning researchers interested in aphasia and other similar disorders.

## 2 Background

### 2.1 Orthographic vs. Phonemic ASR

An automatic speech recognizer (ASR) typically implies an orthographic system (e.g. one that produces words written in the English alphabet). Phonemic ASR, by contrast, uses symbols like ARPABet or the International Phonetic Alphabet (IPA) to indicate how the utterance was pronounced. Unlike their orthographic counterparts, phonemic ASRs might transcribe the same word several different ways, capturing linguistic variability (e.g. dialect, coarticulation) or identifying errors (e.g. mispronunciations, paraphasias).

The previous generation of orthographic ASRs used a layered architecture, with an intermediate layer mapping phoneme sequences to words using a pronunciation dictionary (Mohri et al., 2001). However, the last decade saw a push toward so-called "end-to-end" systems to directly predict orthographic sequences, enabled by deep neural networks, unassisted and unbound by phoneme-to-word constraints (Graves and Jaitly, 2014). This period coincided with the introduction of Librispeech (Panayotov et al., 2015), a corpus extracted from the collection of public domain audiobooks LibriVox. With 1,000 hours of training data available as a free download, LibriSpeech not only became a standard resource in research toolkits, it also came to serve as a primary benchmark for orthographic ASR. In a flurry of activity, the word error rate (WER) for LibriSpeech's `test-clean` dropped from 5.5% (established with its introduction) to under 1.5% just five years later (Zhang et al., 2020). By comparison, the go-to benchmark for phonemic ASR, TIMIT (Garofolo et al., 1993), was less competitive, though phoneme error rates (PER) still halved: an early end-to-end system had a PER of 17.7% (Graves et al., 2013) compared to 8.3% more recently (Baevski et al., 2020).

### 2.2 ASR for Aphasic Speech

End-to-end ASRs rely heavily on statistical methods, learning acoustic and linguistic patterns from large speech corpora. By definition, aphasic speech breaks from typical linguistic patterns, with highly variable error patterns exacerbating the difficulty of the ASR task. Previous work with AphasiaBank reflects these difficulties. Le and Provost (2016) reported 47%–76% PER when grouped by severity of aphasia.[1] Perez et al. (2020) improved on this with a PER of 33%–61%. Le et al. (2018) reported results in terms of WER: 37.4% overall, ranging 34%–63% per severity group.

Small datasets are another hindrance to aphasic ASR, though recent innovations enabled ASRs to be trained on far less data. Similar to recent work in natural language processing (Mikolov et al., 2013; Radford and Narasimhan, 2018; Devlin et al., 2019), the self-supervised methods behind wav2vec 2.0 (Baevski et al., 2020) use large amounts of *unlabeled* speech data, pretraining a model to predict its own abstract feature representations using a contrastive loss function (van den Oord et al., 2018). These pretrained models are intended to be fitted with new output layers and fine-tuned for specific tasks like ASR, and are readily available for download. Using this technique, Baevski et al. (2020) showed how a viable ASR could be trained with as little as 10 minutes labeled data, highlighting its utility for low-resource languages. Applying this to aphasic ASR, Torre et al. (2021) achieved a 22.3–55.5% WER on English AphasiaBank, depending on severity. Remarkably, the authors also trained an ASR using only 1 hour of Spanish AphasiaBank, with a 42.8 WER and a character error rate (CER) of 24.8%.

### 2.3 Automating Aphasia Assessment Tasks

As discussed earlier (§1), the development of a robust ASR system for aphasic speech has the potential to transform clinical practice for the assessment of aphasia. Currently, our group has been developing novel methods to automatically classify clinically relevant types of paraphasias in confrontation picture naming tests (Fergadiotis et al., 2016; Cowan et al., 2021; Casilio et al., 2019; McKinney-Bock and Bedrick, 2019). Our algorithms determine the lexical status of erroneous productions using a word frequency model, use grapheme-to-phoneme analysis to assess phonological similarity between productions and target words, and employ a neural network to measure semantic similarity. Then, information across these three dimensions is combined automatically to classify paraphasias in clinically relevant categories. However, automated analyses of this sort still require that language samples be manually transcribed which represents a major barrier to their translation into practice. Without the ability to automatically produce accurate phonemic transcripts an ASR system, human experts must perform this laborious and error-prone task. For a naming test with dozens or hundreds of items, this is rarely feasible in a clinical setting. As a first step in evaluating the potential of an ASR system for aphasic speech to be used in such a pipeline, Task B in this challenge is focused on

[1] Although they report PER, AphasiaBank's transcripts are (mostly) orthographic. Supplementary materialscontained transcripts tokenized as words (not phonemes) and a pronunciation dictionary, suggesting their ASR targeted a fixed vocabulary instead of free-form phoneme prediction.

a simpler task: assess the ability of the ASR system to generate phonemic transcriptions not for error classification but rather to determine response accuracy.

## 3 Preparing the PSST Corpus

Funding for the dataset preparation and baseline model development activities, and for the shared task itself, originated from the National Institutes of Health's Office of Data Science Strategy, under the "Administrative Supplements to Support Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data" program (NOT-OD-21-094). The goal of this funding mechanism was to support efforts to promote and facilitate the use of existing biomedical datasets by the AI/ML community.

The PSST Corpus is comprised of short speech segments from English AphasiaBank (MacWhinney et al., 2011), specifically responses to the Boston Naming Test Short Form (BNT-SF) (Mack et al., 1992) and Verb Naming Test (VNT) (Thompson, 2012) portions of the protocol. Participants included 107 individuals with aphasia who completed both BNT-SF and VNT as retrieved from English AphasiaBank on September 1, 2021. We defined aphasia as an Aphasia Quotient (AQ) of <93.8 on the Western Aphasia Battery - Revised (Kertesz, 2007) or <11 on the BNT-SF. Participants were right-handed, predominantly English-speaking, with a history of a single, left-hemispheric stroke, adequate hearing and vision, and no significant comorbid neurological or psychiatric illness. Individuals with concomitant motor speech disorders were also included. The extracted segments averaged 3.9 seconds in length, include 3291 utterances from 107 speakers, and total approximately 3.5 hours of audio.

Ground truth phonemic transcriptions for the BNT-SF and VNT were derived from two previous studies and adapted for the purposes of this ASR project. Naming attempts were originally identified and phonemically transcribed by trained research assistants and disagreements resolved by a licensed speech-language pathologist. Transcriptions were entered and time-aligned to audiovisual recordings using ELAN (Max Planck Institute for Psycholingustics, 2022). Using the time alignments, we automatically extracted audio from the full AphasiaBank videos, applying filters for loudness and clarity (see Appendix B.2). A trained research assistant and licensed speech-language pathologist updated transcriptions to reflect the present study's conventions, then the transcripts were normalized and mapped to ARPAbet for ASR purposes (see Appendix B.3). Correctness labels were assigned to all responses by a licensed speech-language pathologist, followed by an audit and resolution process by consensus. We defined correctness as the presence of the target word anywhere within the segmented response. Pronunciation variations of the target word that could be explained by an individual's dialectal pattern and/or typical patterns of coarticulation were scored as correct.

| | Train | Validation | Test |
|---|---|---|---|
| Hours | 2.59 (73%) | 0.36 (10%) | 0.59 (17%) |
| Segments | 2173 (70%) | 325 (10%) | 624 (20%) |
| Speakers | 74 (69%) | 11 (10%) | 22 (21%) |

Table 1: Quantities of data for each split of the PSST dataset in terms of hours of audio, number of segments, and number of speakers.

Data splits targeted train, valid, and test proportions of 70%, 10%, and 20% respectively, with quantities measured as hours of audio. As shown in Table 1, the final proportions were approximately 73%, 10%, and 17%. Each speaker was included in no more than one of the splits. To stratify the splits by overall severity of aphasia, we categorized each participant as mild ($75 < AQ$), moderate ($50 < AQ \leq 75$), severe ($25 < AQ \leq 50$), or very severe ($AQ \leq 25$) per the criteria from the WAB (Kertesz, 2007). To find the optimal split, 1000 candidate configurations were computed, then we chose the configuration with the lowest average KL divergence for duration of audio across the three splits (with a value of about 0.007). Table 1 shows the hours of audio, number of segments, and number of speakers in each split.

## 4 Task A: ASR for Aphasic Speech

Task A asked participants to automatically transcribe the phonemes in each segment of recorded audio. We provided ARPAbet transcripts for the train and validation splits described in §3. We provided code to compute FER and PER for these splits, and we made available our baseline model's source code and pretrained weights. Shortly before the deadline, we released the audio from the test split with the transcripts withheld. Challengers submitted the transcripts their models produced for the test set, and we used the same scripts to compute final metrics. We received entries from two challengers (Yuan et al., 2022; Moël et al., 2022), who submitted transcripts for 7 models apiece.

### 4.1 Evaluation

Task A was evaluated in terms of PER and FER. To calculate PER, we computed the Levenshtein distance (phoneme errors, i.e. the minimum insertions, deletions, and replacements) between target and ASR transcripts. PER is defined as this distance divided by the total length (in phonemes) of the target transcripts.

Like PER, the FER was computed as the errors (in terms of feature distance) divided by the expected length (number of phonemes $\times$ 24 features). Our implementation of feature distance is very similar to one found in `panphon` (Mortensen et al., 2016), specifically the `feature_edit_distance()` algorithm. As discussed in Appendix A, phonological features specified as present/absent ([+feature] / [−feature])

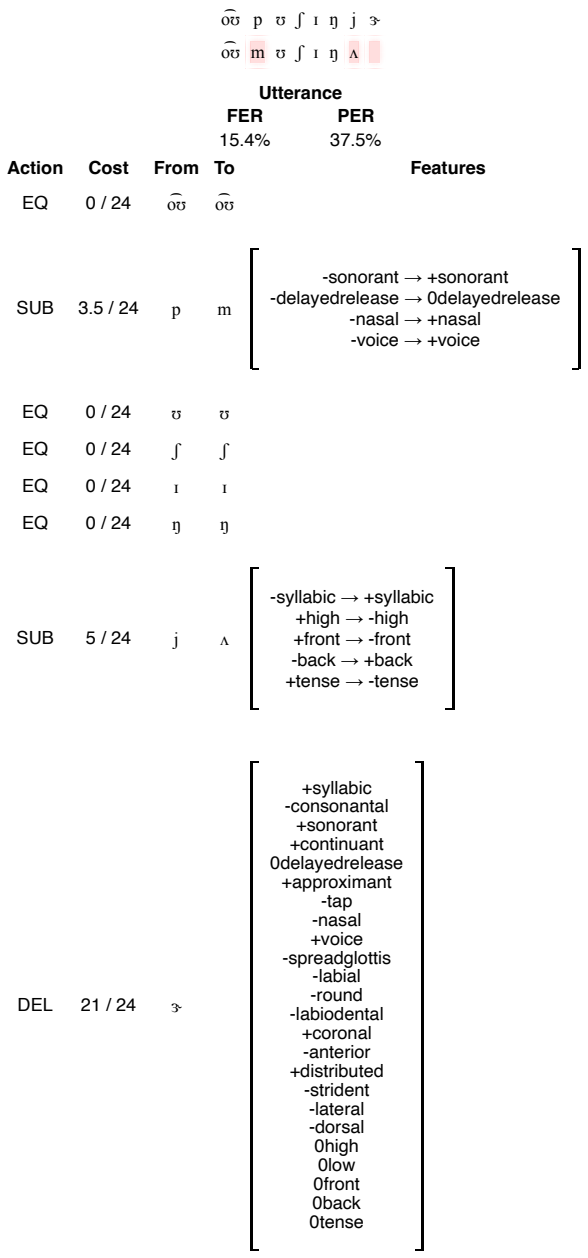


Figure 1: An error analysis generated by the `pssteval-viewer` tool we provided with the challenge materials. The tool shows the PSST transcript (top) aligned to an ASR's output, the FER and PER for the utterance, and then the feature analysis used to compute the FER. This example is utterance id `ACWT01a-VNT20-shove` as transcribed by Moëll/O'Regan et al.'s MO4 model.

or unspecified ([0feature]). If two phonemes differed and the feature was specified in both, that feature error had a cost of 1; if the feature was unspecified in one phoneme, it cost ½. Insertions and deletions were treated as if each feature of missing phoneme was unspecified. The values for each feature align with Hayes (2009), with the exception of diphthongs. While English diphthongs are usually represented by two letters, they behave more like a single sound (Ladefoged and Johnson, 2015); further, each diphthong has only one ARPAbet token. Neither Hayes (2009) nor `panphon` defines features for diphthongs, so we synthesized these definitions, prompting some special rules for feature error calculation. See Appendix A for more details, including the full table of features.

## 4.2 Models

**Baseline Model (*PSST–A*)** For the PSST baseline ASR model (*PSST–A*), we began with a pretrained wav2vec2.0 acoustic model downloaded from fairseq (Ott et al., 2019), specifically the BASE model described in Baevski et al. (2020). This model contains 95m parameters pretrained on 960 hours from the LibriSpeech dataset (Panayotov et al., 2015). We fitted the model with an output layer corresponding to the phoneme inventory of the PSST transcripts, then fine-tuned the model targeting a connectionist temporal classification (CTC) loss. Details on the fine-tuning process can be found in Appendix C.

**Yuan et al. (Y1–Y7)** The approach taken by Yuan et al. focused on data augmentation, exploring outside data sets prepared in a variety of ways. For the challenge, they submitted 7 configurations for our summary, which we will call Y1 through Y7. Each model used a wav2vec2.0 approach comparable to *PSST–A*, but began with the LARGE variant, which uses 315 million parameters, and is trained on 60,000 hours of unlabeled audio from Librivox (from which LibriSpeech is extracted). Y5 was trained only with PSST data, serving as a baseline for their augmentation experiments.

Y2 augmented PSST with 3.9 hours of TIMIT. Adhering to convention, the 61 labels in TIMIT were collapsed to 39 phonemes (Lopes and Perdigao, 2011a; Lee and Hon, 1989), resulting in labels similar to those provided for the PSST challenge, except /ɾ/ was merged with /d/, and /ʒ/ was merged with /ʃ/.

Y4, Y6, and Y7 augmented PSST with LibriSpeech data in various quantities. To prepare LibriSpeech for use with phonemic ASR, Yuan et al. automatically generated pseudo-labels from the orthography using a grapheme-to-phoneme (G2P) model, which had a phoneme inventory nearly aligned with the PSST corpus, although like the TIMIT experiment, the flap symbol /ɾ/ was unused.

Y1 and Y3 augmented PSST with 47 hours of Aphasiabank, taking care to exclude the speakers assigned to the PSST test set. To prepare the (mostly) orthographic corpus for phonemic ASR, Yuan et al. used a technique of iterative self-labeling. First, they produced a set of phonemic labels using a model trained on only PSST data. Then they trained a new model, augmenting PSST with the AphasiaBank samples that exceeded an experimentally determined confidence threshold. Confidence scores were computed in two ways: (a) unweighted, using a standard CTC loss; and (b) weighted,

adjusting confidence with probabilities found during the pseudo-labeling step. This process was repeated until the model no longer improved. Y1 was unweighted with a 0.9 threshold, trimming 47.0 hours of AphasiaBank to the best 33.3. Y3 was weighted, with a 0.7 threshold, yielding 44.0 hours of AphasiaBank.

**Moëll/O'Regan et al. (MO1-MO7)** Moëll/O'Regan et al. (2022) also explored data augmentation strategies for their submissions to the ASR challenge. We refer to their 7 configurations as MO1 through MO7. The authors used two off-the-shelf wav2vec2.0 architectures: BASE, which has 95m parameters, which was pretrained on 960 hours of unlabeled audio; and the LS-960 variant of LARGE, with 315m parameters, which was pretrained on the same 960 hours as BASE. Of those we received, only MO3 and MO7 used BASE, while the rest used LARGE. For MO6, they established an unaugmented baseline with the LARGE architecture.

Much of Moëll/O'Regan et al. focused on expanding PSST and the other datasets with audio perturbation techniques. In MO2 and MO5, they synthesized new PSST data by adjusting the pitch of the audio (while preserving time). In MO4, they synthesized new PSST data by time-stretching the audio (while preserving pitch). For MO6, they augmented PSST by adding Gaussian noise to the signals.

In MO1, MO3, and MO5, Moëll/O'Regan et al. augmented PSST with TIMIT data. They chose to omit utterances that conflicted with the PSST corpus' phoneme inventory, resulting in only 1.1 hours of augmentation data drawn from TIMIT's train and test splits. Noting an acoustic mismatch between the dry, studio-quality recordings of TIMIT and the untreated academic environments of the PSST recordings, the authors experimented with artificial reverb on the TIMIT data: using room impulse response (RIR) convolution, they simulated random rooms by applying filters found in online collections.

### 4.3 Results

Results for the Task A models are shown in Table 2. *PSST–A* showed an FER of 12.1% and a PER of 26.4%. Only two models failed to outperform these metrics: Y6 and Y7, the models using 100 and 960 hours of LibriSpeech. MO1 through MO7 improved on *PSST–A*. Their best-performing model in terms of FER was MO1, which augmented both PSST and TIMIT data using RIR augmentation. MO2 (pitch-shift augmentation) yielded their best PER at 25.1%. The worst-performing models from Moëll/O'Regan et al. were MO6 and MO7 (LARGE, with vs. without noise augmentation) with an FER of 12% each, and a respective PER of 25.9% and 26.1%. Y1 through Y5 were the five best-performing models. The stand-out best was the unweighted pseudo-labeling configuration of the AphasiaBank experiment at 9.9% FER and 20.0% PER. Y2, augmented with TIMIT, was the next best, at 10.3% FER / 21.1% PER. Y3, the weighted AphasiaBank configuration, followed closely behind at 10.4% FER / 21.5% PER. Y5 (no augmentation) had an FER of 11.3% and a PER of 22.3%, and Y4 (3.9 hours of LibriSpeech) improved on this only slightly.

### 4.4 Discussion

Both challengers were primarily focused on augmentation of the PSST dataset, a sensible approach considering the small size of the corpus. Each emphasized a different augmentation strategy: Yuan et al. explored the effects of domain shift and data quantities, while Moëll/O'Regan et al. synthesized additional data with audio perturbation techniques. For Yuan et al., their best model showed a relative improvement of 9% FER / 10% PER against its unaugmented counterpart Y5. The best model from Moëll/O'Regan et al. showed a relative improvement of 5% FER / 3% PER against its unaugmented counterpart M07.

Another difference between the two challengers' submissions was their respective choices of pretrained model. The models from Yuan et al. were all pretrained on 60,000 hours of unlabeled audio, while every model from Moëll/O'Regan et al. pretrained on 960 hours of unlabeled audio. Comparing each challenger's unaugmented LARGE models (Y5 and MO6), the 60,000-hour model improved on the 960-hour model by a relative 9% FER / 14% PER. By comparison, model size had minimal effect: MO6 improved on *PSST–A* by $<$1% FER / 3% PER, having 315 million and 95 million parameters, respectively.

Moëll/O'Regan et al. experimented with several different techniques, laying the foundation for future investigations. One interesting question the authors raise is whether their pitch-shift techniques, which preserve time, could be retaining acoustic markers of phonological features more so than their other techniques. The effects of room acoustics could also be explored in more depth: for example, what if RIR filter selection were more intentional, factoring in room size, shape, and construction material? Finally, Moëll/O'Regan et al. were somewhat conservative with the balance of synthetic data to unmodified PSST. With pitch perturbation, 3-fold augmentation is known to be effective and the recommended practice with last-generation ASR toolkits (Ko et al., 2015). Also, several perturbation techniques could be combined for numerous subtle variations of synthetic data.

The Yuan et al. work also prompts fascinating questions. The paper's narrative centers on the effects of various quantities of in- and out-of-domain data. This effect is clearest between the LibriSpeech-augmented models: Y4 (using only 3.9 hours of LibriSpeech) was fourth-best, whereas Y6 (100 hours) and Y7 (960 hours) were the overall worst. The authors hypothesize this is a consequence of domain mismatch. Indeed, LibriSpeech is audibly different from the PSST corpus

| Model | Arch | Data (hours of audio) Pretrain | PSST | TIMIT | AphasiaBank | Other | ASR FER | PER |
|---|---|---|---|---|---|---|---|---|
| Y1 | LARGE | 60,000 | 2.8 | | $33.3^{U}$ | | **9.9%** | **20.0%** |
| Y2 | LARGE | 60,000 | 2.8 | 3.9 | | | 10.3% | 21.1% |
| Y3 | LARGE | 60,000 | 2.8 | | $44.0^{W}$ | | 10.4% | 21.5% |
| Y4 | LARGE | 60,000 | 2.8 | | | $3.9^{L}$ | 10.6% | 22.2% |
| Y5 | LARGE | 60,000 | 2.8 | | | | 10.9% | 22.3% |
| MO1 | LARGE | 960 | 2.8 | $1.1^{r}$ | | | 11.3% | 25.5% |
| MO2 | LARGE | 960 | $5.6^{p}$ | | | | 11.4% | 25.1% |
| MO3 | BASE | 960 | 2.8 | $1.1^{r}$ | | | 11.7% | 26.3% |
| MO4 | LARGE | 960 | $5.6^{t}$ | | | | 11.7% | 25.4% |
| MO5 | LARGE | 960 | $5.6^{p}$ | $1.1^{r}$ | | | 11.9% | 26.0% |
| MO6 | LARGE | 960 | 2.8 | | | | 12.0% | 25.9% |
| MO7 | BASE | 960 | $5.6^{n}$ | | | | 12.0% | 26.1% |
| *PSST–A* | BASE | 960 | 2.8 | | | | 12.1% | 26.4% |
| Y6 | LARGE | 60,000 | 2.8 | | | $100^{L}$ | 12.5% | 26.0% |
| Y7 | LARGE | 60,000 | 2.8 | | | $960^{L}$ | 16.7% | 38.0% |

$^{L}$ Librispeech, pseudo-labeled with G2P
$^{U}$ iteratively pseudo-labeled (unweighted)
$^{W}$ iteratively pseudo-labeled (weighted)
$^{p}$ with pitch-shifted variants
$^{t}$ with time-shifted variants
$^{n}$ with Gaussian noise augmentation
$^{r}$ RIR reverb applied

Table 2: ASR results for Test set. Results are show in terms of feature error rate (FER), phoneme error rate (PER). Values in gray did not improve on *PSST–A*.

in many ways: speaker demographics, recording conditions, and factors concerning the clinical context of PSST. In contrast to these "bottom-up" characteristics, the authors also describe a "top-down" effect, pointing out how a model like wav2vec2.0 tends to develop an implicit language model (LM) As more out-of-domain data is added, this implicit LM is biased toward out-of-domain transcripts. They support this hypothesis with a principal component analysis, illustrating how the model's contextualized representations visibly shift as more out-of-domain data is added to the training data, more so than the in-domain data from AphasiaBank.

These findings are compelling, though we'd like to emphasize how a segment of speech can be transcribed phonemically in many different ways and still be correct, depending on its context. By ASR standards, TIMIT was transcribed using narrow conventions—extremely narrow in the case of stop consonants (e.g. /b/), which are subdivided as closures (e.g. `BCL`) and releases (e.g. `B`) occurring in isolation or as a sequence (e.g. `BCL B`). In ASR systems, these closures are conventionally relabeled as as silence. (Lopes and Perdigao, 2011a) As a result, the word "maybe" is alternately realized with the stop (when `M EY BCL B IY` becomes /mə͡ɪ bi/) or without (when `M EY BCL IY` becomes /mə͡ɪ i/). In PSST conventions, however, both of these pronunciations are /mə͡ɪbi/. Considering how open-ended transcription can be, we note how Yuan et al. used different techniques to generate pseudo-labels: G2P for Librispeech versus iterative pseudo-labeling for AphasiaBank. The G2P model was trained on a word-to-pronunciation dictionary, and the transcripts are a function of orthography, uninfluenced by the recordings. On the other hand, the AphasiaBank labels were generated by a model trained on the PSST labels themselves, and the transcripts are a function of the audio recordings. Unlike their AphasiaBank model, their G2P model has never been exposed to contextually important phenomena like the mispronunciations, neologisms, inter- and intra-word variation, etc. found in the PSST transcripts. So while the LibriSpeech data is out-of-domain, perhaps its pseudo-labels are better characterized out-of-*range*, with the important distinction that the latter could have a remedy. We could learn more if the LibriSpeech experiments were repeated using the iterative pseudo-labeling methods.

## 5 Task B: Correctness

In Task B, we asked participants to perform a simple example of a downstream task, namely, determining whether a recording contained a target word pronounced correctly. Since the BNT-SF and VNT are confrontation naming tests, they are intended to elicit specific nouns and verbs (respectively) in response. For the challenge, we used the same audio samples as Task A, with true/false labels provided by our annotators (see §3). We also provided a set of acceptable phoneme sequences for each stimulus, including all variations in conjugation, dialect, etc. that we found during data preparation. This allowed us to focus on the question of how to identify and preserve sufficient acoustic-phonetic information from a speech signal to improve on a downstream classification task. Like Task A, we provided scripts for the classification metrics for

the train and valid splits, and we provided the source code for the baseline model.

We received a submission from one challenger (Tran, 2022) for Task B. Our baseline model relies on ASR transcripts from Task A, so we also experimented with the Task A transcripts submitted by challengers.

### 5.1 Evaluation

Task B was evaluated in terms of F1-score, precision, recall, and accuracy. To compute each metric, we tallied the true positives ($TP$), true negatives ($TN$), false positives ($FP$), and false negatives ($FN$). Precision was computed as $\frac{TP}{TP+FP}$, recall as $\frac{TP}{TP+FN}$, and accuracy as $\frac{TP+TN}{TP+TN+FP+FN}$. F1 is the harmonic mean of precision and recall, or $\frac{2TP}{2TP+FP+FN}$.

### 5.2 Models

**Baseline Model (*PSST–B*)** The baseline model for Task B relied on a simple string matching algorithm. We began with the transcripts produced by *PSST–A*, removing any silence and noise labels. If a transcript contained any of the pre-determined "correct" phoneme sequences, uninterrupted and in its entirety, the sample was marked true; otherwise, it was marked false.

**Challenge Submission: Tran (2022)** The approach in Tran (2022) explored acoustic feature engineering as a supplement to the methods used in the baseline model. Motivated by previous work identifying acoustic markers of mild cognitive impairment (MCI) (Roark et al., 2011), Tran conducted a broad search for relevant acoustic features using speech analysis toolkits. These features were aggregated using statistical functions such as mean, minimum, and maximum. Similarly, aggregates of ASR confidence measurements were also explored in the feature set. Features were selected using a T-test, focusing on those deemed statistically significant: mean/standard deviation of loudness, mean/standard deviation of spectral flux, and mean/max of the ASR confidence measures. The features were concatenated with the *PSST–B* predictions into fixed-length vectors, then used to train support vector machine (SVM) and logistic regression classifiers. Hyperparameters were optimized with grid search, and both linear and non-linear SVM kernels were explored.

**The Effects of ASR on Task B** Although neither Yuan et al. (2022) nor Moël et al. (2022) applied their work to Task B, we used their transcripts to observe how each ASR model affected the Correctness task. For this experiment, we followed the same methods as *PSST–B*, swapping out the transcripts for those produced by challengers' models. We also computed metrics using the gold standard transcripts to identify this model's ceiling with hypothetically "perfect" ASR.

### 5.3 Results

The baseline model had an accuracy of 0.903, precision of 0.929, recall of 0.858, and F1 of 0.892. The techniques used in Tran (2022) yielded the same labels as the baseline model, so all metrics were the same.

In the experiment using transcripts from Task A, we found mixed results, which we report in Table 3. The ceiling for "perfect" ASR showed a 0.984 F1, 0.968 precision, 1.000 recall, and 0.985 accuracy. The baseline transcripts had a 0.892 F1, 0.929 precision, 0.858 recall, and 0.903 accuracy. Y2 achieved the top F1 of 0.921, and the top accuracy of 0.985. All of Y1–Y5 improved on the four metrics with one exception: Y5 was below the baseline precision, while obtaining the best recall of 0.914, and the second-best F1 (0.920) and accuracy (0.926). Although Y1 achieved the stand-out best FER and PER in Task A, its transcripts were less effective for identifying correctness, having an F1 of 0.917 and an accuracy of 0.925. Similarly, the gains MO1 through MO7 showed in Task A did not translate to the classification task. Of these, MO7 had the best F1 at 0.888 and accuracy at 0.900. MO4 improved on the baseline in terms of recall (0.865), but at the expense of precision (0.910). MO6, MO3, and MO5 improved on precision (0.934, 0.931, and 0.930, respectively) at the expense of recall (0.842, 0.842, and 0.832, respectively). Y6 and Y7 also improved a bit on precision (0.934 and 0.942, respectively) while taking a heavy hit to recall (0.696 and 0.432, respectively).

### 5.4 Discussion

As Tran points out, the baseline established by *PSST–B* was quite strong. Surprisingly, while 26% of the phonemes produced by *PSST–A* were incorrect, less than 10% of those transcripts were labeled incorrect by *PSST–A*. When planning the challenge, we chose to avoid more difficult (and more clinically informative) tasks, like those that require subtler judgements phonological similarity judgements. In retrospect, we may have designed Task B to be *too* easy, leaving little room to improve on the baseline.

Tran's experiments showed negative results, producing labels identical to *PSST–B*. This suggests that the acoustic features didn't provide more information than the *PSST–B* algorithm could glean from the transcripts. The author also notes that without including the *PSST–B* predictions as a feature, the performance of the acoustic models was only slightly better than chance. Further, Tran discusses the challenge of retaining valuable information while aggregating time sequences to fixed-length vectors. To this, we note some problem formulation differences between Task B and work like Roark et al. (2011) and Fraser et al. (2014). First, their acoustic markers were found in narrative speech tasks, consisting of several successive sentences, containing more prosodic information than a confrontation naming test. Second, the Task B correctness labels describe an event (a paraphasia) as opposed to a condition like aphasia or MCI; thus, the clinical dementia rating used in the cited work is more analogous to the AQ index included with the PSST data.

| Transcripts | F1 | Precision | Recall | Accuracy | FER | PER |
|---|---|---|---|---|---|---|
| *PSST-Gold* | 0.984 | 0.968 | 1.000 | 0.985 | 0% | 0% |
| Y2 | **0.921** | 0.941 | 0.901 | **0.928** | 10.3% | 21.1% |
| Y5 | 0.920 | 0.926 | **0.914** | 0.926 | 10.9% | 22.3% |
| Y1 | 0.917 | 0.941 | 0.894 | 0.925 | **9.9%** | **20.0%** |
| Y3 | 0.903 | **0.949** | 0.861 | 0.914 | 10.4% | 21.5% |
| Y4 | 0.899 | 0.930 | 0.871 | 0.910 | 10.6% | 22.2% |
| *PSST-Baseline* | 0.892 | 0.929 | 0.858 | 0.903 | 12.1% | 26.4% |
| MO7 | 0.888 | 0.928 | 0.851 | 0.900 | 12.0% | 26.1% |
| MO4 | 0.887 | 0.910 | 0.865 | 0.897 | 11.7% | 25.4% |
| MO6 | 0.885 | 0.934 | 0.842 | 0.899 | 12.0% | 25.9% |
| MO1 | 0.884 | 0.912 | 0.858 | 0.896 | 11.3% | 25.5% |
| MO3 | 0.884 | 0.931 | 0.842 | 0.897 | 11.7% | 26.3% |
| MO2 | 0.883 | 0.921 | 0.848 | 0.896 | 11.4% | 25.1% |
| MO5 | 0.878 | 0.930 | 0.832 | 0.893 | 11.9% | 26.0% |
| Y6 | 0.798 | 0.934 | 0.696 | 0.836 | 12.5% | 26.0% |
| Y7 | 0.593 | 0.942 | 0.432 | 0.724 | 16.6% | 38.0% |

Table 3: Correctness results using the *PSST–B* model, using Test transcripts generated by Task A models Y1-Y7 and MO1-MO7. F1, precision, recall, and accuracy scores are shown, alongside the FER and PER shown in Task A. The first row, *PSST-Gold*, used the gold standard transcripts. Values in gray did not improve on *PSST–A*.

In the experiment using transcripts from Task A, we see how improvements to FER and PER do not necessarily ripple out to the downstream task. FER and PER consider the full transcript, so improvements outside the response boundaries have no effect on correctness. Even if improvements occurred within the response boundary, so long as any errors remain, the *PSST–B* algorithm will mark it as false. A correctness algorithm that considered likelihoods for each token in the sequence might better show a relationship to incremental ASR improvements.

The perfect recall and imperfect precision of *PSST-Gold* indicate that with ideal transcripts, 10 false positives account for all the errors. In these samples, the correct sequence of phonemes were present, but the response was incorrect for other reasons. For example, the string /mə͡ilbɑks/ ("mailbox") contains /mə͡il/, it is incorrect because it is a different word, and a noun rather than a verb. Similarly, /klæfɪŋ/ contains the /læfɪŋ/ ("laughing"), but the production is a non-word. This can be seen as a limitation of the algorithm with no sensitivity to word and syllable boundaries. Unlike *PSST-Gold*, the remaining transcripts had worse recall than precision, suggesting they tended to miss correctly pronounced words (false negatives) more so than they smoothed out mispronunciations (false positives).

We reviewed *PSST–B* errors that were common across the transcripts. In one instance, we provided the transcript /pʊʃɪŋ/ ("pushing") and labeled the response as correct, whereas none of the ASR transcripts agreed. In fact, 7 of 12 transcripts had /m/ as the initial consonant, and upon listening to the sample post-hoc, we tend to agree with the ASR. Another particular challenge pertains to matters of motor planning and articulation. One example included a prolongation of the initial /m/ in the word "mixing", i.e. /mː- mɪksɪŋ/. During the prolongation, the participant was also lowering the jaw with lips closed, introducing more oral resonance than typical for /m/, and demonstrating involuntary pitch fluctuations. This seemed to confuse all the ASR systems, though predictably: 6 were transcribed as /**p**ɪksɪŋ/ and 5 as /**bp**ɪksɪŋ/ In their raw form, our transcripts annotated phenomena like phonological fragments and prolongations, but these annotations were removed during pre-processing. Furthermore, none of our annotations addressed deviations in pitch or resonance.

## 6 Conclusion

As we hoped, the PSST participants improved on our baseline approach. The ASR metrics FER and PER were improved by a relative 18% and 24%, respectively. Those improvements alone improved the F1 of the Correctness task by a relative 3.3%. These ideas warrant further experimentation, and we expect progress will continue as a result of expanding the PSST data and refining this work.

To this end, as a next step we will investigate which linguistic and clinical characteristics pose the greatest challenge across the ASR systems. Further, we will assess how FER/PER relate to the performance of downstream tasks; and, explore how different approaches to FER computation could improve its utility. At the same, we intent to continue expanding the PSST dataset using AphasiaBank data while also refining our evaluation methods. Given the opportunity to hold another PSST challenge, we see ample opportunity to raise the bar with the downstream tasks: introducing tasks like phonological and morphological similarity assessment, or leaning in to the complexities associated with accents and dialect.

## 7 Acknowledgements

This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award numbers R01DC015999 and R01DC015999-04S1.

We thank past and present Portland State University students Mia Cywinski, Emily Tudorache, and Khanh Nguyen for their contribution to the preparation and annotation of the transcribed dataset, as well as current and past members of the Oregon Health & Science University research team Alexandra Salem, Linying Li, Dr. Brooke Cowan, and Dr. Katy McKinney-Bock. We thank the challenge participants for taking on these important tasks. We thank Brian MacWhinney for above-and-beyond assistance with hosting our challenge data. Finally, we are grateful for AphasiaBank participants, whose voices make this research possible.

Abel, S., Willmes, K., and Huber, W. (2007). Model-oriented naming therapy: Testing predictions of a connectionist model. *Aphasiology*, 21(5):411–447, April.

Baevski, A., Zhou, H., rahman Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Best, W., Greenwood, A., Grassly, J., Herbert, R., Hickin, J., and Howard, D. (2013). Aphasia rehabilitation: Does generalisation from anomia therapy occur and is it predictable? A case series study. *Cortex*, 49(9):2345–2357, October.

Brookshire, C. E., Conway, T., Pompon, R. H., Oelke, M., and Kendall, D. L. (2014). Effects of intensive phonomotor treatment on reading in eight individuals with aphasia and phonological alexia. *American Journal of Speech-Language Pathology*, 23(2):S300–S311, May.

Carragher, M., Conroy, P., Sage, K., and Wilkinson, R. (2012). Can impairment-focused therapy change the everyday conversations of people with aphasia? A review of the literature and future directions. *Aphasiology*, 26(7):895–916, April.

Casilio, M., Fergadiotis, G., Bedrick, S., and McKinney-Bock, K. (2019). Can machines classify paraphasias? Evidence from Dell's model.

Cho-Reyes, S. and Thompson, C. K. (2012). Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology*, 26(10):1250–1277, October.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York, NY.

Cowan, B., McKinney-Bock, K., Casilio, M., Fergadiotis, G., and Bedrick, S. (2021). An evaluation framework for machine learning models of paraphasia classification.

Dell, G. S. and O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42(1–3):287–314.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4):801–838, October.

Dell, G. S., Chang, F., and Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, 23(4):517–542, October.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3):283–321.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Edmonds, L. A. and Kiran, S. (2006). Effect of semantic naming treatment on crosslinguistic generalization in bilingual aphasia. *Journal of Speech Language and Hearing Research*, 49(4):729, August.

Fergadiotis, G., Gorman, K., and Bedrick, S. (2016). Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology*, 25(4S):S776–S787, December.

Fergadiotis, G., Kapantzoglou, M., Kintz, S., and Wright, H. H. (2019). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology*, 33(5):544–560.

Fraser, K. C., Meltzer, J. A., Graham, N., Leonard, C., and Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

Fridriksson, J., Richardson, J. D., Fillmore, P., and Cai, B. (2012). Left hemisphere plasticity and aphasia recovery. *NeuroImage*, 60(2):854–863, April.

Goodglass, H. and Wingfield, A. (1997). *Anomia: Neuroanatomical and cognitive correlates*. Academic Press, San Diego, CA.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In Eric P. Xing et al., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, pages 1764–1772, Bejing, China, 22–24 Jun. PMLR.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Halpern, B. M., Feng, S., Son, R. v., Brekel, M. v. d., and Scharenborg, O. (2022). Low-resource auto-

matic speech recognition and error analyses of oral cancer speech. *Speech Communication*, 141:14–27.

Hayes, B. (2009). *Introductory Phonology*. Wiley-Blackwell, Malde, MA.

Herbert, R., Hickin, J., Howard, D., Osborne, F., and Best, W. (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology*, 22(2):184–203, January.

Kaplan, E., Goodlgass, H., and Weintraub, S. (2001). *Boston Naming Test*. Lippincott Williams & Wilkins, Philadelphia, PA, 2nd edition.

Kendall, D., Conway, T., Rosenbek, J., and Gonzalez-Rothi, L. (2003). Case study: Phonological rehabilitation of acquired phonologic alexia. *Aphasiology*, 17(11):1073–1095, January.

Kendall, D. L., Rodriguez, A. D., Rosenbek, J. C., Conway, T., and Gonzalez Rothi, L. J. (2006). Influence of intensive phonomotor rehabilitation on apraxia of speech. *Journal of Rehabilitation Research and Development*, 43(3):409–418, June.

Kendall, D. L., Rosenbek, J. C., Heilman, K. M., Conway, T., Klenberg, K., Gonzalez Rothi, L. J., and Nadeau, S. E. (2008). Phoneme-based rehabilitation of anomia in aphasia. *Brain and Language*, 105(1):1–17, April.

Kertesz, A. (2007). *Western Aphasia Battery – R*. Grune & Stratton, New York.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Interspeech 2015*, pages 3586–3589. ISCA, September.

Ladefoged, P. and Johnson, K. (2015). *A Course in Phonetics*. Cengage Learning, Stamford, CT, seventh edition.

Le, D. and Provost, E. M. (2016). Improving automatic recognition of aphasic speech with aphasiabank. In *INTERSPEECH*.

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

Lee, K.-F. and Hon, H.-W. (1989). Speaker-independent phone recognition using hidden markov models. *IEEE Trans. Acoust. Speech Signal Process.*, 37:1641–1648.

Lopes, C. and Perdigao, F. (2011a). Phoneme recognition on the TIMIT database. In *Speech Technologies*. InTech.

Lopes, C. and Perdigao, F. (2011b). Phoneme recognition on the timit database. In Ivo Ipsic, editor, *Speech Technologies*, chapter 14. IntechOpen, Rijeka.

Mack, W. J., Freed, D. M., Williams, B. W., and Henderson, V. W. (1992). Boston Naming Test: Shortened Versions for Use in Alzheimer's Disease. *Journal of Gerontology*, 47(3):P154–P158, May.

Max Planck Institute for Psycholingustics. (2022). *ELAN (Version 6.3) [Computer software]*. The Language Archive, Nijmegen, The Netherlands. URL: https://archive.mpi.nl/tla/elan.

McKinney-Bock, K. and Bedrick, S. (2019). Classification of semantic paraphasias: optimization of a word embedding model. In *3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 52–62, Minneapolis, USA. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*.

Mohri, M., Pereira, F., and Riley, M. (2001). Weighted finite-state transducers in speech recognition. *Departmental Papers (CIS)*, page 11.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. S. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL. URL: https://github.com/dmort27/panphon.

Moël, B., O'Regan, J., Mehta, S., Kirkland, A., Lameris, H., Gustafsson, J., and Beskow, J. (2022). Speech data augmentation for improving phoneme transcriptions of aphasic speech using wav2vec 2.0 for the PSST Challenge. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, June 20-25, 2022. European Language Resource Association (ELRA).

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. URL: https://github.com/facebookresearch/fairseq.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Perez, M., Aldeneh, Z., and Provost, E. M. (2020). Aphasic Speech Recognition Using a Mixture of Speech Intelligibility Experts. In *Proc. Interspeech 2020*, pages 4986–4990.

Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Raymer, A. M. and Rothi, L. J. G. (2001). Impairments of word comprehension and production. In Roberta Chapey, editor, *Language intervention strategies in aphasia and related neurogenic communication disorders*, pages 606–625. Lippincott Williams & Wilkins, 4 edition.

Roach, A., Schwartz, M., Martin, N., Grewal, R., and Brecher, A. (1996). The Philadelphia Naming Test:

scoring and rationale. *Clin. Aphasiol.*, 24:121–133, January.

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

Schuell, H., Jenkins, J. J., and Jimenez-Pabon, E. (1964). *Aphasia in adults*. Harper & Row, New York, NY.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Faseyitan, O., Brecher, A., Dell, G. S., and Coslett, H. B. (2009). Anterior temporal involvement in semantic word retrieval: Voxel-based lesion-symptom mapping evidence from aphasia. *Brain*, November.

Schwartz, M. F., Faseyitan, O., Kim, J., and Coslett, H. B. (2012). The dorsal stream contribution to phonological retrieval in object naming. *Brain*, 135(12):3799–3814, December.

Simmons-Mackie, N. (2018). *Aphasia in North America*. Aphasia Access, Moorestown, NJ.

Tao, T., Yoon, S.-Y., Fister, A., Sproat, R., and Zhai, C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 250–257, Sydney, Australia, July. Association for Computational Linguistics.

Thompson, C. K. (2012). Northwestern Assessment of Verbs and Sentences (NAVS). https://www.scholars.northwestern.edu/en/publications/northwestern-assessment-of-verbs-and-sentences-navs.

Torre, I. G., Romero, M., and Álvarez, A. (2021). Improving Aphasic Speech Recognition by Using Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish. *Applied Sciences*, 11(19):8872, September.

Tran, T. (2022). Post-Stroke Speech Transcription Challenge (Task B): Correctness detection in anomia diagnosis with imperfect transcripts. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, June 20-25, 2022. European Language Resources Association (ELRA).

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Walker, G. M. and Schwartz, M. F. (2012). Short-form Philadelphia naming test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology*, pages S140–S153, May.

Walker, G. M., Schwartz, M. F., Kimberg, D. Y., Faseyitan, O., Brecher, A., Dell, G. S., and Coslett, H. B. (2011). Support for anterior temporal involvement in semantic error production in aphasia: New evidence from VLSM. *Brain and Language*, 117(3):110–122.

Yuan, J., Cai, X., and Church, K. (2022). Data augmentation for the Post-Stroke Speech Transcription (PSST) challenge: Sometimes less is more. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, June 20-25, 2022. European Language Resources Association (ELRA).

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *ArXiv*, abs/2010.10504.

## 8 Language Resource References

Garofolo, John and Lamel, Lori and Fisher, William and Fiscus, Jonathan and Pallett, David and Dahlgren, Nancy and Zue, Victor. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Abacus Data Network.

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307. Supported by NIH-NIDCD R01-DC008524 (2007-2022).

Vassil Panayotov and Guoguo Chen and Daniel Povey and Sanjeev Khudanpur. (2015). *Librispeech: An ASR corpus based on public domain audio books*. OpenSLR (https://www.openslr.org).

# Appendices

## A More on Phonological Features and Feature Error Rate (FER)

Phoneme error rate (PER) is the go-to evaluation metric for phonemic ASR, derived from the edit distance between the predicted and target phonemes. For PSST, we explore a feature error rate (FER) as finer-grained alternative to PER. Instead of the phonemic edit distance, the error in FER is a phonological feature distance (Mortensen et al., 2016; Tao et al., 2006). In short, each phoneme is represented by a quasi-binary vector indicating the presence and absence of each feature described by the system, and these vectors can be used to compute a measure akin to Euclidean distance. Feature distance is then normalized by the sequence length to determine the error rate, much like PER. In other words, FER is a way of giving "partial credit" to an ASR transcript when it produces phonemes which are similar (but not exact) to the target transcript, defining similarity in terms of distinctive phonological features.

Phonological features distill information about how people distinguish the sounds of their language from one another, while also grouping phonemes into natural classes (Chomsky and Halle, 1968). For example, the English words "bead" and "bid" both contain *high*

| ARPAbet | IPA | Example Word | Special diphthong features |
|---|---|---|---|
| EY | /e͡ɪ/ | "bay" | [-+high, +-tense] |
| OW | /o͡ʊ/ | "beau" | [-+high, +-tense] |
| OY | /ɔ͡ɪ/ | "boy" | [-+high, -+front, +-back, +-round] |
| AW | /a͡ʊ/ | "bough" | [-+high, +-low, -+back, -+round] |
| AY | /a͡ɪ/ | "buy" | [-+high, +-low, -+front] |

Table 4: Diphthongs and their unique features used during computation of feature error rate (FER)

| Cost | Feature Changes | | |
|---|---|---|---|
| 1 | [–feature] | ↔ | [+feature] |
| 0.75 | [–feature] | ↔ | [+–feature] |
| | [–+feature] | ↔ | [+feature] |
| 0.5 | [–feature] | ↔ | [0feature] |
| | [–+feature] | ↔ | [+–feature] |
| | [0feature] | ↔ | [+feature] |
| 0.25 | [–feature] | ↔ | [–+feature] |
| | [–+feature] | ↔ | [0feature] |
| | [0feature] | ↔ | [+–feature] |
| | [+–feature] | ↔ | [+feature] |
| 0 | [–feature] | ↔ | [–feature] |
| | [–+feature] | ↔ | [–+feature] |
| | [0feature] | ↔ | [0feature] |
| | [+–feature] | ↔ | [+–feature] |
| | [+feature] | ↔ | [+feature] |

Table 5: Costs associated with each feature difference during computation of feature error rate (FER)

*front vowels*. In the feature system proposed by (Hayes, 2009), high front vowels are a natural class primarily described as [+syllabic, +high, +front]. In fact, the two phonemes share all the same features, save for one distinction: the /i/ in b**ea**d is [+tense], while the /ɪ/ in "b**i**d" is said to be lax, or [−tense], distinguished by only that feature. Some phonemes do not specify a certain feature, for example, the tense/lax distinction only applies to vowels, so /b/ and /d/ are both [0tense].

Distinctive features are thus used in phonological analysis to classify phonemes and describe their linguistic behavior (e.g. allophonic variations or historical sound changes), and they are empirically validated for that purpose. Recently, however, phonological features have found novel applications in computational linguistics, enhancing statistical models with information about phonemes' features and feature distances (Mortensen et al., 2016).

For the PSST challenge, we use FER as an evaluation metric for ASR. Previous research has used a variation of the concept as a metric for automatic phoneme recognition (Halpern et al., 2022), but the practice is not well established. Our motivation here is to gain insight into what makes an ASR a better fit for our tasks. During transcription, certain feature-adjacent phonemes can be quite difficult to distinguish (by an ASR or a human). Yet in some contexts, feature-adjacent phonemes like /t/ and /d/ are functionally interchangeable (e.g. a sound change attributable to dialect), whereas more distant phoneme errors would invalidate an analysis.

Compared to PER, FER is much more difficult to compute and understand, and all the more difficult for those with no background in phonology. For this reason, we put together the `pssteval-viewer` tool to illustrate how FER was computed for each utterance, which we shared with PSST challengers in our evaluation toolkit. An example feature analysis generated by the software is shown in Figure 1.

To build our table of feature values, we began with the system specified by Hayes (2009). We excluded two features which do not contrast in our ARPAbet transcript (nor English, generally): [constrictedglottis] and [trill]. Diphthongs presented a conundrum: with no single entry for diphthongs in the feature table, the two components would be treated as two phonemes. In other words, if a diphthong replaced a monophthong (or vice versa), the distance would always include an insertion or a deletion, and the feature error would be greater than a full phoneme. To rectify this, we treated diphthongs as individual phonemes (as they are in ARPAbet), adding new entries in the feature table for /e͡ɪ/, /o͡ʊ/, /ɔ͡ɪ/, /a͡ʊ/, and /a͡ɪ/ (the vowels in "bay", "bow/beau", "boy", "bow/bough", and "buy", respectively), and new feature values to capture their movement. These all emphasize the first of their two component vowels (Ladefoged and Johnson, 2015), so when a feature has the new value [+⁻feature] (present toward absent) we consider it between [+feature] and [0feature], while [−⁺feature] (absent toward present) is between [0feature and [−feature]. The five diphthongs and their novel features are highlighted in Table 4 All combinations of feature changes and their costs are shown in Table 5. introduced two new symbols to capture how a diphthong's features moved between its components.

# B More Details on Data Preparation

## B.1 More on Data Preparation

Approximately one third of the total number of included responses ($n = 3291$) consisted of BNT-SF first responses ($n = 1074$), defined as single-word first complete attempts according to the scoring guidelines of the

| ARPAbet | IPA | consonantal | delayedrelease | continuant | sonorant | approximant | syllabic | tap | nasal | voice | spreadglottis | labial | round | labiodental | coronal | anterior | distributed | strident | lateral | dorsal | high | low | front | back | tense |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P | p | + | − | − | − | − | − | − | − | − | − | + | − | − | − | 0 | 0 | 0 | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| B | b | + | − | − | − | − | − | − | − | + | − | + | − | − | − | 0 | 0 | 0 | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| T | t | + | − | − | − | − | − | − | − | − | − | − | − | − | + | + | − | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| D | d | + | − | − | − | − | − | − | − | + | − | − | − | − | + | + | − | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| K | k | + | − | − | − | − | − | − | − | − | − | − | − | − | − | 0 | 0 | 0 | − | + | + | − | 0 | 0 | 0 | 0 |
| G | g | + | − | − | − | − | − | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | + | − | 0 | 0 | 0 | 0 |
| CH | t͡ʃ | + | + | − | − | − | − | − | − | − | − | − | − | − | + | − | + | + | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| JH | d͡ʒ | + | + | − | − | − | − | − | − | + | − | − | − | − | + | − | + | + | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| F | f | + | + | + | − | − | − | − | − | − | − | + | − | + | − | 0 | 0 | 0 | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| V | v | + | + | + | − | − | − | − | − | + | − | + | − | + | − | 0 | 0 | 0 | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| TH | θ | + | + | + | − | − | − | − | − | − | − | − | − | − | + | + | + | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| DH | ð | + | + | + | − | − | − | − | − | + | − | − | − | − | + | + | + | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| S | s | + | + | + | − | − | − | − | − | − | − | − | − | − | + | + | − | + | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| Z | z | + | + | + | − | − | − | − | − | + | − | − | − | − | + | + | − | + | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| SH | ʃ | + | + | + | − | − | − | − | − | − | − | − | − | − | + | − | + | + | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| ZH | ʒ | + | + | + | − | − | − | − | − | + | − | − | − | − | + | − | + | + | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| HH | h | − | + | + | − | − | − | − | − | − | + | − | − | − | − | 0 | 0 | 0 | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| M | m | + | 0 | − | + | − | − | − | + | + | − | + | − | − | − | 0 | 0 | 0 | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| N | n | + | 0 | − | + | − | − | − | + | + | − | − | − | − | + | + | − | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| NG | ŋ | + | 0 | − | + | − | − | − | + | + | − | − | − | − | − | 0 | 0 | 0 | − | + | + | − | 0 | 0 | 0 | 0 |
| L | l | + | 0 | + | + | + | − | − | − | + | − | − | − | − | + | + | − | − | + | − | 0 | 0 | 0 | 0 | 0 | 0 |
| DX | ɾ | + | 0 | + | + | + | − | + | − | + | − | − | − | − | + | + | − | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | j | − | 0 | + | + | + | − | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | + | − | + | − | + | + |
| W | w | − | 0 | + | + | + | − | − | − | + | − | + | + | − | − | 0 | 0 | 0 | − | + | + | − | − | + | + | + |
| R | ɹ | − | 0 | + | + | + | − | − | − | + | − | − | − | − | + | − | + | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| ER | ɝ, ɚ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | + | − | + | − | − | − | 0 | 0 | 0 | 0 | 0 | 0 |
| IY | i | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | + | − | + | − | + | + |
| IH | ɪ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | + | − | + | − | − | − |
| UW | u | − | 0 | + | + | + | + | − | − | + | − | + | + | − | − | 0 | 0 | 0 | − | + | + | − | − | + | + | + |
| UH | ʊ | − | 0 | + | + | + | + | − | − | + | − | + | + | − | − | 0 | 0 | 0 | − | + | + | − | − | + | − | − |
| EH | ɛ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | − | − | + | − | − | − |
| EY | e͡ɪ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | $-^{+}$ | − | + | − | $+^{-}$ | |
| AH | ʌ, ə | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | − | − | − | + | − | |
| AO | ɔ | − | 0 | + | + | + | + | − | − | + | − | + | + | − | − | 0 | 0 | 0 | − | + | − | − | − | + | − | |
| OW | o͡ʊ | − | 0 | + | + | + | + | − | − | + | − | + | + | − | − | 0 | 0 | 0 | − | + | $-^{+}$ | − | − | + | $+^{-}$ | |
| OY | ɔ͡ɪ | − | 0 | + | + | + | + | − | − | + | − | + | $+^{-}$ | − | − | 0 | 0 | 0 | − | + | $-^{+}$ | − | $-^{+}$ | $+^{-}$ | − | |
| AE | æ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | − | + | + | − | 0 | |
| AW | a͡ʊ | − | 0 | + | + | + | + | − | − | + | − | − | $-^{+}$ | − | − | 0 | 0 | 0 | − | + | $-^{+}$ | $+^{-}$ | − | $-^{+}$ | 0 | |
| AY | a͡ɪ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | $-^{+}$ | $+^{-}$ | $-^{+}$ | − | 0 | |
| AA | ɑ | − | 0 | + | + | + | + | − | − | + | − | − | − | − | − | 0 | 0 | 0 | − | + | − | + | − | + | 0 | |

Table 6: The 40 phonemes in this ASR system in ARPAbet and IPA, and their associated phonological features. Features align with Hayes (2009), with the exception of diphthong handling, which are treated as individual phonemes here (using special symbols $-^{+}$and $+^{-}$to describe their movement).

| | Split | Mild | | Moderate | | Severe | | Very Severe | |
|---|---|---|---|---|---|---|---|---|---|
| **Hours** | Train | 0.85 | (32.8%) | 1.39 | (53.4%) | 0.33 | (12.5%) | 0.03 | (1.0%) |
| | Validation | 0.11 | (31.5%) | 0.20 | (56.3%) | 0.04 | (12.1%) | 0.00 | (0.0%) |
| | Test | 0.17 | (28.2%) | 0.32 | (55.0%) | 0.08 | (13.9%) | 0.02 | (2.6%) |
| **Segments** | Train | 1073 | (49.3%) | 893 | (41.0%) | 187 | (8.6%) | 20 | (0.9%) |
| | Validation | 121 | (37.2%) | 170 | (52.3%) | 34 | (10.4%) | 0 | (0.0%) |
| | Test | 262 | (41.9%) | 275 | (44.0%) | 67 | (10.7%) | 20 | (3.2%) |
| **Speakers** | Train | 33 | (44.5%) | 32 | (43.2%) | 8 | (10.8%) | 1 | (1.3%) |
| | Validation | 4 | (36.3%) | 6 | (54.5%) | 1 | (9.0%) | 0 | (0.0%) |
| | Test | 8 | (36.3%) | 10 | (45.4%) | 3 | (13.6%) | 1 | (4.5%) |

Table 7: Detailed breakdown of the data illustrating the attempt to balance aphasia severity across each split. The balance is show for each of hours of audio, number of segments, and number of speakers.

| IPA | Example | ARPAbet |
|---|---|---|
| /p/ | "**p**at" | P |
| /b/ | "**b**at" | B |
| /t/ | "**t**en" | T |
| /d/ | "**d**en" | D |
| /k/ | "**c**oat" | K |
| /g/ | "**g**oat" | G |
| /ɾ/ | "bu**tt**er" (allophone of /t/, /d/) | DX |
| /ʔ/ | "co**tt**on" (allophone of /t/) | *(removed)* |
| /t͡ʃ/ | "**ch**u**rch**" | CH |
| /d͡ʒ/ | "**j**u**dge**" | JH |
| /f/ | "**f**an" | F |
| /v/ | "**v**an" | V |
| /θ/ | "**th**in" (voiceless) | TH |
| /ð/ | "**th**en" (voiced) | DH |
| /s/ | "**s**ee" | S |
| /z/ | "**z**oo" | Z |
| /ʃ/ | "**sh**oe" | SH |
| /ʒ/ | "occa**s**ion" | ZH |
| /h/ | "**h**at" | HH |
| /n/ | "**n**ose" | N |
| /ŋ/ | "si**ng**" | NG |
| /m/ | "**m**an" | M |

| IPA | Example | ARPAbet |
|---|---|---|
| /w/ | "**w**in" | W |
| /j/ | "**y**es" | Y |
| /ɹ/ | "**r**ed" | R |
| /l/ | "**l**ate" | L |
| /ɝ/ | "h**ear**d" (stressed) | ER |
| /ɚ/ | "p**er**haps" (unstressed) | ER |
| /i/ | "sh**e**" | IY |
| /ɪ/ | "f**i**t" | IH |
| /u/ | "b**oo**t" | UW |
| /ʊ/ | "w**oo**d" | UH |
| /e͡ɪ/ | "st**a**te" | EY |
| /ɛ/ | "r**e**d" | EH |
| /o͡ʊ/ | "v**o**te" | OW |
| /ɔ͡ɪ/ | "b**oy**" | OY |
| /ɔ/ | "d**a**wn" | AO |
| /ʌ/ | "b**u**t" (stressed) | AH |
| /ə/ | "**a**lone" (unstressed) | AH |
| /ɑ/ | "n**o**t" | AA |
| /æ/ | "c**a**t" | AE |
| /a͡ɪ/ | "k**i**te" | AY |
| /a͡ʊ/ | "c**ow**" | AW |

Table 8: A list of International Phonemic Alphabet (IPA) notations used by our laboratory and their ARPAbet mappings for ASR, with examples.

PNT. For about a quarter of BNT-SF responses ($n = 155$), first complete attempts were segmented to also contain surrounding connected speech when phonemic boundaries between words were blurred. BNT-SF responses that overlapped with examiner speech as well as responses labeled as non-naming attempts (e.g., descriptions of the target, whispered responses, etc.) were excluded.

Approximately two thirds of the response data consisted of VNT first responses ($n = 2217$), defined as any response from the moment following picture stimulus presentation and a first examiner prompt to the moment preceding a second examiner prompt and/or the administration of the next picture stimulus. Nonverbal cues from the examiner, such as gestures indicating the target verb, were treated as second prompts if they occurred. Examiner speech that overlapped with a response was excluded, and, if possible, exactly one segment of participant speech was retained per test item.

### B.2 Audio Preprocessing

When we extracted audio-only segments from the full TalkBank session videos, we applied pre-processing steps using `ffmpeg`. To remove high-energy, low-frequency noise, we used a high-pass filter, rolling off the audio signal below 100Hz (at a rate of 12 dB per octave). Then, we applied adaptive limiting to the audio in two phases. First, we used a applied a filter designed to achieve broadcast-standard loudness normalization (EBU R128), dynamically adjusting to an integrated loudness of –23dB. Second, to remove large peaks (e.g. when a microphone was bumped) we applied a look-ahead limiter set to prevent the signal from exceeding –6dB. Finally, we downsampled and downmixed to a monaural 16KHz (discarding sounds over 8KHz, typical for ASR) and extracted each segment to individual WAV files.

### B.3 Transcription Procedures

Phonemic transcriptions were broad with conventions originally developed by our laboratory for the purposes of use with a computer algorithm. For this project, we aimed to apply previously developed conventions in a way that captured some degree of phonetic detail if and when phonemic boundaries were crossed. To this end, research assistants received training from a licensed speech-language pathologist on some typical coarticulation processes and dialectical patterns observed in the participant sample, namely those that could be represented using broad phonemic notation.

### B.4 Transcript Pre-Processing for ASR

For ASR purposes, the IPA transcripts were converted to ARPAbet. The full mapping of IPA to ARPAbet symbols is shown in Table 8. Similar to conventional ASR preparation (Lopes and Perdigao, 2011b), some phonemes were combined or removed: /ə/ and /ʌ/ became `AH`, /ɝ/ and /ɚ/ became `ER`, and glottal stops (/ʔ/) were removed from the transcripts. We used a special symbol `SPN` for instances where transcribers noted unintelligible words or speech noises (e.g. laughing, coughing).

## C More on baseline model training

The model was fine-tuned for 12,000 total iterations (401 epochs), linearly ramping up to a learning rate of $5 \times 10^{-5}$ over the first 4000 iterations. For the first 2000 iterations, we froze all but the newly-initialized weights, priming only the output layer. For the final model, we restored the model to the point when it showed the minimum PER on the validation set, at 5964 iterations (200 epochs). We used a maximum batch size of 6.4 million frames of audio (400 seconds). Figure 2 shows the progression of the model's loss over the course of training.

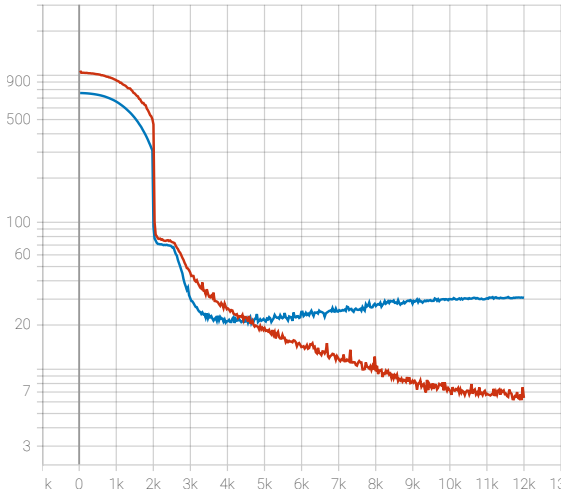


Figure 2: CTC loss over number of updates for the baseline model (at 30 updates per epoch). The red line is loss computed for the train set, and the blue line is loss computed for the test metric.