HHS Public Access

Author manuscript

Int IEEE EMBS Conf Neural Eng. Author manuscript; available in PMC 2024 September 25.

Published in final edited form as:

Int IEEE EMBS Conf Neural Eng. 2023 April ; 2023: . doi:10.1109/ner52421.2023.10123751.

Using Automatic Speech Recognition to Measure the Intelligibility of Speech Synthesized from Brain Signals

Suvi Varshney[*,†,‖], **Dana Farias**[‡], **David M. Brandman**[*,§], **Sergey D. Stavisky**[*,†,§], **Lee M. Miller**[‡‡,¶,‖,§]

[*]Department of Neurological Surgery, University of California, Davis

[†]Computer Science Graduate Group, University of California, Davis

[‡]Department of Physical Medicine and Rehabilitation, University of California, Davis

[‡‡]Department of Neurobiology, Physiology, and Behavior, University of California, Davis

[¶]Department of Otolaryngology / Head and Neck Surgery, University of California, Davis

[‖]Center for Mind & Brain, University of California, Davis

Abstract

Brain-computer interfaces (BCIs) can potentially restore lost function in patients with neurological injury. A promising new application of BCI technology has focused on speech restoration. One approach is to synthesize speech from the neural correlates of a person who cannot speak, as they attempt to do so. However, there is no established gold-standard for quantifying the quality of BCI-synthesized speech. Quantitative metrics, such as applying correlation coefficients between true and decoded speech, are not applicable to anarthric users and fail to capture intelligibility by actual human listeners; by contrast, methods involving people completing forced-choice multiple-choice questionnaires are imprecise, not practical at scale, and cannot be used as cost functions for improving speech decoding algorithms. Here, we present a deep learning-based "AI Listener" that can be used to evaluate BCI speech intelligibility objectively, rapidly, and automatically. We begin by adapting several leading Automatic Speech Recognition (ASR) deep learning models – Deepspeech, Wav2vec 2.0, and Kaldi – to suit our application. We then evaluate the performance of these ASRs on multiple speech datasets with varying levels of intelligibility, including: healthy speech, speech from people with dysarthria, and synthesized BCI speech. Our results demonstrate that the multiple-language ASR model XLSR-Wav2vec 2.0, trained to output phonemes, yields superior performance in terms of speech transcription accuracy. Notably, the AI Listener reports that several previously published BCI output datasets are not intelligible, which is consistent with human listeners.

Index Terms—

Brain Computer Interface (BCI); Automatic Speech Recognition (ASR); Speech Intelligibility; Speech Synthesis; Performance Metrics

Corresponding author: sstavisky@ucdavis.edu.

[§]These senior authors contributed equally to this work.

I. Introduction

Restoring communication is a top priority for people who cannot speak after neurological injuries such as stroke or ALS. One promising approach to restore patients' lost ability to speak is to bypass the damaged parts of their nervous system using a brain computer interface (BCI): a device that links the brain to external devices. Previous intracortical BCIs [1] successfully provided point-and-click and handwriting communication: paralyzed users attempted to move their hand, and algorithms decoded that neural data to make character selections (with speeds of 8 to 19 words per minute). However, a much faster (150 words per minute) and more intuitive approach to restoring communication would be to directly decode attempted speech from the associated brain activity. There have been several exciting recent demonstrations of speech reconstruction from neural activity [2]–[5].

However, there is a critical gap in the field: there are no established metrics for quantifying decoded speech when the ground truth is not known – that is, when the BCI user can't speak. Quantification of BCI-synthesized speech would help us evaluate and compare methods for accurately communicating the user's intended speech. Two different flavors of approaches have been described for quantifying speech: the first is to quantify how similar true and synthesized audio is in terms of low-level physical features, for example using metrics like correlation coefficients of the spectral power in different frequency bands [2]–[4]. These approaches can't be directly applied when there is no true speech available, as would be the case for speech BCI users with anarthria. The second, which *are* applicable to a synthesized-only situation, is to have human listeners report if they understood the speech, for example in forced-multiple choice online questionnaires [3]. However, this is slow, labor-intensive, subjective, and does not assess the BCI speech at a more fine-grained resolution (e.g., individual phonemes). Ideally, the field needs an automated metric that (1) could facilitate comparisons across different studies using the same scoring metric, (2) would scale to larger datasets for BCI speech, and (3) could accurately capture human speech *intelligibility*. Such an automated metric could (4) also be used as a loss function to be incorporated into novel decoding algorithms.

To meet this need, our goal is to create an 'AI Listener' (Figure 1), which takes as inputs an audio signal (such as that generated by a speech BCI) and the written ground truth transcript of what the speaker said (or attempted to say). The output of this software tool is a set of metrics with varying degrees of granularity. A future, longer-term goal is to validate our ASR metrics against intelligibility scores by professionally trained human listeners (i.e., speech-language pathologists).

Here we report our first steps towards this goal. We present a comparison of the performance of several popular ASR techniques [6]–[9] when applied to healthy speech and to BCI synthesized speech from our own work [2] and two other recently published examples [3], [4]. The Wav2vec 2.0 architecture, trained on multilingual speech corpora, provided a highly accurate transcription of a healthy speech dataset and the most accurate transcription of dysarthric speech datasets, which we evaluated as a proxy for speech that is of intermediate intelligibility. We adapted Wav2vec 2.0 to provide phoneme error rates, which we view as a useful finer-grained metric of speech BCI output accuracy. Importantly, however,

we found that each of the ASR techniques reproduces the finding that the state-of-the-art published BCI-synthesized voices are unintelligible. Nonetheless, we believe that our adapted Wav2vec 2.0 provides a promising starting point towards the goal of an 'AI Listener' for speech BCI assessment.

II. Methods

A. Automated speech recognition models

We considered four different candidate models for our "AI Listener": DeepSpeech [8], two different versions of Wav2vec 2.0 [6], [7] and the Kaldi speech recognition toolkit [9]. We use Mozilla's implementation of DeepSpeech and the latest version of their pre-trained model with an included language model. For the Kaldi speech recognition toolkit, we use their Time Delay Neural Network (TDNN) with the included Recurrent Neural Network (RNN) based Language Model. The Kaldi project uses a Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) to align phones to audio followed by the TDNN for sequence prediction. In contrast to Kaldi, the Wav2vec 2.0 ASR is a self-supervised, Transformer-based architecture trained on multiple corpora. We use the XLSR [6] which is trained on multilingual speech corpora, with English as the language most heavily used for training. The XLSR version of the Wav2vec 2.0 large model, fine-tuned on an English corpus [10], uses English characters as the output units of the acoustic model. This model was trained without an implicit language model, and produces quite accurate transcripts, which are grammatically correct in most cases. We also use XLSR [7] to output International Phonetic Alphabet (IPA) phonemes for comparison, which classify audio into English and non-English phonemes. To further produce words and sentences from the phoneme output, we trained our own language model, using the t5 pretrained transformer model [18] as the base and the Librispeech train corpus [11]. The training data was prepared using a phonemizer [15] with eSpeak engine to convert the Librispeech transcripts to phonemes.

B. Speech datasets

We compare these models on a variety of speech corpora, comprised of healthy speech, speech from individuals with dysarthria, and BCI speech generated from neural signals. For healthy speech, we use the Librispeech test set [11]. For the BCI-synthesized speech, we use participant T5's results from ('Wilson2020') [2], participant 5 from ('Herff2019') [4], and the four sentences provided in the supplementary video of ('Anumanchipallli2019') [3]. For the speech of individuals with dysarthria, we use the public Nemours dataset [12], TORGO dataset [13], and UA Speech Dataset [14]. The audio from all of these datasets was converted to single channel at 16Khz. We use a phonemizer [15] with eSpeak engine to convert the English text prompts to their IPA phonemes.

To provide the reader with a sense of the intelligibility of these dysarthric datasets [12]–[14], we scored 51 randomly sampled audio files (20 each from Nemours and UA Speech, 11 from TORGO) using the 9-point scale described in Frenchay Dysarthria Assessment [17]. This method categorizes the speech into 5 categories, ranging from Normal (A) to Profound (E). The categories Normal (A) to Moderate (C) are intelligible, where most words are decipherable with careful or repeated listening. The categories Moderate-Severe (C-D) to

Profound (E) are unintelligible, where the majority of words are undecipherable even after careful and repeated listening. The sample from the Nemours dataset [12] ranges from the Mild-Moderate (B-C) category to the Profound (E) category with the median intelligibility falling into the Severe (D) category. The speech in the Torgo dataset ranges from the Normal (A) category to the Profound (E) category, with a median rating lying in the Moderate (C) category. The samples from the UA Speech dataset range from Mild (B) to Profound (E) category with the median lying in the Moderate-Severe (C-D) category.

C. Performance metrics

To compare the ASR models, we use three metrics: the Word Error Rate (WER), Character Error Rate (CER), and Phoneme Error Rate (PER). We reasoned that representing speech as phonemes, and subsequently PER, is more appropriate than CER since 1) IPA is a larger set than English characters and has more acoustic granularity (and consistency), and 2) current BCI decoding approaches often rely on kinematic decoding of the movements of the human vocal apparatus [3] which is more directly related to acoustic-phonetics than characters. PER and WER are complementary metrics; we propose that PER will more closely align to the level of speech representation as a human perceives it in the moment, while the perception of words (as captured by WER) involves a language model and may be more analogous to how intelligible a human listener finds a whole utterance upon its completion (when context and language priors help the listener make sense of what they heard). Here we use CER as an additional model comparison method to be consistent with the ASR literature and demonstrate that our implementations are performing as expected.

All three of these metrics can yield values above 1. For example, consider if the source sentence is “hello” and the destination is “he will”. The WER here would be the number of edits (insertion, substitution or deletion) divided by the total number of words in the source. So the WER in this example is 2.

We use the Python implementation of DeepSpeech from Mozilla’s GitHub repository. We obtain the output transcripts from the DeepSpeech model to compute the WER and disable the scorer of the model to obtain the CER. Similarly, for Kaldi’s TDNN model, we obtain words and disable the language model to obtain the phonemes. For the XLSR-Wav2vec 2.0 English model, we obtain the characters as the output for computing WER and CER, while on the XLSR-Wav2vec 2.0 model, we get phonemes as output to compute PER. Figure 2 summarizes the models and their available evaluation metrics.

D. Hardware Setup

To train and test the models, we use an Intel Xeon CPU with 48 Cores, 168 GB of RAM, along with a 12 GB Nvidia Titan V GPU. On a test utterance of duration 4.7 seconds, average inference time was 3.2 seconds for DeepSpeech, 450 milliseconds for XLSR-Wav2vec 2.0, and 4.1 seconds for the Kaldi model.

III. Results

We evaluated the four models’ ability to transcribe healthy, dysarthric, and BCI-synthesized speech into output series of characters, phonemes, and words (Figure 3). The XLSR model

performed best on the healthy and dysarthric speech datasets, which we believe indicates it will be particularly useful for assessing speech BCIs. That said, none of the models (including XLSR) suggest that the BCI speech is intelligible - which is accurate, given that all three of the speech BCI papers report that the output was largely unintelligible [2]–[4].

We also present the confusion matrices for the phoneme XLSR model on 3 groups of datasets in Figure 4. Since the XLSR model can predict phonemes outside of the English IPA, we mark the second to last row as “other” (empirically this design choice seems to help the performance of this state-of-the-art architecture [6], perhaps by not forcing every sound to be classified as a legitimate English phoneme). Similarly, for any phonemes in the dataset that are not in the output space of the model, we mark the second to last column as “other”. Since this is a sequence prediction problem, we do not have a one-to-one correspondence between the predicted output and the ground truth. We therefore mark the last row and last columns (separated by the thin white lines) for insertion and deletion operations on the predicted output to reach the ground truth. The “insert” row and “delete” column sum to 100% and all the other columns (excluding the last row elements) sum to 100%.

One interesting trend that we notice in the confusion matrices of the dysarthric speech and the BCI speech is the classification of phonemes into the non-English phoneme group (“other” row). This may be due to the “foreign”-sounding phonemes encountered in the BCI and the dysarthric speech. This also shows that the model is able to recognise sounds as not being English phonemes, rather than forcing them into the most likely English phoneme class, similar to a human evaluator, who may tag the sound as unrecognizable if it does not resemble an English phoneme.

IV. Discussion

Establishing reliable, normative automated quality metrics is important for the nascent speech BCI field: these will allow researchers to compare performance across different methods and between studies from different groups (e.g., comparing speech BCIs that use different algorithms, record from different brain areas, or use different types of brain sensing technologies). For example, other communication BCI subfields using discrete selection (e.g. via SSVEP, P300, or point-and-click cursor control) have benefited from the rigorous use of achieved bitrate [16]. Our results validate that current BCI-synthesized speech and the highly dysarthric speech we evaluated are not intelligible, which is qualitatively consistent with human listeners also not perceiving this speech as intelligible. It remains to be seen how well these ASR methods perform on (hoped for) future higher-accuracy BCI-synthesized speech. We are developing these speech BCI metrics in parallel with the ongoing development of better speech synthesis BCIs [5], under the belief that having the right evaluation tools available will accelerate the overall endeavour.

In this work, we’ve shown that the Wav2vec 2.0 architecture – pre-trained to produce phoneme outputs - has very high accuracy in transcribing healthy spoken English. This is likely because the Transformer architecture and the multilingual training of this state-of-the-art model boosted its performance relative to previous ASR models. Importantly, it should

be easily adaptable to other languages, given that this was a principal motivation behind the original XLSR project [6].

XLSR-Wav2Vec 2.0 as used here is causal (not using future information) when calculating PER, and acausal (it uses the full utterance) when calculating WER. We view the PER as a putative proxy for how well a human listener can identify a phoneme as it is being spoken (where the preceding history of speech is also used to provide a prior in the human brain, as it is in the deep learning model). The WER is a putative proxy for how well a human can identify the words spoken at the end of an utterance, where whole-sentence context is available both to people and to the model (via the post-processing language model). Both PER and WER, we believe, will ultimately be useful for evaluating intelligibility holistically while maintaining high specificity for errors.

Several major steps remain before this AI Listener is ready to serve its intended role as a speech BCI scorer. The most critical is to compare the intelligibility metrics (PER and WER and any additional metrics developed in the future) against the same metrics from annotations provided by a number of diverse human listeners. The goal is to establish a monotonic relationship between AI Listener and human scores. This may require adapting the ASR model; we note that the goal of traditional ASR is to maximize accuracy (potentially even beyond typical human ability), whereas here our explicit goal is to maximize consistency with human listeners. A related limitation of the present work is that we essentially evaluated the model on speech data of two extremes: easily intelligible, and unintelligible. Future work will need to fill out this range (and validate it against human listeners) with speech of intermediate quality; this data could be future BCI-generated voice (if the field advances fast enough), or it may be artificially degraded healthy speech or intermediate dysarthric speech.

Acknowledgment

This work was supported by the National Institutes of Health through the UC Davis Health Clinical and Translational Science Center (CTSC: NIH UL1-TR001860) and a UC Davis Faculty Senate New Research Initiative Grant. S.D.S. is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. We thank Drs. M. Slutzky and C. Herff for providing the Herff2019 speech output data.

References

[1]. Willett FR, Avansino DT, Hochberg LR, Henderson JM & Shenoy KV High-performance brain-to-text communication via handwriting. Nature 593, 249–254 (2021). [PubMed: 33981047]

[2]. Wilson GH, Stavisky SD, Willett FR, Avansino DT, Kelemen JN, Hochberg LR, Henderson JM, Druckmann S, Shenoy KV. Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. J Neural Eng. 2020 Nov 25;17(6):066007. doi: 10.1088/1741-2552/abbfef. [PubMed: 33236720]

[3]. Anumanchipalli Gopala K., Chartier Josh, and Chang Edward F.. “Speech synthesis from neural decoding of spoken sentences.” Nature 568.7753 (2019): 493–498. [PubMed: 31019317]

[4]. Herff C, Diener L, Angrick M, Mugler E, Tate MC, Goldrick MA, Krusienski DJ, Slutzky MW and Schultz T (2019) Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices. Front. Neurosci 13:1267. doi: 10.3389/fnins.2019.01267. [PubMed: 31824257]

[5]. Wairagkar Maitreyee, Hochberg Leigh R, Brandman David M and Stavisky Sergey D. “Decoding Intracortical Neural Activity from Dorsal Motor Cortex.” In 2023 11th International IEEE/EMBS Conference on Neural Engineering (NER), (in press).

[6]. Conneau Alexis, Baevski Alexei, Collobert Ronan, Mohamed Abdelrahman, and Auli Michael. “Unsupervised cross-lingual representation learning for speech recognition.” arXiv preprint arXiv:2006.13979 (2020).

[7]. Xu Qiantong, Baevski Alexei, and Auli Michael. “Simple and effective zero-shot cross-lingual phoneme recognition.” arXiv preprint arXiv:2109.11680 (2021).

[8]. Hannun Awni, Case Carl, Casper Jared, Catanzaro Bryan, Diamos Greg, Elsen Erich, Prenger Ryan et al. “Deep speech: Scaling up end-to-end speech recognition.” arXiv preprint arXiv:1412.5567 (2014).

[9]. Povey Daniel & Ghoshal Arnab & Boulianne Gilles & Burget Lukáš & Glembek Ondrej & Goel Nagendra & Hannemann Mirko & Motlíček Petr & Qian Yanmin & Schwarz Petr & Silovský Jan & Stemmer Georg & Vesel Karel. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

[10]. Grosman J (2021). Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

[11]. Panayotov Vassil, Chen Guoguo, Povey Daniel, and Khudanpur Sanjeev. “Librispeech: an asr corpus based on public domain audio books.” In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5206–5210. IEEE, 2015.

[12]. Menendez-Pidal X, Polikoff JB, Peters SM, Leonzio JE and Bunnell HT, “The Nemours database of dysarthric speech,” Proceeding of Fourth International Conference on Spoken Language Processing. IC-SLP ‘96, 1996, pp. 1962–1965 vol.3, doi: 10.1109/ICSLP.1996.608020.

[13]. Rudzicz Frank, et al. TORGO Database of Dysarthric Articulation LDC2012S02. Web Download. Philadelphia: Linguistic Data Consortium, 2012.

[14]. Kim Heejin, Hasegawa-Johnson Mark, Perlman Adrienne, Gunderson Jon, Huang Thomas S., Watkin Kenneth, and Frame Simone. “Dysarthric speech database for universal access research.” In Ninth Annual Conference of the International Speech Communication Association. 2008.

[15]. Bernard M, & Titeux H (2021). Phonemizer: Text to Phones Transcription for Multiple Languages in Python. Journal of Open Source Software, 6(68), 3958. 10.21105/joss.03958.

[16]. Kao JC, Nuyujukian P, Ryu SI & Shenoy KV A high-performance neural prosthesis incorporating discrete state selection with hidden Markov models. IEEE Transactions on Biomedical Engineering 9294, 1–1 (2016).

[17]. Enderby Pamela M. (Pamela Mary), 1949-. Frenchay Dysarthria Assessment. San Diego, Calif. :College-Hill Press, 1983.

[18]. Raffel Colin, Shazeer Noam, Roberts Adam, Lee Katherine, Narang Sharan, Matena Michael, Zhou Yanqi, Li Wei, and Liu Peter J.. “Exploring the limits of transfer learning with a unified text-to-text transformer.” The Journal of Machine Learning Research 21, no. 1 (2020): 5485–5551.

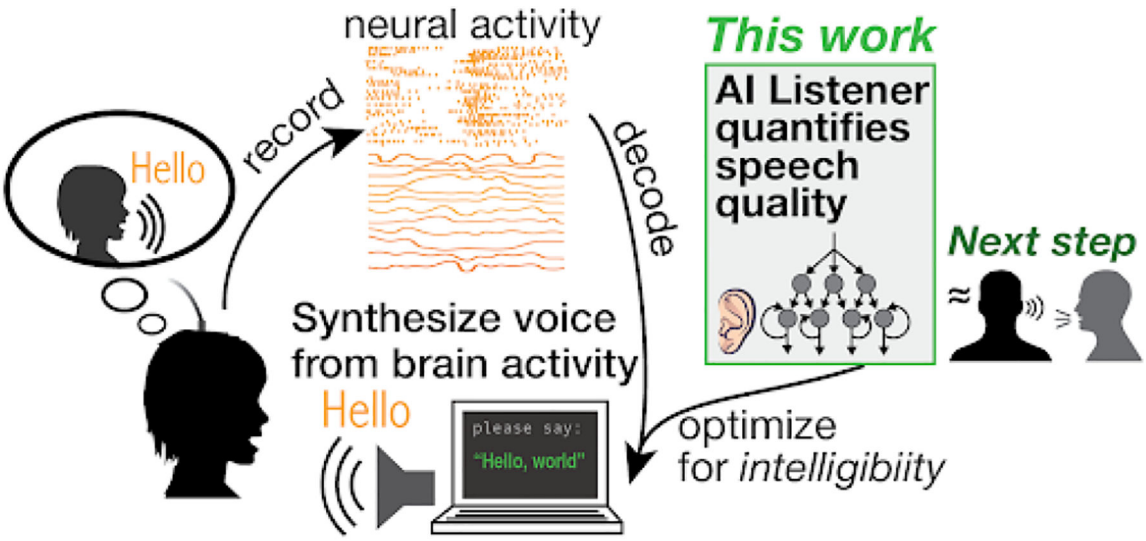

Fig. 1. Approach overview.

This work focuses on developing an ASR tool to score the intelligibility of synthesized speech, which would fill an important gap in the overall effort to build a speech restoration brain-computer interface. An important next step is to validate that this automated metric scores intelligibility in a way that is similar to actual human listeners.

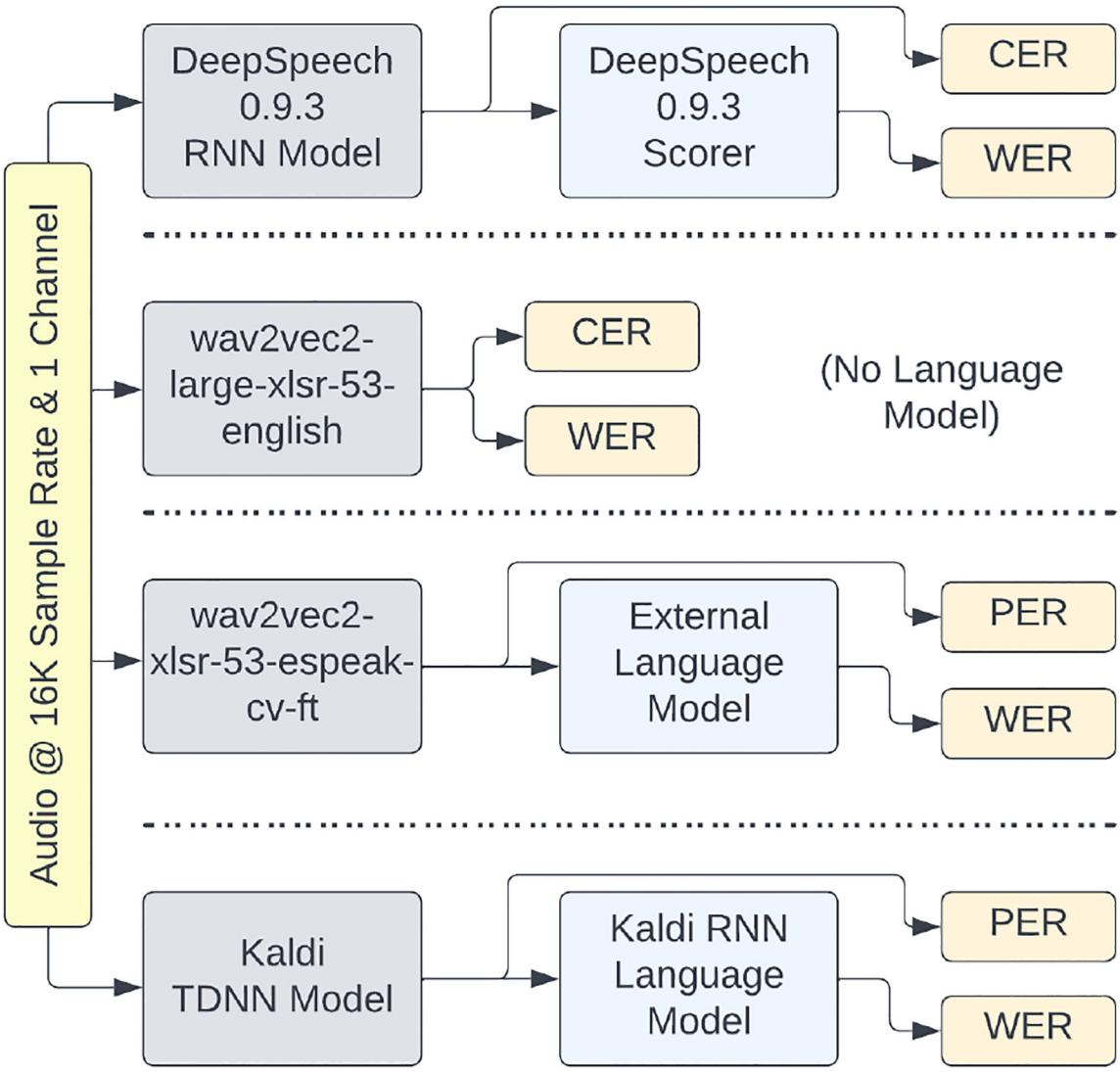

Fig. 2. Methods compared.

Speech audio intelligibility was assessed using four different deep learning automated speech recognition models. Final speech intelligibility metrics are phoneme error rate (PER), character error rate (CER), and word error rate (WER).

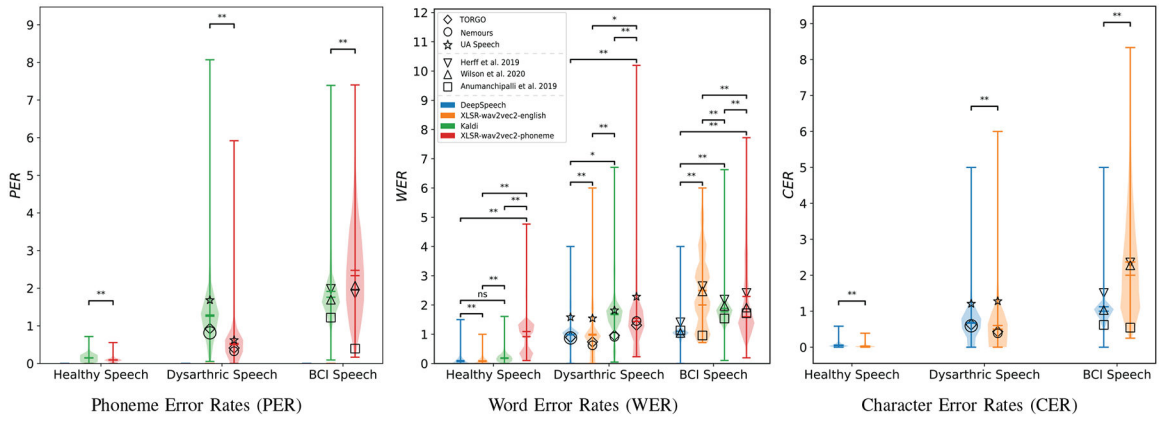

Fig. 3. Candidate Model Performance.

Violin plots show the distribution of performance across all utterances, with colored horizontal lines showing the distribution mean and median. WER is defined as the number of edits (insertion, substitution or deletion) of words in the output of the model to reach the intended sentence. CER and PER are defined in the same way but for characters and phonemes in a sentence, respectively. Note that PER, WER, and CER can be above 1, but any value approaching 1 (or higher) implies that the speech is not intelligible to this model. The black symbols represent the mean error rates of the corresponding individual datasets for that column's model. The healthy speech has only one dataset, Librispeech. (ns: p-value >= 0.01, *: p-value < 0.01, **: p-value < 0.005)

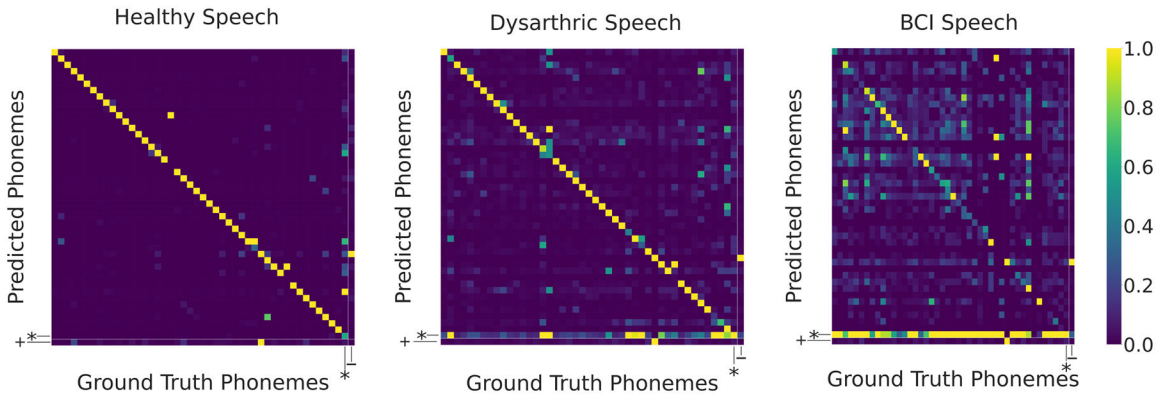

Fig. 4. Confusion matrices for the output phonemes of XLSR-Wav2vec 2.0.

The datasets are grouped into 3 categories: Healthy Speech (the Librispeech test set), Dysarthric Speech (aggregating TORGO, Nemours and UA Speech) and BCI Speech (aggregating the data from Wilson et al. 2020, Herff et al. 2019, Anumanchipalli et al. 2019). Asterisks * indicate rows where the model infers phonemes outside of English IPA or columns where phonemes in the dataset are not in the output space of the model. The "–" symbol indicates when a deletion operation was required to match ground truth, and a "+" indicates when an insertion operation was required. The colorbar corresponds to the proportion of events.