INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

On the benefits of self-supervised learned speech representations for predicting human phonetic misperceptions

Santiago Cuervo, Ricard Marxer

Université de Toulon, Aix Marseille Université, CNRS, LIS, France
santiago.cuervo@lis-lab.fr, ricard.marxer@lis-lab.fr

Abstract

Deep neural networks (DNNs) trained by self-supervised learning (SSL) have recently been shown to produce representations similar to brain activations for the same speech input. Can SSL representations help to explain human speech perception errors? Aiming to shed light on this question, we study their use for phonetic misperception prediction. We extract representations from wav2vec 2.0, a recent SSL architecture for speech, and use them to compute features for a model predicting the presence of phonetic perception errors in speech-in-noise signals. We perform our experiments on a corpus of over 3000 consistent word-in-noise confusions in English. We consider multiple SSL-based features and compare them against conventional acoustic baselines and features obtained from DNNs fine-tuned through supervised learning for ASR. Our results show the superiority of SSL representations when extracted from the proper layer, further suggesting their potential to model human speech perception.

Index Terms: speech perception, intelligibility prediction, sublexical intelligibility, self-supervised learning, speech-in-noise

1. Introduction

A prominent theory of the brain frames it as an inference engine that optimizes for prediction performance based on context [1]. Similarly, the core idea behind *self-supervised learning* (SSL), a recently popularized machine learning paradigm, is to extract useful representations from data by learning to predict it based on the context in which it occurs. Accordingly, there has been growing interest in exploring the potential role of SSL in some brain processes [2, 3].

In [3], the authors exposed significant correlations under the same speech input between representations obtained from wav2vec 2.0 [4], a SSL algorithm for speech processing, and brain activations in response to speech. In this paper, we continue this line of research connecting SSL and human speech perception. We investigate the correlations between SSL representations and consistent human phonetic misperceptions of speech-in-noise. Specifically, we build models that use features computed from wav2vec 2.0 representations to predict the elicitation of human misperceptions at each phone in single-word speech-in-noise recordings.

We consider intrusive (ie. with access to the clean speech waveform) and non-intrusive features used in the intelligibility prediction literature, compare SSL representations against spectrum-based baselines, and analyze the effect of supervised ASR fine-tuning on the quality of representations for phonetic misperception prediction. Our main contributions are:

- Empirically demonstrating the superiority of wav2vec 2.0 representations in phonetic misperception prediction relative to conventional acoustic features and commonly used representations obtained from ASR models.
- Establishing the proper level in wav2vec 2.0 from which to extract representations for phonetic misperception prediction, further illustrating its hierarchical distribution of information.
- Proposing the *masked language modeling* (MLM) loss as a non-intrusive index for speech intelligibility, and demonstrating that it is predictive of phonetic misperceptions.

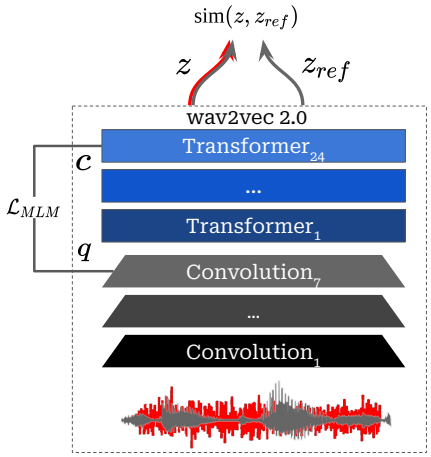

Figure 1: *wav2vec 2.0 architecture and its features for speech-in-noise misperception prediction.*

2. Phonetic misperception prediction

For a sequence of uttered phones $p_1, \ldots, p_m$ perceived by a listener as $\hat{p}_1, \ldots, \hat{p}_n$, we train a model to predict for each phone in p, if it was correctly identified in $\hat{p}$. The prediction is conditioned on a sequence of features extracted from the uttered waveform. Such predictive model can be useful to provide an optimization objective for speech enhancement systems [5], and perhaps to gain insights on human speech perception.

We used self-supervised learned speech representations from wav2vec 2.0 (Figure 1) to compute features for the predictive model. Wav2vec 2.0 consists of three main modules: 1) a strided convolutional network transforms the speech waveform consisting of K samples $x \in \mathbb{R}^{1\times K}$ into a sequence of frame encodings $y \in \mathbb{R}^{d\times T}$, 2) a vector-quantizer discretizes each encoding, producing a sequence of codes $q \in \mathbb{Z}^{1\times T}$, and 3) a stack of transformer blocks with bidirectional attention produces a sequence of context vectors $c \in \mathbb{R}^{d\times T}$.

The model is trained using a MLM loss. During self-

10.21437/Interspeech.2023-1476

supervised training some encodings in y are masked, and the model is optimized to predict the codes in q corresponding to the masked encodings, conditioned on their corresponding context vectors in c. For more details we refer the reader to [4].

After training, the frame encodings and context vectors can be used as speech representations for downstream tasks. We compute multiple features based on them:

Raw speech representations $z \in \mathbb{R}^d$. An encoding or context vector learned by wav2vec 2.0. In an intrusive setup, we process the clean reference waveform to obtain z_{ref} and concatenate it to z: $\text{cat}(z, z_{ref}) \in \mathbb{R}^{2d}$. In [6] the authors demonstrated that raw SSL representations are predictive of macroscopic (word and sentence level) speech intelligibility indices.

Similarity to a clean reference signal $\text{sim}(z, z_{ref}) \in [-1, 1]$. It is the cosine similarity between a representation from the corrupted waveform z and a representation from the noise-free reference signal z_{ref}. In [7] it was shown to correlate with macroscopic speech-in-noise intelligibility for representations extracted from a DNN trained through supervised learning.

Masked language modeling loss $\mathcal{L}_{MLM}(c_i) \in \mathbb{R}$. It is the MLM loss obtained when using c_i for predicting q_i. To compute $\mathcal{L}_{MLM}(c_i)$, we mask y_i and the 4 closest frames to it (with each frame lasting 25 ms, it is slightly above the mean phone duration). The brain seems to perform statistical inferences based on context [2, 8], so we can expect that when faced with uncertainty about a heard phone, it would use contextual information to predict the most likely phonetic category. We hypothesized that the MLM loss, which captures the idea of context-based prediction, correlates with speech-in-noise intelligibility.

3. Experiments and results

3.1. Setup

3.1.1. Data and pre-processing

We performed our experiments on the English Consistent Confusion Corpus [9], a dataset created by gathering perceived responses from 15 listeners to common English words mixed with random noise maskers. The corpus is composed of words-in-noise misperceived in the same way by at least 6 of the 15 listeners. These consistent hearing errors are valuable targets to test models of speech perception. For each of its 3207 consistent misperceptions, the corpus provides the speech and masker 16 kHz waveforms that produced them, and the responses given by the listeners. The corpus uses three different types of noise masker: stationary speech shaped noise, four-speaker speech babble, and three-speaker babble modulated noise.

We split the dataset in train, validation, and test sets in portions of 80/10/10%, respectively. Considering that the properties of the produced misperceptions are dependent on the masker type [10], and that other factors such as speaker gender and identity are roughly balanced across maskers, the splits were made stratified by masker type.

Our task is to predict misperceptions at individual phones. To obtain prediction targets, we computed the edit scripts between the perceived[1] and target phonetic transcripts. An edit operation on a specific phone indicates a phonetic perception error. To locate phones (hence misperceptions) in time, we used the Montreal Forced Aligner [11] to generate phonetic alignments for the clean speech signal. These alignments are publicly available at `https://tinyurl.com/CCalign`. We considered two prediction resolutions: phone-wise and frame-wise. As an example, consider the utterance "bit" (`B IH T`)[2], with each phone lasting 2, 3 and 1 frames, respectively. If the word is confused as "bet" (`B EH T`), the prediction targets phone-wise and frame-wise would be `0 1 0` and `0 0 1 1 1 0`, respectively. Frame-wise predictions are made for each speech representation, according to the sampling rate of the feature extractor. In phone-wise predictions, we obtain a phone representation by averaging the frame representations falling within the time window of a phone according to the alignments.

3.1.2. Self-supervised model

We extract SSL representations from the wav2vec 2.0 LARGE model trained on 60 k hours of speech [4]. The convolutional encoder contains 7 blocks of temporal convolutions with 512 channels, strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,3,3,2,2). This results in an output frequency of 49 Hz with a stride of about 20 ms between each sample, and a receptive field of 25 ms of audio. The transformer is composed of 24 blocks with model dimension 1024, inner dimension 4096, and 16 attention heads. The quantizer uses product quantization with 2 codebooks of 320 entries each, and embedding dimension of 768. The total number of parameters is 317 million. We applied PCA dimensionality reduction to a 256-dimensional space to the representations. This was done after observing overfitting due to the large numbers of input parameters.

3.1.3. Baselines and metrics

We compare the performance on phonetic misperception prediction of SSL features against two time-frequency spectrum features: 1) the *Short-Time Objective Intelligibility* (STOI) [12], a classic intrusive index for speech-in-noise intelligibility, and 2) *Ratemaps*, a neurologically-inspired spectrogram used in intelligibility prediction [13]. When using ratemaps, we use frequency coefficients as z. To compute the STOI index we use the `pystoi`[3] package with the default parameters. For a given phone, we compute its STOI across the audio segment corresponding to it. To compute the ratemaps we use the `python_ratemap` package[4] using 55 channels.

In all experiments we evaluate the F1 score of the predictive model, and report its mean and standard deviation obtained across 10 independent runs. Precision and recall were balanced in all cases, therefore we do not report them separately.

3.1.4. Computing infrastructure and code

All experiments were run on a single node with an AMD EPYC 7313 16-Core Processor, 95 GB of RAM, and an A100 80 GB NVIDIA GPU. We make the code available at `https://github.com/tiagoCuervo/SSLPhoneticConf`.

3.2. Choosing the predictive model

Aiming for simplicity, we initially assumed independence from the rest of the utterance and made phone-wise predictions using only phone-located features. We used the Support Vector Classifier [14] implementation from Scikit-learn [15] as a prediction model. The regularization constant and the kernel used were chosen according to a 5-fold cross validation from the sets $10^{i \in \{-1,0,1,2\}}$ and {linear, RBF}, respectively. We used as fea-

[1] For all utterances we use the most common misperception.

[2] ARPAbet notation.

[3] `https://github.com/mpariente/pystoi`

[4] `https://github.com/rikrd/python_ratemap`

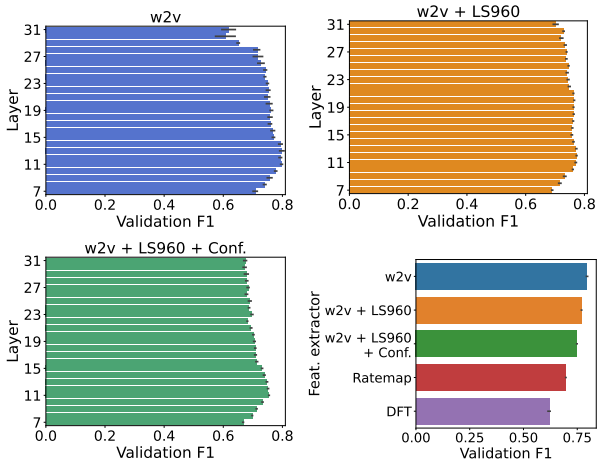

Figure 2: *Validation F1 scores across layers and feature extractors using* $\text{cat}(z, z_{ref})$ *as input features. The same patterns were observed for* z *and* $\text{sim}(z, z_{ref})$*. Wav2vec 2.0 representations from lower transformers (layers 11 to 14) perform best in all models. The plot comparing feature extractors (bottom right) shows the superiority of representations learned through SSL with respect to the ones fine-tuned for ASR and conventional acoustic features (where applies, it considers only the scores between layers 11 and 14).*

tures $\text{sim}(z, z_{ref})$, with z and z_{ref} being SSL representations, and obtained an F1 score of 0.585 ± 0.010.

The brain however exploits context, therefore our next step was to test if the prediction model could benefit from it. We experimented with a neighborhood of radius r centered around the phone. We used zero padding to handle edges. For $r = 1, 2,$ and 3, we obtained F1 scores of 0.603 ± 0.007, 0.624 ± 0.008, and 0.630 ± 0.007, respectively. This showed that the elicitation of phone misperceptions is dependent on context, and therefore on higher level linguistic features (as expected from [16, 10]), and that the self-attention mechanism in wav2vec 2.0 either does not capture the relevant context for phone perception error prediction, or this information is lost by phone-wise averaging and/or the computation of $\text{sim}(z, z_{ref})$.

According to these results, in the following experiments we used full sequences as inputs and a bidirectional LSTM [17] sigmoid classifier as predictive model. We used a single hidden layer with dropout regularization [18] ($p = 0.5$). For experiments using scalar features we used 32 hidden units. For models processing vectors we used 256 units and input dropout ($p = 0.2$). The model is trained using the Adam optimizer [19] with a learning rate of 0.0001 and a batch size of 256 to minimize a binary cross entropy loss. The model is trained until the validation F1 score does not improve for more than 15 epochs.

LSTM networks can handle long input sequences, therefore in all the following experiments we used frame-wise predictions to avoid the dependency on alignments at inference time.

3.3. Predictive power across layers

The features learned by wav2vec 2.0 are hierarchical, with intermediate transformer blocks capturing more phonetic information [20]. We evaluated representations extracted at each block $l \in \{7, 8, 9, \ldots, 31\}$ (ie. output of the convolutional encoder and all transformer blocks) in phonetic misperception prediction. Figure 2 and Figure 3 illustrate the obtained results using

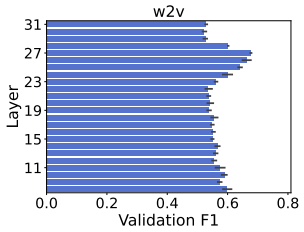

Figure 3: *Validation F1 scores across layers using* $\mathcal{L}_{MLM}(c)$ *as input features. Contrary to other features (Figure 2), the best performance occurs around layer 27, a high transformer layer.*

Table 1: *Phone misperception prediction performance using a bidirectional LSTM with various features and feature extractors. Best performing layer according to validation F1 shown in parenthesis, where applicable. Scores reported on test set. Representations learned purely through self-supervised learning are better predictors of human phonetic misperceptions.*

	Feature	Feature extractor	Test F1
Intrusive	STOI	DFT	0.595 ± 0.008
	$\text{sim}(z, z_{ref})$	Ratemap	0.661 ± 0.008
		w2v ($l = 12$)	**0.709 ± 0.006**
		w2v + LS960 ($l = 18$)	0.685 ± 0.001
		w2v + LS960 + Conf.($l = 23$)	0.619 ± 0.026
	$\text{cat}(z, z_{ref})$	Ratemap	0.710 ± 0.010
		w2v ($l = 12$)	**0.795 ± 0.004**
		w2v + LS960 ($l = 19$)	0.780 ± 0.004
		w2v + LS960 + Conf.($l = 17$)	0.766 ± 0.002
Non-intrusive	$\mathcal{L}_{MLM}(z)$	w2v ($l = 27$)	0.674 ± 0.004
	z	Ratemap	0.672 ± 0.002
		w2v ($l = 14$)	**0.756 ± 0.008**
		w2v + LS960 ($l = 10$)	0.739 ± 0.006
		w2v + LS960 + Conf. ($l = 19$)	0.689 ± 0.005

as features $\text{cat}(z, z_{ref})$ and $\mathcal{L}_{MLM}(z)$, respectively.

In Figure 2 (top-left), representations from the lower transformer blocks perform best, consistently outperforming the representations from the convolutional encoder. The prediction score drops sharply for $l > 28$. The highest transformer block is trained to directly predict quantized representations from the convolutional encoder, and therefore the top blocks might be encouraged to encode lower-level, less phonetically-relevant information. Similar results are reported in [20] in terms of phonemic categorization performance. The same trends were observed when using $\text{sim}(z_{ref}, z)$ and z (not shown).

The scores obtained when using $\mathcal{L}_{MLM}(z)$ (Figure 3) initially tend to decrease, but rise after $l = 24$ and have a peak around $l = 27$. As before, performance drops to its lowest for $l > 28$. When computing $\mathcal{L}_{MLM}(z)$, the phone at which the misperception is predicted is mostly masked, and therefore the feature is computed from context. In this case, we hypothesize that patterns between phones given by lexical features could be most relevant, eg. in a phone bigram with the first phone masked and the second phone being 'Z', it could be predicted that the occluded phone is 'IH', in order to form the common word "is". Layer $l = 27$, being the highest before the performance drop, could be the one capturing these higher level patterns.

3.4. The effect of supervised fine-tuning

To determine if self-supervised learning is important for the prediction of phonetic perception errors, we compare the per-

formance of wav2vec 2.0 representations against models fine-tuned through supervised learning for ASR. We considered two of such models. The first used wav2vec 2.0 as backbone, and was fine-tuned to predict letter transcriptions using a Connectionist Temporal Classification (CTC) [21] loss on 960 hours of the LibriSpeech [22] (LS960) dataset. The second used the model fine-tuned on LS960 as backbone, and was fine-tuned to predict phonetic transcriptions using a CTC loss on the noisy utterances from the English Confusion Corpus.

For the model fine-tuned on LS960 we used the checkpoint provided in `https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec`. To fine-tune the model on the Confusion Corpus we followed the recipe available under the same link for fine-tuning on 1 h of labeled data (roughly the length of the Confusion Corpus).

The results presented in Figure 2 and Table 1 show that supervised fine-tuning does not bring any improvement upon the best performing SSL representations. On the contrary, across most layers ASR fine-tuning degrades predictive performance. ASR fine-tuned representations only outperform SSL representations consistently on the highest transformer blocks (the worst performing SSL features). Contrary to MLM, the CTC loss forces last blocks to focus on phonetic information, which results useful for predicting phonetic misperceptions.

The model fine-tuned on both LS960 and the Confusion Corpus performed the worst. This indicates that more fine-tuning and/or fine-tuning on speech-in-noise data results in further degradation of prediction performance. ASR training could make models more robust to noise by forcing them to focus on discrete (phonetic) properties. This robustness might harm their ability to detect confusing regions in the speech signal, which could be useful to predict the likelihood of a misperception.

4. Discussion

4.1. On the meaning of our results

In [3], the authors identified similarities between different wav2vec 2.0 layers and different regions of the brain associated to speech perception. Our results in Section 3.3 could be connected to their findings on the hierarchy of speech processing. Lower transformer blocks performing best in phonetic perception prediction could be connected to their reports on higher correlations between low transformers and activations in the Heschl's and Superior Temporal gyruses, regions of the cortex believed to play a role in phonetic categorization [23, 24]. [3] also showed that mid-high transformer blocks are more strongly correlated with activations in the Superior Temporal sulcus, a region of the cortex related to high level linguistic features [25, 26]. This could support our hypothesis on the peak performance in $l = 27$ with $\mathcal{L}_{MLM}(z)$ (Figure 3, Section 3.3) being due to exploiting linguistic patterns above the acoustic level.

The practical implications for intelligibility prediction research are evident from the study. Firstly, the results suggest that SSL representations should be chosen over supervised-learned ones, contrary to what has been done in [7, 27] for instance. Secondly, $\text{cat}(z, z_{ref})$ consistently and significantly outperforming $\text{sim}(z, z_{ref})$ as a feature for intelligibility prediction, indicates that learned non-linear functions over raw features should be preferred over linear similarity measures. This could improve the results reported by [7]. It is left to be seen if raw non-intrusive features also outperform non-intrusive methods based on ASR systems' uncertainty, such as [28, 29].

The hypothesis that the MLM loss correlates with intelligibility was also validated by our results. The MLM loss is a promising candidate for non-intrusive intelligibility prediction, as it does not require any speech transcriptions at any point of the pipeline used for its computation. In contrast, other currently proposed non-intrusive methods require ASR systems trained on transcribed speech [28, 29], which prevents them from being applied to low-resource languages. Unsupervised ASR [20] holds potential for improving such methods.

4.2. On limitations and future work

We would have liked to perform evaluations on more data, but we are limited by the scarcity of corpora for speech perception [27]. We consider extending our analyses to the Clarity Challenge corpus [5] of responses from hearing impaired listeners.

In Section 3.4 there should be an additional comparison with a model only fine-tuned on the Confusion Corpus in order to disentangle the effects of further fine-tuning, and fine-tuning on noisy data. However, we had issues to successfully fine-tune the base wav2vec 2.0 LARGE checkpoint trained on 60k hours of speech. This problem has been reported by other users[5].

More experiments should be done using other feature extractors to further validate the benefits of SSL. In this study we focused on wav2vec 2.0, motivated by its demonstrated similarities with human speech processing [3]. In the future, we plan to analyze representations from w2v-BERT [30] and Whisper [31], state-of-the-art SSL and ASR models, respectively.

5. Related work

This study shares similar goals with [3], which aimed to correlate wav2vec 2.0 representations with brain activations recorded using fMRI. Similarly, we correlate wav2vec 2.0 representations and human phonetic perception errors, another signal related to the human speech processing system. We arrive to similar conclusions on the benefits of SSL representations and the hierarchical distribution of information in wav2vec 2.0.

Our work also shares connections with the literature on intelligibility prediction based on DNN representations [27, 7, 28, 6, 32]. More relevantly, in [6], SSL representations were used and optimized to predict multiple speech intelligibility indices. However, in contrast to our phonetic sublexical focus, all these methods are aimed for intelligibility over words and sentences.

6. Conclusions

We have empirically demonstrated that SSL representations obtained from the lower transformer blocks in wav2vec 2.0 are better predictors of human phonetic perception errors, relative to conventional acoustic features, and representations from DNNs fine-tuned through supervised learning for ASR. Overall, our results reinforce the candidacy of self-supervised learning as a mechanism for speech perceptual learning in the brain. Additionally, we have proposed the MLM loss as a non-intrusive method for intelligibility prediction. This method does not require speech transcriptions for its computation, and could be useful on low-resource languages. Finally, we have analyzed our findings and outlined possible research directions.

7. Acknowledgements

We are grateful to the French National Research Agency for their support through the ANR-20-CE23-0012-01 (MIM) grant.

[5] `https://tinyurl.com/issueW2V`

8. References

[1] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010. [Online]. Available: https://doi.org/10.1038/nrn2787

[2] C. Caucheteux and J.-R. King, "Brains and algorithms partially converge in natural language processing," *Communications Biology*, vol. 5, no. 1, p. 134, 2022. [Online]. Available: https://doi.org/10.1038/s42003-022-03036-1

[3] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J.-R. King, "Toward a realistic model of speech processing in the brain with self-supervised learning," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=Y6A4-R_Hgsw

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, pp. 12 449–12 460.

[5] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech 2022*, 2022, pp. 3508–3512.

[6] R. E. Zezario, S. wei Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MTI-Net: A Multi-Target Speech Intelligibility Prediction Model," in *Proc. Interspeech 2022*, 2022, pp. 5463–5467.

[7] Z. Tu, N. Ma, and J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners," in *Proc. Interspeech 2022*, 2022, pp. 3488–3492.

[8] J. T. Hale, L. Campanelli, J. Li, S. Bhattasali, C. Pallier, and J. R. Brennan, "Neurocomputational models of language processing," *Annual Review of Linguistics*, vol. 8, no. 1, pp. 427–446, 2022. [Online]. Available: https://doi.org/10.1146/annurev-linguistics-051421-020803

[9] R. Marxer, J. Barker, M. Cooke, and M. L. Garcia Lecumberri, "A corpus of noise-induced word misperceptions for english," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL458–EL463, 2016. [Online]. Available: https://doi.org/10.1121/1.4967185

[10] M. L. G. Lecumberri, J. Barker, R. Marxer, and M. Cooke, "Language Effects in Noise-Induced Word Misperceptions," in *Proc. Interspeech 2016*, 2016, pp. 640–644.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[12] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[13] Y. Tang and M. Cooke, "Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions," in *Proc. Interspeech 2016*, 2016, pp. 2488–2492.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

[16] M. Cooke, M. L. García Lecumberri, J. Barker, and R. Marxer, "Lexical frequency effects in english and spanish word misperceptions," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. EL136–EL141, 2019. [Online]. Available: https://doi.org/10.1121/1.5090196

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[20] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=QmxFsofRvW9

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, p. 369–376.

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

[23] B. Khalighinejad, P. Patel, J. L. Herrero, S. Bickel, A. D. Mehta, and N. Mesgarani, "Functional characterization of human Heschl's gyrus in response to natural speech," *NeuroImage*, vol. 235, p. 118003, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811921002809

[24] H. G. Yi, M. K. Leonard, and E. F. Chang, "The encoding of speech sounds in the superior temporal gyrus," *Neuron*, vol. 102, no. 6, pp. 1096–1110, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627319303800

[25] S. M. Wilson, A. Bautista, and A. McCarron, "Convergence of spoken and written language processing in the superior temporal sulcus," *NeuroImage*, vol. 171, pp. 62–74, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811917310923

[26] P. E. Turkeltaub and H. Branch Coslett, "Localization of sublexical speech perception components," *Brain and Language*, vol. 114, no. 1, pp. 1–15, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0093934X10000556

[27] Ľuboš Marcinek, M. Stone, R. Millman, and P. Gaydecki, "N-MTTL SI Model: Non-Intrusive Multi-Task Transfer Learning-Based Speech Intelligibility Prediction Model with Scenery Classification," in *Proc. Interspeech 2021*, 2021, pp. 3365–3369.

[28] Z. Tu, N. Ma, and J. Barker, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction," in *Proc. Interspeech 2022*, 2022, pp. 3493–3497.

[29] A. M. Castro Martinez, C. Spille, J. Roßbach, B. Kollmeier, and B. T. Meyer, "Prediction of speech intelligibility with dnn-based performance measures," *Computer Speech & Language*, vol. 74, p. 101329, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821001224

[30] Y. Chung, Y. Zhang, W. Han, C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *CoRR*, vol. abs/2108.06209, 2021. [Online]. Available: https://arxiv.org/abs/2108.06209

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[32] J. Roßbach, B. Kollmeier, and B. T. Meyer, "A model of speech recognition for hearing-impaired listeners based on deep learning," *The Journal of the Acoustical Society of America*, vol. 151, no. 3, pp. 1417–1427, 2022. [Online]. Available: https://doi.org/10.1121/10.0009411