

Transfer Learning from Whisper for Microscopic Intelligibility Prediction

Paul Best, Santiago Cuervo, Ricard Marxer

▶ **To cite this version:**

Paul Best, Santiago Cuervo, Ricard Marxer. Transfer Learning from Whisper for Microscopic Intelligibility Prediction. Interspeech 2024, Sep 2024, Kos, Greece. pp.3839-3843, 10.21437/Interspeech.2024-2258 . hal-04683361

HAL Id: hal-04683361
https://hal.science/hal-04683361v1

Submitted on 4 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Interspeech 2024
1-5 September 2024, Kos, Greece

Transfer Learning from Whisper for Microscopic Intelligibility Prediction

Paul Best, Santiago Cuervo*, Ricard Marxer*

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

Abstract

Macroscopic intelligibility models predict the expected human word-error-rate for a given speech-in-noise stimulus. In contrast, microscopic intelligibility models aim to make fine-grained predictions about listeners' perception, e.g. predicting phonetic or lexical responses. State-of-the-art macroscopic models use transfer learning from large scale deep learning models for speech processing, whereas such methods have rarely been used for microscopic modeling. In this paper, we study the use of transfer learning from *Whisper*, a state-of-the-art deep learning model for automatic speech recognition, for microscopic intelligibility prediction at the level of lexical responses. Our method outperforms the considered baselines, even in a zero-shot setup, and yields a relative improvement of up to 66% when fine-tuned to predict listeners' responses. Our results showcase the promise of large scale deep learning based methods for microscopic intelligibility prediction.

Index Terms: speech perception, intelligibility prediction, Whisper, deep learning.

1. Introduction

Traditional speech intelligibility models measure how comprehensible speech is. Generally, they make predictions about the average number of words that may be heard correctly given a noisy speech stimulus. These models are referred to as *macroscopic*, as their prediction target is an expected value aggregated across many words or sentences, and multiple listeners. Macroscopic intelligibility prediction has been of great interest for its importance in developing speech enhancement related applications, such as hearing aids.

In the last decade, a different paradigm for intelligibility modeling has emerged: *microscopic intelligibility prediction* [1, 2]. These models aim to predict human perception of speech at a higher level of detail than macroscopic models. For instance, such model might be expected to predict a confusion matrix characterising phonetic misperceptions, or to predict a listener's lexical response to a specific noisy speech token. It is hoped that building these models will contribute to understanding the mechanisms underlying human speech perception, since they must reproduce more specific listener's behavior.

Throughout the last decade, we have also witnessed great advances in speech processing applications thanks to the popularization and scaling of deep learning methods. Training large-scale deep learning models from scratch on intelligibility prediction is usually not feasible due to the small amount of labeled data available. However, recent advances on representation learning showed that transfer learning from large-scale self-supervised or weakly-supervised pre-trained models is highly effective across a wide range of speech applications [3, 4, 5, 6]. These methods yielded the state-of-the-art models for macroscopic intelligibility prediction [7, 8], but little work has been done on applying them to microscopic modeling. Moreover, there is evidence that the representations learned by large scale pre-trained models are predictive of human brain speech processing [9], likely making them relevant to model speech perception with tasks such as microscopic intelligibility prediction.

In this paper, we explore the use of transfer-learning from a state-of-the-art large-scale deep learning model for automatic speech recognition (ASR) to microscopic intelligibility prediction. We focus on *Whisper* [5], an ASR model that approaches the accuracy and robustness of humans. We use it to tackle the task of predicting lexical responses, which is considered to be the most challenging on microscopic intelligibility modeling [1]. Our method outperformed the considered baselines, even in a zero-shot setup. When fine-tuned to predict listener's responses, it yielded a relative improvement of 66%. Overall, our results showcase the promise of transfer learning from large pre-trained models to microscopic intelligibility prediction.

2. Problem

We approach the problem of microscopic intelligibility prediction using the framework proposed by Marxer et al. [1]. They consider a setup in which there are a set of tuples (x, y, k), where $x \in \mathbb{R}^T$ is a noisy single-channel speech waveform of duration T, $y = [y_1, \ldots, y_n]$ with $y_i \in \mathcal{V}$ for $i = 1, \ldots, n$ is a vector of n unique responses perceived by a set of m listeners from a finite vocabulary $\mathcal{V}$, and $k = [k_1, \ldots, k_n]$ with $k_i \in \{1, \ldots, m\}$ is a vector indicating the frequency of each response in y. They propose three tasks of increasing detail and difficulty to evaluate microscopic intelligibility models: confusion frequency prediction, confusion characterization, and full confusion prediction. The first task consists in predicting the probability of confusions occurring at the resolution of individual phonemes. In the second, the target is to predict what each phoneme will be confused with, i.e., predicting a probability distribution over possible phoneme substitutions. Finally, the last task is to predict the complete word perceived by the listeners, represented as a probability distribution over responses.

In this work, we focus on the most challenging task of full confusion prediction. Our model is therefore expected to predict a probability distribution over lexical responses conditioned on the speech waveform $P(y|x)$. As proposed by Marxer et al. [1], the performance of the model is evaluated as the likelihood of the observed data, namely the (x, y, k) tuples, given the probabilities predicted by the model. The likelihood is the probability mass function of a multinomial distribution. Let i

*Equal contribution.

10.21437/Interspeech.2024-2258

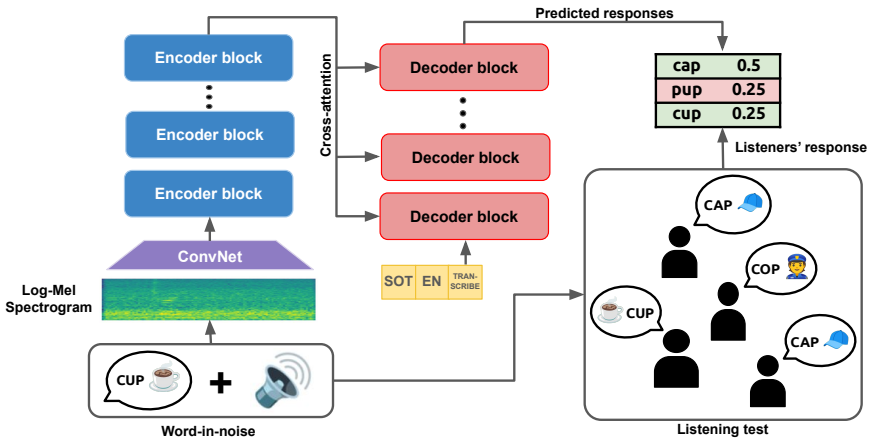


Figure 1: *Illustration of the task studied in this paper and our proposed method. We aim to predict the distribution of responses from a set of listeners to noisy speech stimuli. We use Whisper, an ASR model, to predict a distribution over posible listeners' responses.*

index the tuples in a dataset of size N, $y_{i,j}$ and $k_{i,j}$ be the j-th response and its frequency in the i-th tuple, and let us denote $P(y_{i,j}|x_i)$ as $p_{i,j}$ for brevity. To avoid numerical underflow errors we use the log-likelihood. Then, the score G that we seek to maximize is:

$$G = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{m_i!}{k_{i,1}!\dots k_{i,n_i}!} p_{i,j}^{k_{i,1}} \dots p_{i,n_i}^{k_{i,n_i}}\right) \quad (1)$$

3. Method

Our method is illustrated in Figure 1. We use Whisper [5], a recent ASR model trained on a 680,000 hours labeled speech corpus recorded in diverse conditions. Whisper outputs a probability distribution over word-level transcriptions given a speech input, therefore its application to lexical response prediction is straightforward. We chose Whisper mainly for two reasons: first, it is a state-of-the-art model whose accuracy and robustness approaches that of humans. We are interested on determining if its comparable performance to humans results in similar behavior under noisy speech stimuli. Second, its architecture is well suited for our task. Whisper consists of an encoder, which extracts contextual acoustic features c, and a decoder that predicts $P(y|c)$. Its encoder has been shown to produce useful features for macroscopic intelligibility prediction [8]. We would like to test if those representations are also useful for microscopic modeling. Having a separate decoder allows us to adapt it to our task without modifying the encoder. This is in contrast to other state-of-the-art encoder-only ASR systems in which the output is directly predicted from the encoder, e.g. [3].

3.1. Whisper architecture

Whisper is an encoder-decoder transformer model [10]. The Mel-spectrogram of the speech waveform $x \in \mathbb{R}^T$ is fed to a strided convolutional neural network, which produces a sequence of frame representations $z \in \mathbb{R}^{T' \times d}$, where d is the inner dimension of the model and $T' < T$. The encoder processes z through a stack of transformer blocks with bidirectional self-attention, yielding a sequence of contextual features $c \in \mathbb{R}^{T' \times d}$. The decoder is a stack of transformer blocks with causal self-attention that predicts the probability over transcriptions conditioned through cross-attention on c. Whisper tokenizes transcriptions using subword units [11], therefore single words can be tokenized as a sequence of multiple tokens. For instance, the word `enjoyed` can be represented as the sequence of tokens [`enjoy`, `ed`]. Let y be a single word tokenized as a sequence of tokens indexed by t, then the decoder outputs $P(y_t|y_{<t}; c)$. By the chain rule of probability $P(y|c) = \prod_t p(y_t|y_{<t}; c)$.

Whisper was trained on multiple tasks and languages, and it uses special tokens as a prefix in the decoder to inject information about the desired setup. We apply the model only for the task of ASR in English, therefore our prefix is in all cases the sequence of tokens `SOT` (start of transcription), `EN` (English), and `Transcribe` (ASR task).

3.2. Transfer learning for lexical response prediction

Whisper can be directly applied to lexical response prediction in a zero-shot fashion. Indeed, we can simply compute the likelihood of the listeners' responses predicted by the decoder. In this setup, we are essentially measuring the agreement between pre-trained Whisper and human speech perception.

We also fine-tune Whisper in two setups: to predict the top (most commonly reported) response, i.e. to match the majority of human listeners; and to model the distribution of listeners' responses. In the first case, we train the model to do ASR as in [5], using the top response as target. In the latter, we train it to minimize the absolute difference between the predicted likelihood of the observed responses and their relative frequency:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\left|\log(p_{i,j}) - \log\left(\frac{k_{i,j}}{\sum_{j=1}^{n} k_{i,j}}\right)\right| \quad (2)$$

4. Experimental Setup

4.1. Dataset

In our experiments, we used the publicly available English Consistent Confusion Corpus [12] (ECCC), a dataset created by gathering perceived responses from 15 listeners to common English words mixed with noise maskers of different types: stationary speech shaped noise (`SSN`); four-speaker speech babble (`BAB4`); and three-speaker babble modulated noise (`BMN3`)). The corpus is composed of words-in-noise misperceived in the same way by at least 6 of the 15 listeners, resulting in over 3,000

Table 1: *Performance on the* `test` *set for the different baselines (top and bottom) and transfer learning paradigms using* WHISPER-LARGE-V3 *(center).*

Model	Avg. log likelihood	Top-1 acc.	Avg. top-n coverage	Kendall corr.
Random	-160	0.00	0.00	0.00
Multinomial	-147	0.00	0.00	0.11
Rnd. init. (pred. top)	-175	0.00	0.02	0.07
Rnd. init. (pred. all)	-176	0.00	0.01	0.05
Zero-shot	-140	0.05	0.25	0.28
Fine-tuned (pred. top)	-81	0.07	0.39	0.38
Fine-tuned (pred. all)	**-46**	**0.13**	**0.48**	**0.42**
Oracle	-5	1	1	1

Table 2: *Ablation of fine-tuned Whisper's modules. Models are trained to predict the distribution of listeners' responses.*

Fine-tuned modules			Avg. log likelihood
Decoder	Encoder Transformer	Encoder ConvNet	
✗	✗	✗	-140
✓	✗	✗	-59
✓	✓	✗	-57
✓	✓	✓	**-46**

consistent misperceptions. For each sample the corpus makes available the original utterance, the noise masker, and the list of listeners' responses.

We split the dataset in `train` (80%), `dev` (10%) and `test` (10%) sets. Considering that the masker type induces different perceptual phenomena [13], and that other factors of variation such as speaker gender and identity are roughly balanced across maskers, the splits were made stratified by masker type.

4.2. Evaluation

As described in Section 2, the criterion of performance for the task of lexical response prediction is the log-likelihood of the observed responses under the predictive model averaged across the dataset (Eq. 1). However, this metric is not easily interpretable, which motivated us to also report three other metrics: the accuracy of the model to predict the most common response as the most likely (top-1 accuracy), the average percentage of the n responses present among the top-n predictions of the model (average top-n coverage), and the Kendall-Tau statistic [14], which measures the correlations between the ground truth and predicted rankings of the observed responses.

For the top-1 and top-n metrics, we compare phonemic representations, so that homophones are considered a match. We use the CMU pronunciation dictionary[1] implemented in the `nltk` library [15] to obtain phonemic representations.

For comparison, we report the same metrics for the naive baselines suggested by Marxer et al. [1]: a random model (the frequency of a response is uniform across any possible word); a multinomial model empirically fitted to the responses independent from the speech stimuli; and a topline oracle model in which the predicted probabilities perfectly match the relative frequencies of listeners' responses. To assess the effect of transfer learning, we also consider a baseline in which we trained a model from scratch (i.e. with random initialization) on our task.

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

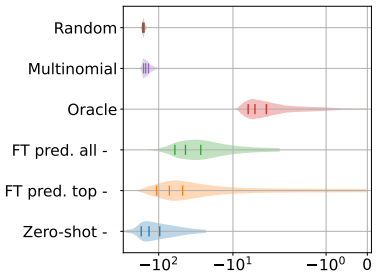


Figure 2: *Distribution of responses' log-likelihoods predicted by different models. Vertical bars show quartiles.*

Table 3: *Performance across model sizes when fine-tuning to predict the distribution of listeners' responses.*

Pre-trained model	Parameters (millions)	Avg. log likelihood
WHISPER-SMALL	244	-53
WHISPER-MEDIUM	769	-51
WHISPER-LARGE-V3	1550	**-46**

4.3. Models and Training

We used the official Whisper checkpoints available at `https://huggingface.co/openai`. For our fine-tuning experiments we used the Adam optimizer [16] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size of 16, and a learning rate schedule including a warm up phase. We did a hyperparameter search over the peak learning rate ($\{$1e-3, 1e-4, 1e-5$\}$), the fraction of training steps devoted to the warm up phase ($\{10\%, 50\%\}$), the learning rate schedule ($\{$linear, cosine$\}$) and the number of training epochs ($\{1, 4, 8, 12, 16\}$). As criterion to choose the best hyperparameters we used the log-likelihood of the data in the `dev` set. We did a full search using the WHISPER-LARGE-V3 model, from which the resulting best hyperparameters were: peak learning rate of 1e-5, warm up phase of 10%, cosine learning rate decay, and 12 epochs of training. We used this setup for all the reported experiments. After training, we keep the checkpoint with the best performance on the `dev` set.

For all experiments we used a single NVIDIA A100 GPU with 80 GB of VRAM. A training epoch for WHISPER-LARGE-V3 takes 12 minutes and requires up to 37 GB of memory.

5. Results

5.1. Effects of transfer learning

On Table 1 and Figure 2 we report the results obtained on the `test` set using the WHISPER-LARGE-V3 model. Our method outperformed all the considered baselines, even in a zero-shot setup. Models trained from scratch performed worse than the naive baselines, which shows the benefits of transfer learning on this task. Interestingly, the model fine-tuned to match the full distribution of responses (`FT pred. all`) performs better on top response prediction than the model trained specifically for it (`FT pred. top`). This shows that the extra supervision obtained from predicting the full distribution of responses

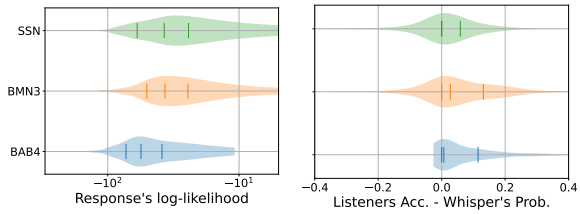


Figure 3: *Variation of performance across masker types. (left) Distribution of responses' log-likelihoods. (right) Difference between listeners' relative frequencies and predicted probabilities for the spoken word. Vertical bars show quartiles.*

leads to a better model of the majority's perception.

5.2. Ablation of fine-tuning across modules

As mentioned in Section 3, we wanted to determine the usefulness of each of the pre-trained Whisper modules for our task. In Table 2 we show the performance when keeping some of the modules frozen. We saw little performance gain from fine-tuning the transformer in the encoder in addition to the decoder. However, there is a large relative improvement of over 40% when fine-tuning the convolutional network in the encoder. This suggests that low-level acoustic features relevant for human perception are not being captured by Whisper's pre-training.

5.3. Performance scaling with model size

We also studied the scaling of performance with model size. We present the results in Table 3. Empirical results show that larger models, which exhibit increased accuracy and robustness, are also better for human lexical response prediction.

5.4. Performance by noise maskers

Motivated by the distinct effects on humans of different types of noise maskers in the ECCC [13], we studied their impact on performance. Figure 3 shows the obtained results. Our model has the most difficulty predicting human perception for speech corrupted with `BAB4` maskers. In order to assess if this mismatch was due to better or worse robustness relative to humans, we also evaluated the difference in accuracies between our model and humans, i.e. the difference between the probability of the correct response estimated by the listeners and that predicted by our model. For `BAB4`, the model was consistently worse than humans in perceiving the original spoken word (the distribution of differences is largely positive for `BAB4`). For the other masker types, in the majority of cases, humans also appear better than Whisper at perceiving the right word.

6. Related work

Previous works have used transfer learning from Whisper for macroscopic intelligibility prediction [17, 18, 8]. Mogridge et al. [17] showed that Whisper's decoder representations outperformed those from the encoder, indicating that high level linguistic patterns heavily impact macroscopic intelligibility.

As for microscopic intelligibility, Cuervo et al. [19] approached the task of phone-level prediction of misperception probability using transfer learning from wav2vec 2.0 [3], a large scale self-supervised trained deep learning model. To the best of our knowledge, we are the first to use transfer learning from a large scale deep learning model for the micro-intelligibility task of lexical responses characterization.

7. Discussion

7.1. On the implications of our results

Our method significantly outperformed the baselines in all the considered metrics. This shows that Whisper does capture useful information for microscopic intelligibility prediction at the lexical level. Indeed, zero-shot experiments showed some degree of analogous behaviour between humans' and Whisper's perceptions of speech-in-noise (25% of agreement on average). Moreover, we saw predictive performance increase with the size of the pre-trained Whisper model, similarly as to how accuracy and robustness scale [5]. Considering that the performance of deep learning models often scales predictably with the amount of compute (model size and amount of data) used for training [20], perhaps we can expect larger models, which are expected to be more accurate and robust, to increasingly converge with humans on speech perception.

Fine-tuning to match listeners' responses on the ECCC helped to significantly increase performance. Interestingly, we observed that fine-tuning to predict the full distribution of listeners' responses results in overall better performance than training to predict the mode of the distribution, even if one is only interested on predicting the mode. This shows that modeling fine-grained phenomena can bring benefits on coarser tasks, which cannot be obtained from training on the coarse task alone. Perhaps similar gains can be obtained on macroscopic intelligibility prediction by training on microscopic intelligibility prediction.

We obtained the most performance gains from fine-tuning the convolutional encoder, which suggests that the largest difference between Whisper's and human speech processing is at the low acoustic level. It could be interesting to see if a convolutional encoder fine-tuned for human speech perception would yield increased accuracy and/or robustness on ASR.

7.2. On limitations

We are limited mainly by the amount of data available for microscopic intelligibility prediction. We believe that part of the performance gap between our method and the topline can be due to the topline being a noisy, biased estimate of the real distribution of lexical responses that our model (by virtue of regularization) approximates. A larger sample of stimuli and listeners' responses is likely to meaningfully increase performance.

In terms of assessing the quality of our model, we are also limited by the lack of strong baselines that employ state-of-the-art methods for speech processing, against which we can compare our results. We hope this work will motivate other researchers to improve upon our work within a standardized evaluation framework, which will lead to the construction of a proper benchmark for microscopic intelligibility models.

8. Conclusions

We applied transfer learning methods from Whisper, a state-of-the-art automatic speech recognition system, for the task of predicting human lexical responses to speech-in-noise stimuli. We evaluated our model within a standardized framework in a zero-shot setup and when fine-tuned specifically for the task. Additionally, we studied the effects on performance of the granularity of fine-tuning, model size, fine-tuning of each pre-trained module, and different noise maskers. Results are promising, with our methods significantly outperforming naive baselines.

9. Acknowledgements

We are grateful to the French National Research Agency for their support through the ANR-20-CE23-0012-01 (MIM) grant, and the Institute of Convergence ILCB, supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

10. References

[1] R. Marxer, M. Cooke, and J. Barker, "A framework for the evaluation of microscopic intelligibility models," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[2] M. I. Mandel, "Learning an intelligibility map of individual utterances," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[3] 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*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[6] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *Proc. Interspeech*, 2023, pp. 2798–2802.

[7] 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.

[8] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," 2024.

[9] 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

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[12] 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.

[13] 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.

[14] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[15] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69–72.

[16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[17] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," 2024.

[18] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Utilizing Whisper to enhance multi-branched speech intelligibility prediction model for hearing aids," 2023.

[19] S. Cuervo and R. Marxer, "On the Benefits of Self-supervised Learned Speech Representations for Predicting Human Phonetic Misperceptions," in *Proc. Interspeech*, 2023, pp. 1788–1792.

[20] J. Hestness, S. Narang, N. Ardalani, G. F. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *CoRR*, vol. abs/1712.00409, 2017. [Online]. Available: http://arxiv.org/abs/1712.00409