Phoneme-Level Speech Intelligibility Reduction

Aine Drelingyte*, Romain Serizel*, Mathieu Lagrange‡
*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

‡Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Abstract—**As noise masking is central to reducing fatigue in open-plan offices, identifying which elements of speech contribute to intelligibility is essential. This paper employs phoneme-level noise masking techniques to examine how the intelligibility of consonants and vowels degrades under noisy conditions, and to determine which aspects of the speech signal are truly compromised. We introduce and compare two methods: one that applies noise between all phoneme boundaries and another that respectively applies noise to consonants and vowels, based on phoneme-specific signal-to-noise ratios. Evaluations on Harvard sentence corpora, using both automatic speech recognition systems and objective intelligibility metrics, reveal that although aggregated word error rates indicate significantly higher degradation for consonants compared to vowels, the direct phoneme error rate analysis does not reflect this disparity. This suggests that the marked decline in word-level intelligibility may not be solely due to differences in phone class, but also to other factors such as ASR contextual compensation mechanisms.**

I. Introduction

Noise pollution greatly undermines our quality of life. In office environments, it not only triggers stress that hampers performance but also breeds annoyance and disrupts social interactions [1]. Background speech, in particular, can be extremely distracting, reducing overall productivity [2]. While active noise cancellation methods are effective at diminishing steady, predictable sounds, they often falter when faced with the fluctuating frequencies of human speech [3]. As a result, noise generation machines have emerged as an alternative; these devices mask speech intelligibility by elevating the overall noise floor. However, they often result in excessively high sound levels due to neglecting the varying contributions of individual phonemes [4].

Typically, noise maskers are set to achieve a $0\,\mathrm{dB}$ signal to noise ratio (SNR). In practice, this often corresponds to a noise level around $45\,\mathrm{dB(A)}$, which simulates interfering speakers positioned a few meters away. This level is chosen as a compromise between effectively masking speech and maintaining a comfortable ambient sound [5]. However, this approach has notable drawbacks: it elevates overall noise levels, potentially contributing to long-term stress, annoyance, and cognitive fatigue [6], [7].

Additionally, research in speech perception further reveals that although intelligibility begins to decline significantly as the SNR lowers, masking is not significant until about $-12\,\mathrm{dB}$, as the residual speech cues remain detectable [8]. Supporting this observation, the study by Henriques et al. [9] on phrase recognition thresholds in noise (PRTN) demonstrated that normal-hearing individuals typically recognize approximately 50% of speech stimuli at an average SNR of about $-8\,\mathrm{dB}$, with individual thresholds ranging from $-4\,\mathrm{dB}$ to $-13\,\mathrm{dB}$. Similar investigations have reported PRTN values between $-2\,\mathrm{dB}$ and $-12\,\mathrm{dB}$, further indicating that even under adverse conditions, some speech cues remain audible [10]. Consequently, noise generation machines operating at a $0\,\mathrm{dB}$ SNR may not fully conceal speech, as they leave enough residual cues for partial intelligibility under focused listening. This residual intelligibility may contribute to increased cognitive load and fatigue, consistent with findings that listening in noisy environments—such as the typical SNRs found in school classrooms—imposes considerable cognitive demands [11].

At a more granular level, research indicates that different phonemes exhibit varying susceptibility to noise masking. For example, consonants are generally more vulnerable to noise interference than vowels, primarily due to their lower energy and shorter duration [12]. Moreover, the nature of the noise (such as steady-state versus fluctuating noise) and its spectral characteristics can differentially affect various speech components [13]. Building on this, we believe that this phoneme-specific variability offers promising avenues for developing more sophisticated targeted speech masking strategies.

In light of these findings, this paper leverages phoneme-level noise masking techniques to investigate how noise degrades phoneme types differently, and to identify which aspects of the speech signal are truly compromised under noisy conditions. Our objective is to understand whether the observed phoneme sensitivity to noise is solely due to differences between phoneme classes or if the ability of contextual compensation mechanisms—plays a significant role.

In our evaluation, automatic speech recognition (ASR) systems served as a proxy for speech intelligibility, supported by prior work showing strong correlations with human listening tests [14], [15]. To complement this, we used objective metrics targeting different intelligibility aspects: Short-Time Objective Intelligibility (STOI) [16], which measures temporal and spectral similarity between clean and degraded speech; Speech Distortion Ratio (SDR); and the Hearing-Aid Speech Perception Index (HASPI) [17], which models auditory processing to assess noise impact on quality.

This study was conducted as part of Reducing Noise with Augmented Reality – ReNAR project, funded by ANR.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria, and including CNRS, RENATER, and several universities as well as other organizations (see https://www.grid5000.fr).

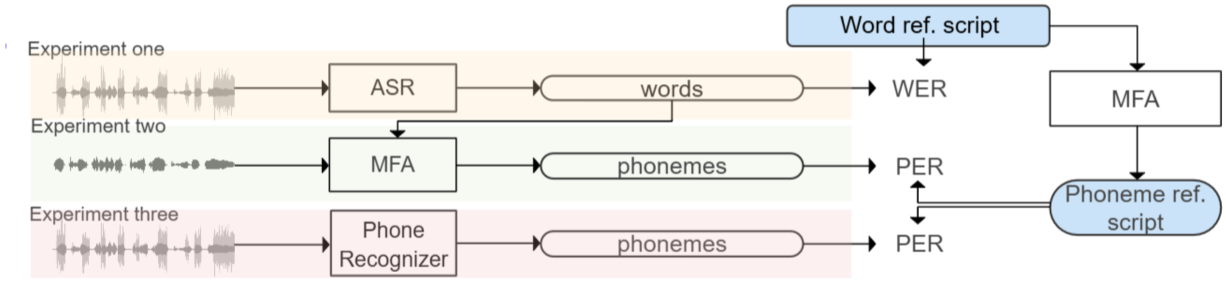

Fig. 1: Three experiments: masked speech is processed by (1) an ASR system for WER, (2) the ASR output aligned with clean speech via MFA for PER, and (3) a phone recognizer for direct PER computation.

II. Methodology

Our study is structured into three blocks displayed in Figure 1, each contributing a perspective on how noise application on phoneme level affects speech intelligibility and recognition.

The first part of the study involves measuring the Word Error Rate (WER) of sentences that have been subjected to various noise conditions. This measurement provides a broad, macro-level assessment of how added noise decreases overall intelligibility by considering whole-word accuracy; even a single misrecognized phoneme can result in an error for an entire word. By analyzing WER, we explore the hypothesis that certain phonemes may contribute more significantly to the perceptual quality of speech, and their masking might lead to greater declines in recognition performance.

The second part uses the Allosaurus phone recognizer [18], which has no language model and relies solely on acoustic cues, to measure phoneme error rates (PER). This allows us to examine how noise affects individual sounds and whether consonants or vowels are more impacted. Since earlier research suggests consonants are more vulnerable [12], we test if they show higher error rates under noisy conditions compared to vowels or complete phonemic set. Unlike word error rate (WER), which may aggregate small phoneme errors at a word level, PER gives a clearer view of how noise disrupts speech at the sound level, helping to identify which phonemes are inherently more susceptible to noise and if they contribute to the overall degradation of speech intelligibility disproportionately.

The third part of the study compares phoneme error rates (PER) from two systems: a basic phone recognizer and an Automatic Speech Recognition (ASR) system. For the ASR system, transcripts are aligned with clean speech using the Montreal Forced Aligner (MFA) to get the phoneme sequences. This comparison is important because it shows how phoneme-level errors differ between a system that only recognizes sounds and one that also uses a language model. It helps to reveal whether language modeling compensates for or amplifies the impact of noise.

III. Experimental Design

We conducted our experiments using two English corpora. The first corpus consists of Harvard sentences spoken by a female native British English speaker and sampled at 48kHz [19]. It is considered for its phonetic balance and widespread use in speech intelligibility research. The second corpus is LibriSpeech, a large-scale audiobook corpus commonly employed in training ASR systems, sampled at 16kHz. Each utterance is masked using speech-shaped noise (SSN) across a spectrum of signal-to-noise ratios ranging from highly adverse to nearly noise-free environments.

Our approach incorporates two phoneme-aware masking strategies. In the first, Phoneme-Based Masking, the speech signal is segmented into individual phonemes, and SNR is calculated and applied between their boundaries obtained with MFA. This method differentiates between vowels and consonants by calculating and applying SNR adjustments separately for each category with no noise on the other. In the second approach, Level-Based Masking, the SNR for each phoneme is computed and applied between phoneme boundaries on all phonemes. This yields a tailored noise profile that adapts to the varying loudness of different phonemes.

To assess the impact of masking methods on speech intelligibility, we also use Uniform Masking, a baseline method, to evaluate ASR performance under standard conditions. In total, we assess the performance of 12 ASR systems. Eight of these systems are known to have been trained on LibriSpeech, while the remaining four were trained on alternative datasets. We consider those two classes separately to offer an out-of-domain perspective and explore biases potentially introduced toward the corpus content. WER was calculated for each SNR category (ranging from $-40\,\mathrm{dB}$ to $40\,\mathrm{dB}$ in $10\mathrm{dB}$ increments).

This experimental design allows us to examine the influence of dataset familiarity on WER-based evaluations. Furthermore, it provides insights into the differential sensitivity of vowels and consonants to noise masking and their respective roles in speech intelligibility.

IV. Experimental Results and Discussion

A. Bias

We first study the ASR behavior using different corpora. Table I presents average WER values calculated on both LibriSpeech and Harvard corpora at an SNR of $0\mathrm{dB}$ using SSN as the noise masker. Models trained on LibriSpeech consistently demonstrate superior performance on LibriSpeech

Corpus	Not Trained on Librispeech	Trained on Librispeech
Harvard	0.789	0.649
Librispeech	0.767	0.460

TABLE I: Average WER calculated on Librispeech and Harvard corpora for Models trained and not trained on Librispeech.

corpus, achieving a substantially lower WER of 0.460, compared to 0.767 for models not trained on this corpus. On the Harvard corpus, the difference is narrower with LibriSpeech-trained models attaining a WER of 0.649 versus 0.789 for models not trained on Librispeech. This performance gap highlights a clear bias favoring familiar data which probably does not reflect a human behaviour. This suggests that in-domain evaluations might overestimate a model's true capability. Thus, these findings emphasize the necessity of incorporating evaluations on out-of-domain datasets to provide a more comprehensive assessment of masking strategy effectiveness. Henceforth, to avoid the bias introduced by Librispeech trained models on Librispeech corpus, we will continue our evaluation by considering the Harvard corpus.

B. Keyword Reception Threshold vs. Human Reception Thresholds

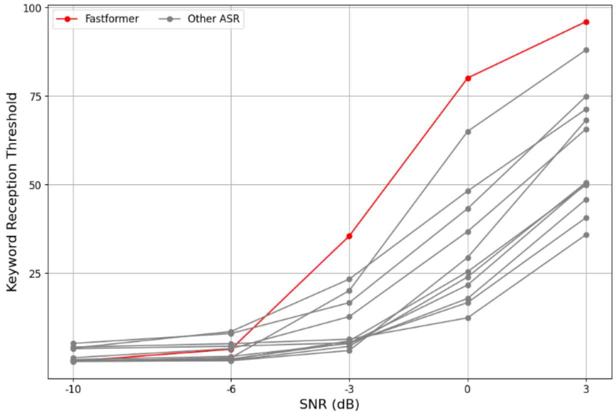

Fig. 2: ASR-based Keyword Reception Threshold (KRT), calculated by measuring the percentage of correct content words in the transcripts—excluding pronouns, prepositions, and articles—using a bag-of-words approach.

The following will compare the human Speech Reception Threshold (SRT) graph presented by Aubanel et al., 2020 [20], with ASR performance in order to identify the ASR system that best matches human speech intelligibility under noise conditions. To facilitate this comparison, we use KRT as an ASR-based equivalent to the SRT, providing a more direct alignment with human intelligibility benchmarks.

Although ASR performance broadly follows similar trends to human intelligibility on the Harvard corpus, there remains a difference between SRT and ASR outcomes. Notably, even the closest ASR system—**Fastformer**—lags behind human performance. Figure 2 illustrates this discrepancy. For example, while humans transition gradually from near-chance to near-perfect understanding—reaching approximately 75% recognition around $-3\,\mathrm{dB}$ —Fastformer maintains KRT near 40%, reaching the same 75% recognition only around $0\,\mathrm{dB}$. These differences emphasize that although ASR-based measures capture general trends, they may not fully reflect the true perceptual impact of noise. Nevertheless, staying mindful of the consistent $3\,\mathrm{dB}$ shortfall relative to human performance, we will rely on Fastformer as our ASR baseline for the upcoming experiments, complemented by additional metrics (STOI, HASPI, SDR) to ensure a thorough intelligibility evaluation.

C. Metrics Comparison

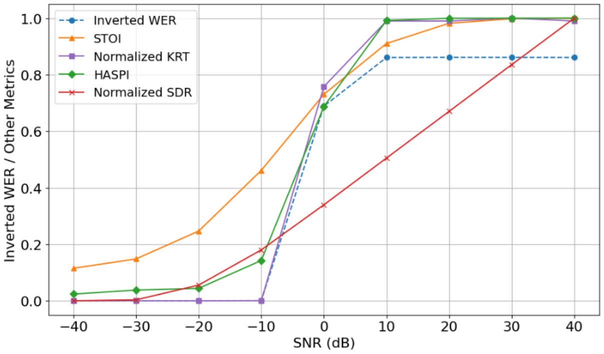

Fig. 3: Comparison of average inverted WER across SNR for uniform masking approach, plotted alongside HASPI, STOI and SDR.

Figure 3 compares different intelligibility metrics, including WER, across varying noise conditions. HASPI and WER show the greatest intelligibility loss at extremely negative SNRs, reflecting the impact of severe noise. Notably, HASPI aligns more closely with WER trends than STOI, likely because it accounts for psychoacoustic processes that mirror actual listening conditions, thereby providing a more faithful intelligibility estimate than STOI [1] in moderate SNR environments. It is important to note that here the horizontal axis is showing the masker gain which is the overall SNR level to which the signal was adjusted. Additionally, in this and the following experiments, Word Error Rate (WER) has been inverted—referred to as inverted WER (iWER)—for clearer comparison, where a value of 1 indicates higher intelligibility and lower values reflect poorer recognition performance.

D. Masking Strategy Impact

Figure 4 illustrates the impact of different masking strategies on speech intelligibility across a range of SNRs.

[1] It is important to note that while the speech was sampled at 48kHz, STOI only reflects intelligibility up to 5kHz. As such, any masking effects in higher frequencies are not captured by this metric, which may limit the interpretation of results involving high-frequency noise components.

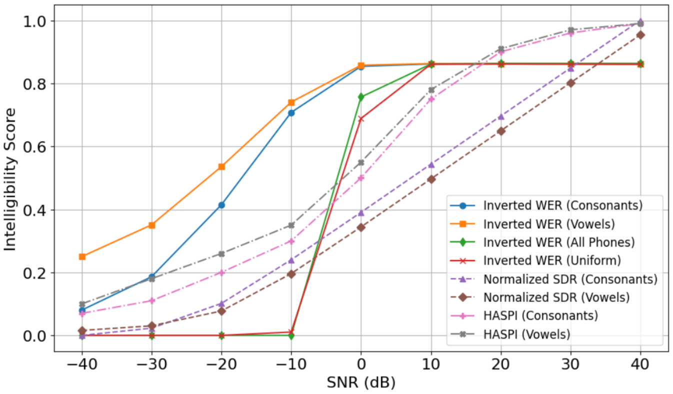

Fig. 4: iWER of level-based, uniform masking approaches, and phone-based with noise applied only on consonants or vowels under varying SNRs.

To begin, both level-based[2] and uniform masking yield low iWERs at extremely negative SNRs ($-40\,\mathrm{dB}$ to $-20\,\mathrm{dB}$), but they diverge significantly from about $0\,\mathrm{dB}$ onward. The level-based approach shows a faster increase in iWER as the noise conditions improve, raising above the uniform method throughout the mid-range SNRs. Thus, indicating a better intelligibility reduction using the uniform approach.

Within phone-based masking, where noise was applied selectively to either consonants or vowels, both phoneme classes exhibit low iWER in extreme noise conditions ($-40\,\mathrm{dB}$ to $-20\,\mathrm{dB}$), with consonants generally exhibiting lower error rates—likely due to their shorter duration and higher-frequency bursts. As the SNR improves beyond $-10\,\mathrm{dB}$, both curves rapidly increase and effectively converge, indicating minimal performance difference at moderate to high SNRs.

To evaluate whether the differences observed under extreme noise conditions (from $-40\,\mathrm{dB}$ to $0\,\mathrm{dB}$) are statistically significant, we performed a *paired t-test* comparing iWERs for consonants and vowels. The t-test yielded a t-statistic of 2.8854 with a p-value of 0.0432, which is below the conventional significance threshold of $p < 0.05$. This indicates that the iWER for consonants is significantly lower than for vowels, implying that consonants have higher error rates and are more susceptible to noise degradation in extreme conditions.

E. Consonant and Vowels, CER and VER

In order to assess the impact of phoneme-level masking on WER, we computed PERs separately for consonants (CER) and vowels (VER) on the Harvard corpus. Specifically, we compared only the insertions, deletions, and substitutions relevant to each phoneme class, using a phone recognizer's output against the clean transcripts. As shown in Figure 5, consonants consistently exhibit slightly higher error rates—likely due to their shorter durations and higher-frequency components—while vowels recover more robustly at higher SNRs, reflecting their stronger formant energy.

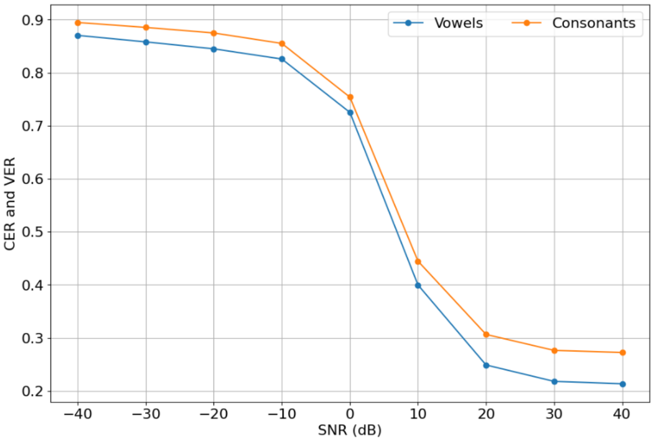

Fig. 5: Comparison of CER and VER under uniform masking across a range of SNRs.

However, the difference between CER and VER was not statistically significant. This suggests that the observed degradation in overall intelligibility is not solely attributable to phonetic class differences but may also be influenced by linguistic context or other factors. Furthermore, our findings align with the HASPI and SDR data in Figure 4, reinforcing that vowel–consonant differences are negligible. This suggests that using other metrics over PER might have been sufficient.

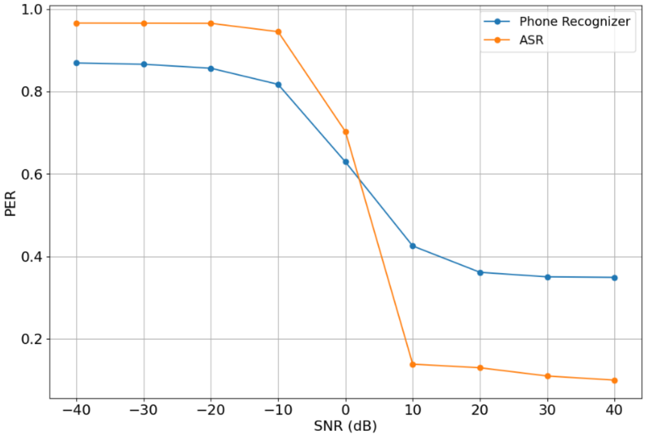

Fig. 6: Comparison of PER between a phone recognizer and an ASR system across different SNRs.

F. Linguistic Context Impact on Phone Recognition

In this subsection we will examine PERs obtained from ASR transcripts after aligning them with clean audio references using MFA. This allows us to determine whether the ASR system mitigates some phoneme-level errors through contextual information. From Figure 6, the phone recognizer consistently shows higher error rates thorough the high-range SNRs than the ASR system. This does suggest that the ASR's additional modeling— most likely its language model—helps it recover from or "fill in" some phoneme-level errors that

[2] The SNR for each phoneme is computed and applied between phoneme boundaries on all phonemes, refer to Section III

the phone recognizer cannot. However, it does not strictly prove that "context" is the sole factor. ASR systems differ in acoustic modeling and training objectives as well, so their better performance at moderate and high SNR could reflect both acoustic and linguistic advantages.

V. Discussion

This study investigated phoneme-level noise masking strategies aimed at reducing speech intelligibility, with a focus on understanding how errors accumulate at the word level and whether these errors differ between consonants and vowels. Our experiments revealed that, under extreme noise conditions, the WER was significantly higher for consonants than for vowels. However, when directly examining phoneme error rates (PER) obtained with a phoneme recognizer, we observed no significant difference between the two phoneme classes. This discrepancy suggests that, although consonants and vowels exhibit similar error rates at the phoneme level, the aggregation of phoneme errors at the word level (captured by WER) disproportionately impacts consonants, creating an impression of greater consonant vulnerability.

Furthermore, the analysis of PER obtained from ASR systems suggests that ASRs may be more effective at compensating for vowel errors compared to consonant errors. Vowels, characterized by their robust acoustic structure and prominent formant energy, likely provide sufficient contextual cues for the ASR's language model to recover or "fill in" missing information, thereby mitigating their impact on overall word recognition. In contrast, consonants, which are transient and acoustically less redundant, become more vulnerable to noise-induced degradation, resulting in higher aggregated errors at the word level. Hence, the interpretation of these findings is that overall intelligibility degradation (as captured by WER) may not be determined solely by phoneme classes.

These results underscore the complexity of phoneme-aware noise masking: although phoneme-level measures like PER may not capture significant differences between consonants and vowels, the word-level performance clearly reflects the cumulative detrimental effect of consonant degradation under extreme noise. Consequently, our findings suggest that improvements in noise masking strategies should consider not only isolated phoneme errors but also their combined effect on overall speech intelligibility.

VI. Conclusion

By employing phoneme-level masking techniques and analyzing both word and phoneme error rates, this study demonstrates that while consonants show greater degradation at the word level, the differences at the phoneme level are not as pronounced.

We believe these insights pave the way for improved noise masking strategies by revealing that speech intelligibility degradation is not solely driven by phoneme classes.

References

[1] L. Brocolini, E. Parizet, and P. Chevret, "Effect of masking noise on cognitive performance and annoyance in open plan offices," *Applied Acoustics*, vol. 114, Jul. 2016.

[2] M. Pierrette, E. Parizet, P. Chevret, and J. Chatillon, "Noise effect on comfort in open-space offices: Development of an assessment questionnaire," *Ergonomics*, vol. 58, no. 1, 2014.

[3] D. Bhatia, F. V. Francis, P. Panging, Hitanshu, and B. Khyllait, "Revolutionizing noise management: Active noise cancellation headphones in healthcare and beyond," *Journal of Biomedical Science and Engineering*, 2024.

[4] L. Lenne, P. Chevret, and J. Marchand, "Long-term effects of the use of a sound masking system in open-plan offices: A field study," *Applied Acoustics*, vol. 159, 2019.

[5] L. Lenne, P. Chevret, and J. Marchand, "Long-term effects of the use of a sound masking system in open-plan offices: A field study," *Applied Acoustics*, vol. 158, 2020.

[6] H. Shu, Y. Song, and H. Zhou, "Assessment of music and water sounds for urban noise masking," in *IEEE TENCON 2018*, 2018.

[7] J. Cai, J. Liu, N. Yu, and B. Liu, "Effect of water sound masking on perception of the industrial noise," *Applied Acoustics*, 2019.

[8] G. A. Miller and P. S. Nicely, "An analysis of perceptual confusions among some english consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, 1955.

[9] M. O. Henriques, E. C. Miranda, and M. J. Costa, "Speech recognition thresholds in noisy areas: Reference values for normal hearing adults," *Brazilian Journal of Otorhinolaryngology*, vol. 74, no. 2, 2008.

[10] S. Anderson, T. White-Schwoch, A. Parbery-Clark, and N. Kraus, "A dynamic auditory-cognitive system supports speech-in-noise perception in older adults," *Hearing Research*, vol. 300, 2013.

[11] C. S. Howard, K. J. Munro, and C. J. Plack, "Listening effort at signal-to-noise ratios that are typical of the school classroom," *International Journal of Audiology*, 2011.

[12] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.

[13] M. A. Stone, C. Füllgrabe, and B. C. Moore, "Notionally steady background noise acts primarily as a modulation masker of speech," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 317–326, 2012.

[14] S.-E. Kim, B. Chernyak, O. Seleznova, J. Keshet, M. Goldrick, and A. Bradlow, "Automatic recognition of second language speech-in-noise," *JASA express letters*, vol. 4, Feb. 2024.

[15] A. Author, B. Author, and C. Author, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the ICML*, PMLR, 2023.

[16] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, Oct. 2011.

[17] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Commun.*, vol. 65, 2014.

[18] X. Li, S. Dalmia, J. Li, *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP*, IEEE, 2020.

[19] P. Demonte, *Harvard speech corpus - audio recording 2019*, University of Salford. Collection, 2019.

[20] V. Aubanel, C. Bayard, A. Strauss, and J.-L. Schwartz, "The fharvard corpus: A phonemically-balanced french sentence resource for audiology and intelligibility research," *Speech Communication*, vol. 124, 2020.