



Towards Adaptable and Intelligible Speech Synthesis in Noisy Environments

Lubos Marcinek, Jonas Beskow, Joakim Gustafson

¹Department of Speech, Music and Hearing, KTH, Sweden

lubosm@kth.se, beskow@kth.se, jocke@speech.kth.se

Abstract

We present an investigation into adaptable speech synthesis for noisy environments. Leveraging a zero-shot TTS we synthesized a corpus of 1,200 speech samples from 100 sentences of varying complexity, each generated at six distinct levels of vocal effort. To simulate realistic listening conditions, the synthesized speech is merged with environmental noise recordings from a diverse range of indoor and transportation settings at nine different signal-to-noise ratios. We assess the intelligibility of the resulting noisy speech using the ASR word error rates across conditions. Additionally, the input text was evaluated using four metrics on sentence complexity and word predictability. A number of regression models that used noise type, SNR, vocal effort and text as input were trained to predict ASR WER. Results show that increased vocal effort improves intelligibility, with benefits up to 30% in adverse conditions, most most pronounced in environments with competing speech at low SNRs.

Index Terms: speech synthesis, speech intelligibility, speech adaptation, noisy environments

1. Introduction

Speech synthesis technology has become an essential component in modern communication systems, transforming how we interact with devices and enabling more accessible communication across diverse applications [1, 2]. While significant advances have been made in Text-to-Speech (TTS) systems, maintaining intelligibility in noisy environments remains a fundamental challenge that affects both human listeners and automated systems [3, 4]. Environmental noise significantly impacts synthetic speech intelligibility, particularly in real-world applications such as public announcements, virtual assistants, and assistive technologies [5]. This challenge is especially pronounced in environments with varying noise levels, where the mutual information between the intended message and the received signal is compromised, leading to degraded intelligibility and reduced communication effectiveness [6, 7]. Current speech synthesis systems face several intrinsic limitations. Traditional systems, designed primarily for controlled, quiet environments, often struggle to maintain intelligibility in the presence of background noise [8, 9]. Moreover, their lack of adaptability to dynamic acoustic conditions results in suboptimal performance in real-world scenarios. This limitation is particularly critical for individuals with hearing impairments, who may already experience difficulties with high-frequency sounds and speech perception [10]. Various strategies have been developed to enhance speech intelligibility in noisy conditions. These include time-frequency masking techniques, adaptive speech synthesis systems that modulate speech style based on environmental conditions, and methods for enhancing excitation and for-

mant prominence [11, 12, 13]. These approaches aim to mimic natural human adaptations to noisy environments, such as the Lombard effect [14], where speakers instinctively modify their vocal effort to maintain intelligibility. However, a critical gap in the literature is the lack of comprehensive studies comparing how these adaptations affect perception of speech in different noise environments. The complexity of creating effective speech synthesis systems extends beyond noise handling to include challenges in achieving naturalness, supporting multiple languages, and expressing emotions in synthesized speech [15, 16, 17]. Recent advancements, such as the use of real-world data from platforms like YouTube and podcasts, have improved the naturalness and intelligibility of synthesized speech [18], yet significant challenges remain in creating truly adaptive and robust systems for diverse acoustic environments. Recent research has explored various strategies to enhance speech intelligibility under adverse acoustic conditions. This includes investigating linguistic modifications like paraphrasing to introduce noise-robust acoustic cues that improve human speech perception [19]. Our study focuses on adaptive vocal efforts in TTS as a direct means to improve intelligibility, with the novel contribution of comparing ASR and human perception of the same stimuli. We use a zero-shot TTS system named Llasa, which is built upon a Llama-based architecture [20]. It leverages audio prompts to generate intelligible speech with natural prosodic variations. Llasa's training incorporates an ASR system using Word Error Rate as a key metric to ensure intelligibility standards. By continuously monitoring WER, the training process ensures that the synthesized speech not only captures the intended linguistic content but also meets stringent intelligibility standards. Our work addresses these research questions:

- how do vocal effort variations affect speech intelligibility across different environmental noise types and SNR levels
- are machine learning models able to effectively predict intelligibility based on acoustic and linguistic features, and if so which factors are most important.

To address these questions, we synthesized 100 utterances of varying textual complexity using audio prompts that differed in vocal effort levels. We then systematically applied different noise types across a range of signal-to-noise ratios (SNR), generating a comprehensive dataset of 88,020 synthesized speech samples. We evaluated all 88,020 samples through automatic speech recognition (ASR) systems under diverse noise conditions to quantify the relationship between vocal effort variations and intelligibility improvements, as measured by ASR performance. This methodology provides a complementary perspective to previously proposed linguistic strategies while establishing a practical framework for developing robust speech synthesis systems optimized for real-world noisy environments.

2. Methodology

2.1. Corpus Creation

Our corpus creation process was informed by established noisy speech databases designed for training speech enhancement algorithms and TTS models [21]. We selected 100 sentences with varying lexical and syntactic complexity to ensure linguistic representation across different grammatical structures and vocabulary levels. Using the zero-shot TTS system LLasa, 1,200 speech samples were synthesized by incorporating six distinct vocal effort levels to simulate comprehensive speaking intensities. Audio prompts combined with input text were leveraged by LLasa to control speaker identity and speaking style characteristics. To maintain consistent speaker identity while systematically varying vocal effort, speech samples were selected from two iconic movie scenes characterized by progressive vocal intensity increases: Reese Witherspoon’s first trial scene in “Legally Blonde” and Samuel L. Jackson’s Ezekiel 25:17 recitation in “Pulp Fiction.” Two audio samples at three vocal effort levels were extracted from each speaker, with durations of 4.5 to 9 seconds. Six hundred speech samples per gender (100 sentences \times 6 vocal effort levels) were yielded, totaling 1,200 synthesized samples. A robust foundation for evaluating speech intelligibility across diverse acoustic conditions was provided by maintaining linguistic consistency while systematically varying acoustic properties. Acoustic features across all 1,200 clean synthesized speech samples were measured using the OpenSmile toolkit [22] with the eGeMAPSv02 feature set. Systematic acoustic changes consistent with natural vocal effort adaptation were produced by the TTS system. Fundamental frequency was increased substantially with higher vocal effort: female speakers showed increases from 32.2 Hz to 38.5 Hz and male speakers demonstrated larger changes from 19.3 Hz to 40.8 Hz. Intensity was increased progressively for both genders (females: 0.46 to 0.74; males: 0.24 to 0.75). The alpha ratio was increased from 9.6 dB to 5.3 dB for females and 15.2 dB to 4.0 dB for males, indicating enhanced high-frequency energy important for speech clarity in noise. Voice quality patterns differed between genders: female speakers showed maximum harmonic-to-noise ratio at intermediate effort levels (6.4 dB, 7.2 dB, 6.3 dB), while male speakers showed continuous increase from 1.1 dB to 7.5 dB. The Hammarberg index was decreased for both genders from 19.3 dB to 14.9 dB for females and 24.8 dB to 12.9 dB for males. Spectral characteristics showed increased high-frequency content with higher effort levels: female speakers demonstrated spectral slope increases from 0.04 to 0.06 and male speakers showed changes from near-zero to 0.05. These measurements confirm that speech with authentic vocal effort characteristics was generated by the TTS system. Baseline intelligibility measures were established by transcribing all 1,200 synthesized speech samples using the Whisper 3 Turbo automatic speech recognition (ASR) system [23]. Transcription accuracy was evaluated using Word Error Rate (WER), where perfect intelligibility is indicated by WER of 0. Only the 1,035 synthesized speech samples that achieved perfect transcription accuracy (WER = 0) were selected for subsequent processing, ensuring that intelligibility degradation could be attributed to noise conditions rather than synthesis artifacts. These high-intelligibility samples were then systematically mixed with various noise types to assess the impact of different acoustic conditions on speech perception.

2.2. Noise Integration and SNR Control

To simulate realistic listening conditions, synthesized speech was combined with environmental noise recordings from the DEMAND corpus [24], representing diverse acoustic environments across domestic, office, public, and transportation settings. These noise categories include: Washing, Kitchen, Living, Hallway, Meeting, Office, Station, Cafeteria, Restaurant, Metro, Bus, and Car environments. To ensure precise and reproducible SNR measurements while preserving the natural acoustic characteristics of different vocal effort levels, we used a methodology that preserves the original energy level of each synthesized utterance. Critically, we did not use energy normalization, as this would artificially eliminate the natural energy differences that are fundamental to vocal effort’s intelligibility benefits. For each speech sample, we calculated the root mean square (RMS) energy without normalization, then randomly selected noise segments matching the utterance duration and calculated their RMS energy. The target SNR was achieved by scaling only the noise amplitude to match the desired signal-to-noise ratio, creating the final mixed signal by combining the original clean speech with the scaled noise, with dynamic range compression applied when necessary to prevent digital clipping while maintaining SNR targets. This approach is methodologically superior compared to energy normalization techniques because it reflects real-world scenarios where increased vocal effort naturally produces higher acoustic energy, ensuring fair comparison across vocal effort levels without artificially removing the very acoustic advantages we aim to study. Each synthesized speech sample was mixed with these twelve background noise types at nine signal-to-noise ratios (SNRs): -24, -18, -12, -6, 0, 6, 12, 18, and 24 dB, covering a wide range of acoustic challenges from extremely adverse to highly favorable listening conditions, generating a total of 88,020 speech-in-noise files.

2.3. Intelligibility Evaluation and Text Feature Extraction

To evaluate the intelligibility of synthesized speech under various noise conditions, we employed the wav2vec 2.0 automatic speech recognition (ASR) system, a self-supervised model pre-trained on large-scale speech data that learns latent speech representations directly from raw waveforms, enhancing robustness against noise distortions. Each of the 88,020 speech samples was processed using wav2vec 2.0 [25] to generate automatic transcriptions, with intelligibility assessed using Word Error Rate (WER), where lower WER indicates higher intelligibility and WER = 0 denotes perfect transcription. The results enable an in-depth analysis of how noise type, vocal effort, and SNR influence ASR performance.

Text was evaluated using four key metrics: Flesch-Kincaid Grade Level (FK) normalized from its 0-20 scale, Word Complexity (WC) measuring the percentage of complex words, Syntactic Complexity (SC) assessing sentence structure, and Predictability (P) evaluating pattern repetition and word frequency. Higher weights (30%) were assigned to Flesch-Kincaid and Word Complexity metrics as they represent fundamental aspects of text accessibility, while lower weights (20%) were given to Syntactic Complexity and Predictability. The final Overall Textual Feature Score (OTFS) is calculated as:

$$OTFS = 0.3 \cdot \min(100, \max(0, (20 - FK) \cdot 5)) + 0.3 \cdot (100 - WC) + 0.2 \cdot (100 - SC) + 0.2 \cdot P$$

The final score ranges from 0 to 100, with higher scores indicating better intelligibility, is a balanced assessment of intelligibility as it combines validated readability metrics with structural and pattern-based features of the text.

3. Results and Analysis

3.1. WER: Intelligibility Scores

The experimental results in Figure 1 show the relationship between Word Error Rate (WER) and Signal-to-Noise Ratio (SNR) across twelve noise environments with three speaking effort levels. Analysis covered indoor settings (kitchen, living room, office, hallway), social spaces (meeting room, cafeteria, restaurant), and transportation environments (station, bus, car, metro), with SNR values from -24 dB to 24 dB. Findings show WER consistently decreasing as SNR increases. This improvement is most significant from -24 dB to 0 dB, becoming more gradual thereafter. Environmental factors strongly influenced recognition performance, with meeting rooms and cafeterias showing the highest WER peaks (~ 1.5) at low SNR levels, while car and office environments maintained lower WER across all SNR values. Transportation environments displayed similar WER patterns, indicating comparable acoustic challenges. Higher speaking effort levels (2 and 3) generally produced lower WER than effort level 1, particularly in challenging environments like cafeterias and meeting rooms. However, this advantage diminished at SNR values above 6 dB. Most environments showed performance plateaus with WER below 0.2 at SNR levels above 12 dB, with convergence across effort levels suggesting diminishing returns in favorable conditions. Results indicate that increased speaking effort effectively compensates for adverse noise conditions, though its effectiveness varies by environment. This strategy benefits challenging acoustic settings with low SNR, while becoming negligible in favorable signal-to-noise conditions.

3.2. Ablation study: Linguistic-Acoustic Interactions

This ablation study examines feature combinations' impact on WER prediction across machine learning models, identifying key features and comparing performance metrics. We test various feature combinations for WER prediction: All Features (voice type, effort, noise type, SNR, Overall Text Features Score), five exclusion sets (each removing one feature), and two specialized sets—Only Categorical (voice type, noise type) and Only Numerical (effort, SNR, Overall Text Features Score). For each set, we deploy seven machine learning models (Linear, Ridge, Lasso Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost) and a feedforward neural network with input matching feature set size, two hidden layers (64/32 units, ReLU activation), and single output neuron. The network uses Adam optimization (0.001 learning rate, MSE loss) trained for 100 epochs. All models are evaluated using MAE, RMSE, and R^2 on an 80/20 train-validation split. The all-features analysis (Table 1) shows XGBoost achieved best performance ($R^2 = 0.65$), followed by Neural Network ($R^2 = 0.61$), with excellent MAE (0.19/0.20) and RMSE (0.33/0.35). Gradient Boosting performed similarly ($R^2 = 0.61$), while Linear/Ridge Regression showed moderate results ($R^2 = 0.47$). Lasso Regression consistently underperformed across metrics. When comparing XGBoost and Neural Network models, Table 2 reveals important insights about performance across feature configurations. Both models show similar degradation patterns when features are removed, with XGBoost maintaining slight advantage. The performance gap narrows significantly when the Overall Text Features Score feature is excluded. The analysis identifies SNR as critical, with both models experiencing substantial performance deterioration upon its removal. The noise type feature also proves important, though its removal results in less severe per-

formance decline. In contrast, removing other features such as voice type, effort, and Overall Text Features Score has minimal impact on model performance. Regarding feature type importance, both models struggle equally when limited to categorical features only. However, XGBoost demonstrates slightly better handling of numerical-only features compared to the Neural Network. Overall, while both models prove robust performers, XGBoost consistently outperforms the Neural Network across most configurations, though they achieve very similar performances in response to feature modifications.

Table 1: Performance Metrics for All Features Configuration

Model	MAE	RMSE	R^2 Score
Linear Regression	0.28	0.41	0.47
Ridge Regression	0.28	0.41	0.47
Lasso Regression	0.33	0.47	0.29
Decision Tree	0.25	0.39	0.50
Random Forest	0.20	0.36	0.58
Gradient Boosting	0.21	0.35	0.61
XGBoost	0.19	0.33	0.65
Neural Network	0.20	0.35	0.61

Table 2: Comparison of R^2 Scores

Feature Configuration	XGBoost	Neural Network
All Features	0.65	0.61
No Voice type	0.64	0.60
No Effort	0.64	0.60
No Noise type	0.40	0.37
No SNR	0.20	0.17
No Overall Text Score	0.59	0.59
Only Categorical	0.14	0.14
Only Numerical	0.39	0.36

4. Discussion

This study demonstrates the enhanced intelligibility of adaptive speech synthesis systems that modulate vocal characteristics based on environmental acoustics, particularly in populated areas like meeting rooms and environments with dynamic noise like passing buses. Public announcements, assistive technologies, and voice assistants could implement context-aware strategies—using moderate vocal effort increases for transportation environments and more dramatic modifications for spaces with competing speech. The non-monotonic patterns observed, especially in cafeteria noise at mid-range SNRs, indicate that simplified linear adaptation approaches may be suboptimal. Environment-specific strategies accounting for unique spectro-temporal masking characteristics would likely achieve better intelligibility outcomes, challenging current one-size-fits-all approaches to speech enhancement. Our findings highlight complex interactions between vocal properties, noise characteristics, and intelligibility. The distinctive patterns across noise types support theories distinguishing informational masking (speech-like noises in meeting rooms and cafeterias) from energetic masking (predictable environmental sounds in kitchens and buses). Performance plateaus at low SNR in environments like café, restaurant, meeting room and station show that competing background speech is more detrimental to intelligibility than other types of noise.

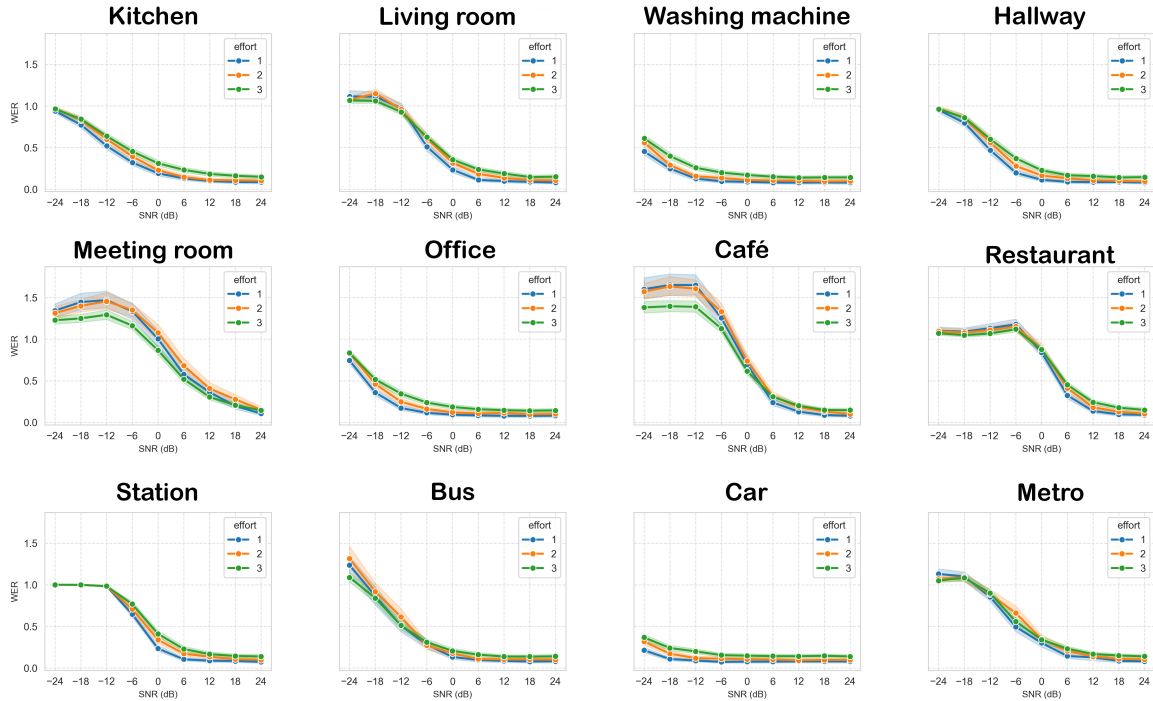


Figure 1: WER vs SNR for different noise types and effort levels.

To supplement our ASR results on noise-adaptable speech synthesis, we conducted an online listening experiment where 35 native English speakers with normal hearing were recruited through Prolific. Participants transcribed 40 speech samples from the same TTS-in-noise-corpus used in the ASR study, focusing on challenging SNR conditions (-24, -18, -12, -6, and 0 dB) across four representative noise environments: kitchen, bus, cafeteria, and meeting room, with unlimited audio replays. Samples varied across three vocal effort levels, with the 30-35 minute experiment designed to prevent fatigue while ensuring representative condition coverage. Results revealed systematic patterns with WER ranging from 0.1 to 1.2, generally improving as SNR increased. Meeting room noise consistently proved most challenging due to informational masking from competing speech. Cafeteria noise showed non-monotonic patterns at mid-range SNRs, suggesting complex noise-speech interactions. Bus and kitchen noise resulted in lower WER values with more predictable SNR-related improvements, with kitchen noise showing relatively stable performance across mid-range SNRs. Vocal effort effects were consistently noise-dependent. In bus noise, increased vocal effort was most beneficial at lower SNRs, with effort levels converging at higher SNRs. Kitchen noise showed consistent high vocal effort advantages. Meeting room environments demonstrated the most dramatic vocal effort effects. Comparing human and ASR results showed improving performance with increasing SNR and benefited from increased vocal effort, but humans exhibited more complex, non-linear patterns, particularly in cafeteria and kitchen conditions. Humans demonstrated greater sensitivity to meeting room noise and distinct "threshold effects" where performance improved dramatically beyond certain SNR values, particularly in bus noise at -15 dB and meeting room noise at -10 dB. These differences reflect that humans are still able to handle acoustic challenges better than current ASR.

5. Conclusions and implications

In conclusion, our study on adaptive vocal effort speech synthesis reveals four key findings: increased vocal effort improves intelligibility by up to 30% in adverse acoustic conditions; these benefits vary systematically by noise type and signal-to-noise ratio, with greatest improvements observed in speech-like noise environments at low SNRs; machine learning analysis identifies SNR and noise type as the most critical predictive factors for intelligibility outcomes; and human listeners demonstrate more complex perceptual patterns than ASR systems, exhibiting threshold effects and non-monotonic response patterns that reflect sophisticated auditory processing mechanisms. Future research should prioritize developing real-time adaptation mechanisms and investigate additional vocal adaptations beyond effort level modulation. The implications of this research extend directly to developing conversational AI assistants capable of dynamically adjusting speech production based on environmental acoustic conditions. Through real-time environmental sound classification, these systems could achieve adaptive speech modifications. This approach enables more robust human-computer interaction in challenging acoustic environments. We are currently implementing these findings in a situation-aware kitchen assistant designed to provide step-by-step cooking instructions to elderly users. This system employs the BEATs sound classifier [26] to perform dual functions: identifying when users are actively following instructions (e.g., chopping or rinsing vegetables) and detecting when environmental noise sources activate, such as ventilation fans or electrical appliances. When elevated noise levels are detected, the assistant automatically selects audio prompts with increased vocal effort for its zero-shot TTS to maintain intelligibility and ensure critical cooking instructions remain accessible despite adverse acoustic conditions.

6. Acknowledgements

This work is funded by the WASP-funded project PerCorSo and the Digital Futures-funded project AAIS (Advanced Adaptive Intelligent Systems).

7. References

- [1] N. Chang, Y. Liu, J. Zhang, and L. Zhang, "The significance of speech synthesis technology for the disabled," *Advances in Information Technology Research*, 2024.
- [2] A. V. Kadam, "Text-to-speech in voice assistants: Challenges and mitigation strategies," *Journal of Engineering and Applied Sciences Technology*, 2023.
- [3] R. W. Peters, B. C. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 577–587, 1998.
- [4] B. W. Hornsby, T. A. Ricketts, and E. E. Johnson, "The effects of speech and speechlike maskers on unaided and aided speech recognition in persons with hearing loss," *Journal of the American Academy of Audiology*, vol. 17, no. 06, pp. 432–447, 2006.
- [5] Y. Ma and Y. Tang, "The intelligibility benefits of modern computer-synthesized speech for normal-hearing and hearing-impaired listeners in non-ideal listening conditions," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 5, no. 1, p. 5, 2024.
- [6] P. N. Petkov and W. B. Kleijn, "Preservation of speech spectral dynamics enhances intelligibility," in *Conference of the International Speech Communication Association*, 2013.
- [7] S. P. López-Peláez and R. A. J. Clark, "Speech synthesis reactive to dynamic noise environmental conditions," in *Conference of the International Speech Communication Association*, 2014.
- [8] Y. Tang, "The intelligibility benefits of modern computer-synthesized speech for normal-hearing and hearing-impaired listeners in non-ideal listening conditions," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, 2024.
- [9] G. K. Anumanchipalli, P. K. Muthukumar, U. Nallasamy, A. Parlikar, A. W. Black, and B. Langner, "Improving speech synthesis for noisy environments," in *Speech Synthesis Workshop*, 2010.
- [10] E. M. Johnson and E. W. Healy, "Maximizing environmental sound recognition and speech intelligibility using time-frequency masking," *Berkeley Program in Law & Economics*, 2023.
- [11] B. Sharma and S. R. M. Prasanna, "Speech synthesis in noisy environment by enhancing strength of excitation and formant prominence," *Conference of the International Speech Communication Association*, 2016.
- [12] D. Erro, T.-C. Zorila, and Y. Stylianou, "Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications," *IEEE Transactions on Audio, Speech, and Language Processing*, 2014.
- [13] I. Thoidis and T. Goehring, "Using deep learning to improve the intelligibility of a target speaker in noisy multi-talker environments for people with normal hearing and hearing loss," *The Journal of the Acoustical Society of America*, vol. 156, no. 1, pp. 706–724, 2024.
- [14] H. Brumm and S. A. Zollinger, "The evolution of the lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11-13, pp. 1173–1198, 2011.
- [15] B. Sofronievski *et al.*, "Macedonian speech synthesis for assistive technology applications," in *European Signal Processing Conference*, 2022.
- [16] W. Bian, Y. Zhou, K. Zhang, and X. Gu, "Emospeech: A corpus of emotionally rich and contextually detailed speech annotations," 2024. [Online]. Available: <https://arxiv.org/abs/2412.06581>
- [17] X. Zhu, Y. Lei, T. Li, Y. Zhang, H. Zhou, H. Lu, and L. Xie, "Metts: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer," 2023. [Online]. Available: <https://arxiv.org/abs/2307.15951>
- [18] L.-W. Chen, S. Watanabe, and A. I. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," *arXiv preprint*, 2023.
- [19] A. Chingacham, M. Zhang, V. Demberg, and D. Klakow, "Human speech perception in noise: Can large language models paraphrase to improve it?" *arXiv preprint arXiv:2408.04029*, 2024.
- [20] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. DAI *et al.*, "Llms: Scaling train-time and inference-time compute for llama-based speech synthesis," *arXiv preprint arXiv:2502.04128*, 2025.
- [21] C. Valentin, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association*, Sept 2016, pp. 352–356.
- [22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [23] 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>
- [24] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," (*No Title*), 2013.
- [25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [26] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *ICML 2023*, June 2023.