Analyzing the Correlation and Impact of Speech Evaluation Metrics on Real-World Speaker Verification and Speech Recognition

Tan-Loc LE
VinBigData, VinGroup
Ha Noi, Vietnam
v.loclt8@vinbigdata.org

Van-Huy NGUYEN
VinBigData, VinGroup
Thai Nguyen University of Technology, Vietnam
Ha Noi, Vietnam
huynguyen@tnut.edu.vn

Tri-Nhan DO
VinBigData, VinGroup
Ha Noi, Vietnam
v.nhandt23@vinbigdata.org

Trung-Kien PHAN
VinBigData, VinGroup
Ha Noi, Vietnam
v.kienpt@vinbigdata.org

Dang-Khoa MAC
VinBigData, VinGroup
Ha Noi, Vietnam
v.khoamd@vinbigdata.org

***Abstract*—This study investigates the relationship between speech quality assessment metrics and the performance of downstream tasks such as Automatic Speech Recognition (ASR) and Speaker Verification (SV). We evaluate non-intrusive metrics and estimated versions of traditional intrusive metrics across multiple languages and noise conditions, focusing on their correlation with real-world task performance where clean reference audio is typically unavailable. Our experiments span Vietnamese and English datasets with varying noise types and levels. Results indicate that while estimated intelligibility metrics like STOI show strong correlation with ASR performance, the effectiveness of various non-intrusive metrics varies across languages and noise conditions. We explore enhanced prediction models combining multiple metrics to better estimate downstream performance. This work provides insights for optimizing speech processing systems in real-world applications.**

***Index Terms*—Speech Quality Assessment, Metric Correlation Analysis, Automatic Speech Recognition (ASR), Speaker Verification (SV).**

I. Introduction

In recent years, advancements in speech processing technologies have led to significant improvements in applications such as Automatic Speech Recognition (ASR) and Speaker Verification (SV) [1]–[3] . However, the effectiveness of these systems often hinges on the quality of the audio input. Various speech scoring metrics aim to assess audio quality in noisy environments, including Perceptual Evaluation of Speech Quality (PESQ) [4], Short-Time Objective Intelligibility (STOI) [5], Signal-to-Noise Ratio (SNR), Mean Opinion Score (MOS), and more recent measures like DNSMOS [6] and MOSA-Net [7]. These metrics are commonly used to evaluate speech enhancement models and predict system performance.

Despite their utility, many established metrics (like PESQ and STOI) depend on clean reference audio for comparison, creating significant challenges in real-world applications where such references are unavailable. This limitation is particularly problematic in production environments for voice assistants, call centers, and security systems, where audio quality varies widely. Recent research has attempted to address this gap through estimated models for traditional metrics [8] and non-intrusive assessment methods [6], [9]. However, these approaches have primarily been validated on English-language datasets, and their effectiveness for other languages remains uncertain.

More critically, there is limited understanding of how these various quality metrics, whether reference-based or estimated, intrusive or non-intrusive, correlate with actual ASR and SV performance across diverse real-world conditions. Previous work by Chiang et al. [10] explored correlations between objective measures and subjective quality / intelligence ratings, but did not extend to examining how these metrics predict downstream task performance. This gap limits our ability to optimize speech systems for real-world deployment.

In this paper, we address these limitations by: (1) compiling and analyzing speech scoring methods across different datasets, including both objective and subjective metrics, intrusive (estimated) and non-intrusive approaches; (2) evaluating the performance of ASR and SV models on diverse multilingual datasets with varying levels of noise; and (3) analyzing the correlation between the performance of ASR and SV tasks and various audio quality metrics to develop more reliable predictors of real-world system performance.

II. Background

A. Speech Quality Assessment

Speech quality assessment methods can be broadly categorized into subjective and objective approaches. Subjective measures like the Mean Opinion Score (MOS) are based on human evaluators rating speech samples, providing direct insight into perceived quality, but requiring significant time

and resources. Objective measures attempt to algorithmically quantify quality aspects without human intervention.

Objective measures fall into two categories: intrusive (full reference) and non-intrusive (no reference). Intrusive metrics such as PESQ [4], STOI [5], and SI-SDR [11] compare degraded speech with clean reference signals, calculating distortion or quality scores. Although accurate in controlled environments, these metrics require clean references, limiting their applicability in real-world scenarios. In practice, estimated versions of these metrics calculated without a reference are often used.

Non-intrusive metrics address the limitation of requiring reference signals by directly evaluating audio quality from degraded signals. Traditional methods, such as P.563 [12] and ANIQUE+ [13], have been widely used, but recent deep learning-based approaches, including Distill-MOS [14] and MOSA-Net [7], have demonstrated better alignment with human subjective judgments. These deep learning methods extract features from the degraded signal to estimate quality scores. Recent advancements also focus on combining multiple metrics within a single framework. For example, TorchAudio-Squim [8] uses transformer-based models to predict three objective metrics: PESQ, STOI, and SI-SDR. Similarly, Audiobox-Aesthetics [9] predicts four distinct metrics: Content Enjoyment, Content Usefulness, Production Complexity, and Production Quality. Chiang et al. [10] further showed that combining objective measures using deep learning can improve the prediction of subjective ratings, even when training data is limited. However, their work did not explore the impact of these metrics on downstream tasks.

B. Downstream tasks

Automatic Speech Recognition (ASR) and Speaker Verification (SV) represent two critical downstream applications affected by speech quality. ASR systems transcribe spoken language into text, with performance typically measured by Word Error Rate (WER) and Character Error Rate (CER). Modern ASR systems employ end-to-end deep learning architectures like Conformer [15] and Whisper [1] but remain sensitive to acoustic conditions. Speaker Verification determines whether a speech sample matches a claimed identity, measured by Equal Error Rate (EER) and Detection Cost Function (DCF). Current SV systems use embeddings like x-vectors [16] or ECAPA-TDNN [3], but their performance degrades significantly under noisy conditions. This sensitivity makes quality metrics relevant for predicting SV performance. While both ASR and SV are known to be affected by speech quality, the correlation between quality metrics and task performance is not well understood, particularly across diverse languages and conditions, motivating this work.

III. Methodology

A. Datasets

To ensure a comprehensive evaluation, we utilized multiple datasets spanning different languages and acoustic conditions for both ASR and SV tasks.

For Automatic Speech Recognition (ASR), the following datasets were used:

- **Our Internal Vietnamese Dataset**: This dataset consists of 5,000 clean audio recordings collected from Vietnamese speakers and 5,000 noisy audio samples created by adding background noise from the MUSAN dataset at random Signal-to-Noise Ratios (SNRs) between -5 dB and 20 dB. All clean recordings were captured in studio-quality conditions.
- **VoiceBank+DEMAND** [17]: A widely used English dataset comprising utterances from 28 speakers (Voice Bank corpus) mixed with various noise types from the DEMAND database.
- **URGENT-Challenge-2024** [18]: An English-only dataset featuring speech recordings captured in challenging acoustic environments characterized by varying levels of background noise, reverberation, and channel effects.
- **URGENT-Challenge-2025**: This is an expanded version of the URGENT-Challenge-2024 dataset, featuring multilingual speech recordings. It includes a greater variety of distortions (7 types compared to 4 in 2024 version), along with more diverse acoustic conditions and a wider range of speakers.

For Speaker Verification (SV), we employed the VoxCeleb1 dataset [19]. This is a large-scale dataset containing 148,642 utterances from 1,251 celebrities, extracted from interviews and videos uploaded to YouTube. To further investigate the impact of noise on SV performance, we generated corrupted versions of VoxCeleb1 by adding general additive noise with SNRs ranging from 1 dB to 20 dB.

B. Automatic Speech Recognition (ASR)

For ASR evaluation, we utilized two state-of-the-art models:

- **Whisper** [1]: A multilingual ASR model developed by OpenAI, pretrained on over 680,000 hours of multilingual and multitask data. Whisper is capable of handling a wide range of languages and was fine-tuned for our targeted tasks.
- **PhoWhisper** [2]: A fine-tuned version of Whisper designed specifically for Vietnamese automatic speech recognition. PhoWhisper was trained on an 844-hour dataset that includes diverse Vietnamese accents and has demonstrated state-of-the-art performance on benchmark Vietnamese ASR datasets.

ASR performance was evaluated using two standard metrics:

- **Word Error Rate (WER)**: The percentage of words that are incorrectly recognized.
- **Character Error Rate (CER)**: The percentage of characters that are incorrectly recognized, which is particularly critical for tonal languages like Vietnamese.

Both models were tested across all datasets under various noise conditions to assess their robustness and accuracy.

C. Speaker Verification (SV)

For speaker verification (SV) evaluation, we employed WavLM [20], a pre-trained transformer-based model specifically designed for learning robust speech representations. WavLM builds on the transformer architecture and is trained with a large-scale dataset of diverse speech and noise conditions. It leverages a self-supervised learning approach to extract high-quality speaker embeddings that are effective for speaker verification tasks.

SV performance was measured using the following metrics:

- **Equal Error Rate (EER)**: The error rate at which false acceptance and false rejection rates are equal. This metric provides a balanced measure of overall system performance.
- **Detection Cost Function (DCF)**: A weighted measure of verification errors, balancing false acceptance and false rejection rates. DCF is particularly useful in applications where different types of errors may have unequal costs.

D. Quality Metrics

We calculated the following objective quality and intelligibility metrics for all audio samples. Note that intrusive metrics were estimated without a clean reference.

- **Estimated objective metrics** :
  - **Torchaudio-squimm** [8]: A toolkit in the TorchAudio library that provides reference-less speech quality assessment through deep neural networks. It uses a single DNN operating on time-domain signals to simultaneously estimate PESQ, STOI, and SI-SDR without requiring clean reference signals. Based on dual-path recurrent neural networks (DPRNNs), it enables practical non-intrusive assessment for real-world applications where clean speech references are unavailable.
  - **Estimated Signal-to-Noise Ratio (Est-SNR)** [21]. A deep learning approach for frame-level SNR estimation with 20ms frame lengths, utilizing unidirectional RNNs for causal estimation and bidirectional RNNs for non-causal estimation. The method systematically examines 18 different monophonic features from various speech processing domains.
  - **Waveform Amplitude Distribution Analysis SNR (WADA-SNR)** [22]: A computationally efficient algorithm that estimates SNR without requiring clean reference signals. It assumes clean speech amplitude follows a Gamma distribution with shape parameter 0.4 and additive noise follows a Gaussian distribution. This method shows significantly less bias and variance compared to standard NIST STNR algorithms, even with non-Gaussian noise types.
- **Estimated subjective metrics**:
  - **Distill-MOS** [14]: Distilled MOS prediction model. This model provides subjective quality assessments without requiring reference signals, developed through knowledge distillation techniques where a larger "teacher" model (based on XLS-R embeddings) provides pseudo-labels for training more compact "student" models, making it practical for real-world scenarios.
  - **CE, CU, PC, PQ**: Confidence scores from Meta AI's Audiobox-Aesthetics models [9]. These represent a novel approach to quantifying audio aesthetics by separating human listening perspectives into four distinct axes: Content Enjoyment (how enjoyable the audio is), Content Usefulness (how informative the audio is), Production Complexity (level of audio production complexity), and Production Quality (technical quality aspects).

E. Correlation Analysis

To assess the relationship between quality metrics and downstream task performance, we calculated Pearson correlation coefficients (ρ) between each metric and the corresponding ASR/SV performance measure (WER, CER, EER). This analysis was performed separately for each dataset, noise type, SNR level, and across all conditions combined.

IV. Experiment Results

A. Automatic Speech Recognition (ASR)

We investigated the correlation between the calculated metrics and ASR model performance (WER, CER) on the test sets using Pearson Correlation. The results are summarized in Table I.

Based on Table I, we observe:

- **STOI consistently shows the highest correlation** with both WER and CER across all datasets, suggesting intelligibility-focused metrics are highly predictive of ASR performance.
- **Performance correlation decreases in challenging conditions**. All metrics show weaker correlations on the URGENT datasets compared to the cleaner Vietnamese and VoiceBank+DEMAND datasets.
- **Estimated intrusive metrics generally outperform non-intrusive ones** in correlation strength (e.g., STOI, STOI, SI-SDR vs. Distill-MOS), although this advantage diminishes in more challenging environments.
- **SNR-based metrics show poor correlation in challenging conditions**, especially WADA-SNR, suggesting they fail to capture complexities affecting ASR in difficult acoustics.
- **Subjective metrics exhibit strong correlation across all noise conditions**: The Meta AI's new metrics (CE, CU, PC, PQ) show lower correlations compared to Distill-MOS, but outperform PESQ and SI-SDR.

B. Speaker Verification (SV)

For Speaker Verification, we analyzed the correlation between quality metrics and matching scores for positive (same speaker) and negative (different speaker) pairs. Results are in Table II.

From the results above, it is evident that:

TABLE I
CORRELATION BETWEEN QUALITY METRICS AND ASR PERFORMANCE (PEARSON CORRELATION COEFFICIENT, ρ). WE PRESENT THE ABSOLUTE VALUES FOR CLARITY ON STRENGTH. HIGHER ABSOLUTE VALUES INDICATE STRONGER CORRELATION.

	Vietnamese Dataset		**VoiceBank+DEMAND**		**URGENT-Challenge-2024**		**URGENT-Challenge-2025**	
Metric	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**
PESQ	0.399288	0.369979	0.307051	0.325640	0.256634	0.216948	0.111967	0.117588
STOI	**0.737747**	**0.721363**	**0.398421**	**0.487822**	**0.468690**	**0.431914**	**0.193198**	**0.212607**
SI-SDR	0.603361	0.585338	0.259532	0.283733	0.349860	0.314995	0.145206	0.160612
EST-SNR	0.356820	0.329748	0.062199	0.065986	0.166371	0.141807	0.079400	0.088776
WADA-SNR	0.437042	0.418519	0.151336	0.160584	0.184393	0.171405	0.003034	0.012664
DISTILL-MOS	0.561118	0.528435	0.330058	0.393112	0.409577	0.361289	0.158678	0.177419
CE	0.545518	0.509238	0.341971	0.366184	0.349778	0.273659	0.174521	0.175045
CU	0.422625	0.384213	0.287556	0.296821	0.260358	0.191318	0.161874	0.156699
PC	0.385574	0.362968	0.270683	0.295064	0.263588	0.239510	0.088883	0.110178
PQ	0.453275	0.411855	0.254865	0.263796	0.201392	0.133808	0.159955	0.157810

TABLE II
CORRELATION BETWEEN AVERAGE QUALITY METRICS AND SV MATCHING SCORES (PEARSON ρ). HIGHER SCORES ARE DESIRED FOR POSITIVE PAIRS, LOWER FOR NEGATIVE PAIRS.

Metric	Positive Pairs (Same Speaker)	Negative Pairs (Different Speakers)
PESQ_avg	0.275169	-0.079335
STOI_avg	0.375938	-0.096068
SI-SDR_avg	0.307812	-0.079471
EST-SNR_avg	0.191972	-0.064522
WADA-SNR_avg	0.183765	-0.061650

- **STOI shows the strongest correlation** with matching scores for positive pairs and the most negative correlation for negative pairs, aligning with ASR findings.
- Correlations are modest overall, particularly for negative pairs, suggesting metrics are less predictive of impostor rejection.
- **SNR-based metrics show the least correlation**.

The results indicate that speech intelligibility significantly affects the accuracy of the accepted rate. In other words, the accuracy of the embedding model is more closely related to the STOI metric than to other quality based metrics such as SNR.

1) Impact of Signal-to-Noise Ratio on EER and Quality Metrics: We created 20 test sets with SNR from 1 dB to 20 dB. Table III shows the EER, threshold, and metric values. Table IV shows the correlation between EER and metrics across 5 SNR levels.

According to Table III and Table IV, we note that:

- **Strong negative correlation between all metrics and EER**. Better quality strongly predicts lower SV error.
- **STOI demonstrates the highest correlation with EER** (-0.9957), reinforcing its strong relationship with SV performance found earlier.
- EER decreases consistently as SNR increases.
- The verification threshold shifts higher with better quality, suggesting calibration should account for quality.

V. DISCUSSION

Our findings have several implications for speech processing applications:

A. Metric Selection for ASR and SV

Our results indicate that the estimated **STOI shows the strongest correlation with performance for both ASR (WER/CER) and SV (matching scores, EER)** across the tested conditions. While STOI focuses on intelligibility crucial for recognition, its strong correlation with SV suggests intelligibility is also paramount for preserving speaker characteristics or enabling the SV system to function effectively. Estimated SI-SDR also shows strong correlation, particularly for SV, highlighting the relevance of signal separation. However, STOI demonstrated the most consistent high correlation across tasks in our study. This suggests prioritizing intelligibility metrics like STOI might benefit both ASR and SV system evaluation, although task-specific nuances remain important.

B. Limitations of Current Metrics

All metrics show significantly reduced correlation with performance in challenging acoustic conditions (URGENT datasets). This indicates a gap in predicting real-world performance using existing metrics, whether estimated intrusive or non-intrusive.

C. Language Independence

We found that although the estimated models were trained on English data, they still performed effectively on other languages (in our case, Vietnamese), as long as the data was sufficiently clean. This suggests that deep learning models can achieve better-than-expected generalization, allowing them to be applied to various languages without needing to retrain from scratch. However, evaluating and adjusting the models for other languages must be done carefully to ensure that these models do not degrade recognition quality. Careful control for confounding factors like model maturity for each language is still needed.

TABLE III
RELATIONSHIP BETWEEN SNR, EER, VERIFICATION THRESHOLD, AND SPEECH QUALITY METRICS.

SNR (dB)	EER	Threshold	PESQ	STOI	SI-SDR	EST-SNR	WADA-SNR
1	0.1402	0.8208	1.1589	0.7600	0.4074	-1.7605	5.1121
5	0.0968	0.8437	1.2148	0.8214	4.2526	1.0816	7.3492
10	0.0697	0.8589	1.3667	0.8822	8.7517	4.6996	10.2863
15	0.0578	0.8676	1.6215	0.9236	12.6708	8.0717	13.0217
20	0.0530	0.8705	1.9435	0.9479	15.7345	10.9520	15.2866

TABLE IV
CORRELATION BETWEEN EER AND SPEECH QUALITY METRICS ACROSS SNR LEVELS (FROM TABLE III DATA). HIGHER QUALITY CORRELATES WITH LOWER EER.

Metric	Correlation with EER (ρ)
PESQ	-0.9823
STOI	-0.9957
SI-SDR	-0.9872
EST-SNR	-0.9815
WADA-SNR	-0.9789

D. Practical Applications

For practical deployment scenarios lacking clean references, using estimated metrics like STOI or SI-SDR provides valuable insights. Furthermore, exploring composite models combining multiple non-intrusive metrics (as briefly investigated via regression) offers a promising approach to potentially achieve prediction accuracy closer to reference-based estimates, enabling better quality monitoring and adaptive processing.

VI. CONCLUSION AND FUTURE WORK

This study provides a comprehensive analysis of the relationship between various speech quality/intelligibility metrics and downstream ASR/SV task performance across multiple languages and acoustic conditions. Our findings demonstrate that estimated intelligibility metrics like STOI show strong correlation with task performance, particularly in controlled to moderately noisy environments, but predictive power diminishes in highly challenging conditions. Quality-based metrics offer practical alternatives but generally show weaker correlations.

In the future, we will focus on:

- Expanding the analysis to more languages, tasks, and diverse acoustic environments.
- Further developing and evaluating composite prediction models combining multiple non-intrusive features.
- Exploring the integration of real-time quality prediction into adaptive speech processing systems (e.g., adjusting ASR models or SV thresholds based on estimated quality).
- Investigating task-specific quality aspects (e.g., how different distortion types affect ASR vs. SV differently).

By bridging the gap between quality assessment and task performance prediction, we can develop more reliable and adaptable speech processing systems for diverse real-world conditions.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[2] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, "PhoWhisper: Automatic Speech Recognition for Vietnamese," in *Proceedings of the ICLR 2024 Tiny Papers track*, 2024.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[4] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq) - a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[5] 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.

[6] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[7] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[8] A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," 2023. [Online]. Available: https://arxiv.org/abs/2304.01448

[9] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," 2025. [Online]. Available: https://arxiv.org/abs/2502.05139

[10] H.-T. Chiang, K.-H. Hung, S.-W. Fu, H.-C. Kuo, M.-H. Tsai, and Y. Tsao, "Study on the correlation between objective evaluations and subjective speech quality and intelligibility," 2023. [Online]. Available: https://arxiv.org/abs/2307.04517

[11] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[12] L. Malfait, J. Berger, and M. Kastner, "P.563—the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[13] D.-S. Kim and A. Tarraf, "Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007.

[14] B. Stahl and H. Gamper, "Distillation and pruning for scalable self-supervised representation-based speech quality assessment," 2025. [Online]. Available: https://arxiv.org/abs/2502.05356

[15] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, 2020, pp. 5036–5040.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[17] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152.

[18] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, "Urgent challenge: Universality, robustness, and generalizability for speech enhancement," in *Interspeech 2024*, 2024, pp. 4868–4872.

[19] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819302712

[20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[21] H. Li, D. Wang, X. Zhang, and G. Gao, "Frame-level signal-to-noise ratio estimation using deep learning," in *Interspeech 2020*, 2020, pp. 4626–4630.

[22] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis." in *Interspeech*, 2008, pp. 2598–2601.