ViNonverbal: The First Vietnamese Nonverbal Speech Dataset via Scalable Pseudo-Labeling

Minh-Ngoc Nguyen[1], Tu-Anh Nguyen[2], Tri-Nhan Do[2], Hoang-Yen Nguyen[2], Dang-Khoa Mac[2]
[1]Hanoi University of Science and Technology, Hanoi, Vietnam
[2]VinFast Joint Stock Company, Hanoi, Vietnam
Email: ngocminhgb123@gmail.com, v.anhnct2@vinfast.vn, v.nhandt23@vinfast.vn, v.yennth90@vinfast.vn, v.khoamd@vinfast.vn

***Abstract*—Nonverbal vocalizations, such as laughter and fillers, are essential for expressive speech recognition and synthesis, yet they are largely neglected in Vietnamese ASR and TTS systems. We introduce ViNonverbal, a Vietnamese nonverbal speech resource consisting of a manually annotated 500-sample benchmark and an additional 4,693 pseudo-labeled speech segments covering eight nonverbal event types. We fine-tune Qwen3-ASR and Whisper-large-v3 using extended nonverbal tokens to bridge this gap. Experimental results demonstrate that Qwen3-ASR achieves a superior F1-score of 57.21% for nonverbal event recognition, significantly outperforming the Whisper baseline. Furthermore, the model maintains high transcription stability with a WER of 6.16% on VIVOS and 13.61% on the ViNonverbal test set. These findings provide a robust foundation for developing more natural, human-like Vietnamese ASR and TTS systems that can understand and generate expressive nuances.**

***Index Terms*—Nonverbal, vietnamese, Text-to-speech, Speech-to-text, ViNonverbal**

I. INTRODUCTION

Recent advances in Text-to-Speech (TTS) have significantly improved speech naturalness through end-to-end neural models such as Tacotron2 [1] and VITS [2]. However, generating truly human-like speech remains a challenge, as existing systems primarily focus on lexical content while overlooking paralinguistic signals such as laughter, breathing, and fillers. These nonverbal vocalizations (NVs) are essential for conveying emotions and conversational dynamics, playing a critical role in natural communication [**?**].

Recent studies suggest that incorporating paralinguistic information can substantially enhance TTS expressiveness. For instance, emotion-controllable models [3] and specialized datasets [4] have enabled more natural synthesis. More recently, explicit modeling of NVs in datasets like NonverbalTTS [5] and NonVerbalSpeech-38K [6] has demonstrated that NV-aware training significantly improves both speech synthesis and recognition.

For Vietnamese, existing datasets such as VIVOS and VLSP primarily consist of clean, read speech lacking nonverbal annotations, which limits their applicability for expressive tasks. This highlights a critical gap for low-resource languages. Meanwhile, the emergence of robust multilingual models like Whisper [7], wav2vec 2.0 [8], and Qwen3-ASR [9] offers a promising solution. These models' strong cross-lingual generalization enables the transfer of nonverbal knowledge from high-resource languages to low-resource settings via scalable pseudo-labeling pipelines.

In this paper, we introduce ViNonverbal, the first Vietnamese speech dataset featuring explicit nonverbal annotations, constructed from in-the-wild audio. Our approach follows a two-stage strategy. First, we manually annotate a 500-sample Vietnamese benchmark and use it to compare bilingual nonverbal-aware Qwen3-ASR and Whisper-large-v3 models. Based on its superior nonverbal recognition performance on this benchmark, Qwen3-ASR is selected as the teacher model to generate pseudo-labels for an additional 4,693 Vietnamese speech segments. The dataset, comprising eight conversational nonverbal categories, along with a demo showcase, is publicly available at: https://nmngoc06.github.io/ViNonverbal-Demo-Showcase/.

Our primary contributions are summarized as follows:

1) A scalable pipeline for constructing nonverbal-aware speech datasets from large-scale, in-the-wild Vietnamese audio.
2) A cross-lingual pseudo-labeling framework utilizing fine-tuned multilingual ASR models to bridge the resource gap for Vietnamese.
3) The construction of **ViNonverbal**, comprising a manually annotated 500-sample benchmark and 4,693 additional pseudo-labeled Vietnamese speech segments with explicit nonverbal tokens.

II. RELATED WORK

A. Specch Recognition Dataset

Large-scale speech recognition datasets have played a fundamental role in the development of modern ASR systems. Widely used benchmarks such as LibriSpeech [10] provide approximately 1000 hours of transcribed speech derived from audiobooks , while Common Voice [11] offers a massively multilingual corpus collected via crowdsourcing . More recent multilingual datasets such as VoxPopuli [12] and Multilingual LibriSpeech (MLS) [13] further enable cross-lingual learning by providing large-scale speech data across multiple languages.

Despite their scale and diversity, these datasets primarily focus on verbal content and do not explicitly model nonverbal vocalizations, such as laughter, sighs, or fillers. This limitation restricts their ability to capture the full richness of human communication.

B. Nonverbal Speech Datasets

Nonverbal vocalizations constitute an essential component of human communication, conveying rich information about emotional states, speaker intent, and interaction dynamics beyond lexical content. Prior research in paralinguistics has consistently demonstrated that incorporating nonverbal cues can significantly improve performance in tasks such as emotion recognition and speaker state modeling [14]. From a psychological perspective, models such as the circumplex model of affect [15] further support that a substantial portion of affective information is expressed through nonverbal signals rather than purely through words.

Motivated by the importance of these signals, recent works have introduced datasets that explicitly model nonverbal vocalizations. Representative examples include NonverbalTTS [5], NonVerbalSpeech-38K [6], and Emilia-NV [16], which provide annotations for a variety of nonverbal events collected from in-the-wild sources such as videos and conversational speech. More recent studies have also explored scalable pipelines and modeling frameworks for nonverbal-aware speech processing [5], [6], highlighting the growing interest in integrating paralinguistic information into speech systems.

Despite these advances, several limitations remain. First, existing datasets are largely concentrated on high-resource languages such as English and Chinese, leaving low-resource languages underexplored. Second, inconsistencies in annotation schemes across datasets hinder interoperability and the development of unified models. Furthermore, many existing approaches rely heavily on fully automatic annotation pipelines, which may introduce noise and reduce reliability in low-resource scenarios. Most importantly, to the best of our knowledge, there is currently no Vietnamese speech dataset specifically designed to model nonverbal vocalizations. As a result, current Vietnamese ASR systems remain limited in their ability to capture expressive and contextual aspects of speech.

C. Multilingual and Semi-supervised ASR

Recent advances in self-supervised and multilingual learning have significantly improved ASR performance. Models such as wav2vec 2.0 [8], HuBERT [17], and WavLM [18] demonstrate strong capability in learning speech representations from large-scale unlabeled data. In particular, multilingual models such as Whisper [7] have shown robust cross-lingual generalization and improved performance across diverse languages .

Additionally, semi-supervised approaches like pseudo-labeling are widely used to leverage weakly labeled data, reducing reliance on manual annotation by utilizing strong pretrained models [8], [17], [18].

These developments provide a foundation for integrating nonverbal signals into ASR. We adopt a semi-supervised strategy, fine-tuning a multilingual model to generate pseudo-labels for nonverbal vocalizations.

Despite these gains, nonverbal-aware ASR remains unaddressed for Vietnamese. To bridge this gap, we introduce **ViNonverbal**, the first Vietnamese dataset specifically designed for modeling nonverbal vocalizations.

III. DATASET CONSTRUCTION

We propose a scalable pipeline to construct the first Vietnamese speech corpus featuring nonverbal annotations, leveraging diverse real-world movie dialogues. The construction workflow, illustrated in **Fig. 1**, integrates a fine-tuned **Qwen3-ASR** for automated pseudo-labeling. This four-stage process—comprising collection, preprocessing, automated labeling, and refinement—significantly enhances annotation efficiency while yielding a high-fidelity dataset optimized for both expressive TTS and robust, nonverbal-aware ASR systems.

A. Data Collection and Processing

Audio data are collected from YouTube movie dialogues to capture diverse conversational scenarios and rich nonverbal signals. The raw audio is standardized into a unified format, including mono conversion, resampling,

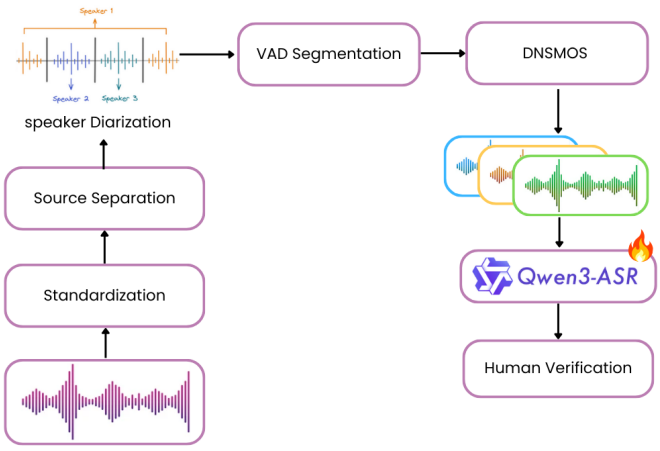

Fig. 1. End-to-end pipeline for building a Vietnamese nonverbal speech dataset with ASR-based pseudo labeling

and loudness normalization, to ensure consistency across sources.

To improve speech quality, we apply a speech separation model to isolate vocal components from background noise. The processed audio is then segmented using a combination of speaker diarization and voice activity detection (VAD). Specifically, we employ the pyannote.audio diarization model [19] to obtain speaker-aware segments, followed by VAD to produce fine-grained speech segments suitable for ASR processing, as commonly adopted in modern speech recognition systems [7], [8].

The resulting segments are further refined through post-processing, including merging adjacent segments, splitting overly long segments, and enforcing duration constraints (e.g., 3–30 seconds). This step ensures speaker-consistent and length-controlled segments that are suitable for downstream modeling.

To ensure data quality, we adopt DNSMOS [20] as an objective metric to evaluate each segment. For each segment s_i , the quality score is computed as:

$$q_i = \text{DNSMOS}(s_i) \tag{1}$$

Segments with scores below a predefined threshold τ are discarded to reduce noise and improve dataset reliability.

Finally, the retained segments are exported as individual audio files with associated metadata (e.g., start time, end time, speaker labels), forming the input for the subsequent ASR-based pseudo-labeling stage. The entire pipeline is designed to be scalable and suitable for large-scale data as well as low-resource languages.

B. *ASR-based Pseudo Labeling*

To automatically annotate both speech content and nonverbal signals, we adopt a pseudo-labeling approach based on a multilingual ASR model. We fine-tune Qwen3-ASR [9] on a mixture of multilingual datasets with nonverbal annotations, including NonverbalTTS [5], NonVerbalSpeech-38K [6], and Emilia-NV [16], together with Vietnamese speech data without nonverbal labels. This strategy leverages the cross-lingual consistency of nonverbal vocalizations [11] and the effectiveness of multilingual and semi-supervised training in low-resource ASR [7], [10].

Due to inconsistencies in annotation schemas across datasets, we normalize all nonverbal labels into a unified set of eight categories, including `<gasp>`, `<laugh>`, `<cough>`, `<sigh>`, `<sniffle>`, `<owh>`, `<uhm>`, and `<ah>`.

For example, `[[Laughter]`$\rightarrow$ `<laugh>]`, `[[Sigh]` $\rightarrow$ `<sigh>]`, `[[sniffing]` $\rightarrow$ `<sniffle>]`and `[[coughing]` $\rightarrow$ `<cough>]`. This normalization ensures consistency across datasets and improves model robustness.

We select Qwen3-ASR [9] as the backbone due to its strong multilingual capability and robustness in real-world conditions. Nonverbal events are represented as special tokens (e.g., `<laugh>`) embedded directly in the transcript, enabling joint modeling of verbal and nonverbal information.

C. *Human Verification*

To ensure high data quality, we employ a human-in-the-loop paradigm where annotators manually verify and refine pseudo-labels generated by the ASR model [21], [22]. Annotators are tasked with correcting transcription errors and validating the accuracy of nonverbal tokens. Missing, redundant, or mislabeled signals are manually adjusted according to the predefined label set, significantly improving the reliability of the resulting dataset.

D. *Vietnamese Pseudo-Label Generation*

Transcripts are normalized into a unified format with consistent nonverbal token representations. We enhance data quality and diversity by removing duplicates and segments with severe noise, misalignment, or annotation errors.

The resulting dataset comprises paired audio and transcripts containing both lexical content and embedded nonverbal labels. Each sample is enriched with metadata, including timestamps and speaker information. Finally, the corpus is partitioned into training, validation, and

test sets to support downstream ASR and TTS tasks, following standard practices in large-scale speech dataset construction [21].

IV. EXPERIMENTS

A. Experimental setup

We design experiments to evaluate the effectiveness of the proposed pipeline in constructing nonverbal-aware speech data and supporting ASR models for Vietnamese.

TABLE I
DATASET COMPOSITION FOR TRAINING AND EVALUATION (SAMPLE-BASED)

Dataset	Language	Orginal	Filtered
NonverbalTTS [5]	EN	6256	1,809
NonverbalSpeech [6]	EN&ZH	38.718	30,625
Emilia-NV [16]	ZH	146,099	34,749
Genrated	EN	-	1,300
VN-dataset $*$	VN	-	40,000

$*$Vietnamese dataset without nonverbal annotations

- We fine-tune multilingual pre-trained models, including **Qwen3-ASR-1.7B** and **Whisper-large-v3**, where the models are augmented with an extended vocabulary of nonverbal tokens.
- The training data consist of multilingual datasets with nonverbal annotations, including **NonverballTTS**, **NonVerbalSpeech-38K**, and **Emilia**, which are normalized into a unified set of **8 nonverbal labels**.
- To further balance the distribution of nonverbal categories, we augment the English data using **orpheus-3b-0.1-ft** to generate additional nonverbal samples.
- In addition, Vietnamese speech data without nonverbal annotations are included to preserve language-specific characteristics.

All audio data are resampled to a uniform sampling rate of 16kHz to ensure consistency across datasets and compatibility with ASR models; detailed dataset composition is summarized in **Table I**

We evaluate nonverbal event recognition using ViNonverbal-Benchmark, an independently manually annotated set of 500 Vietnamese conversational speech samples. This benchmark is separate from the additional 4,693 samples pseudo-labeled by the selected Qwen3-ASR teacher model. For standard ASR performance, we report results on the VIVOS benchmark. Our models are compared against strong baselines, including PhoWhisper [23] and wav2vec2-large-vi-vlsp2020 [23], using Word Error Rate (WER) as the primary metric for transcription quality.

For nonverbal prediction, we evaluate performance using token-level Precision, Recall, and F1-score. Following the evaluation protocol in event extraction [24], a predicted nonverbal token is considered correct only if both its category and positional offset in the transcription exactly match the ground truth. This strict matching criteria ensures that the model not only identifies the correct nonverbal event but also places it accurately within the lexical context. All metrics are reported as micro-averaged scores over the entire evaluation set.

B. Nonverbal-Aware ASR Evaluation

The evaluation results are summarized in **Table II**. We report Word Error Rate (WER) for ASR performance on both the VIVOS benchmark and the proposed ViNonverbal test set, and token-level precision, recall, and F1-score for nonverbal prediction.

On VIVOS, Qwen3-ASR achieves a WER of 6.16%, which is competitive with strong Vietnamese ASR baselines, indicating that nonverbal-aware training does not degrade general ASR performance. On the proposed dataset, Qwen3-ASR outperforms Whisper in WER (13.61% vs. 19.54%), demonstrating better robustness in conversational speech with nonverbal events.

For nonverbal prediction, Qwen3-ASR shows consistently better performance than Whisper. The higher F1-score suggests a more balanced trade-off between precision and recall, indicating stronger capability in modeling nonverbal vocalizations.

Overall, the results confirm that the proposed dataset enables nonverbal-aware speech recognition while preserving competitive ASR performance.

C. ViNonverbal Dataset Analysis

1) Dataset Overview: The **ViNonverbal** dataset comprises approximately 10 hours of Vietnamese conversational speech collected from in-the-wild sources, including movie dialogues. After preprocessing and filtering, the dataset contains 4,693 audio segments with diverse conversational contexts and spontaneous nonverbal (NV) events, making it suitable for real-world speech applications.

2) Nonverbal Label Distribution: The dataset covers eight types of nonverbal vocalizations, including `<laugh>`, `<cough>`, `<sigh>`, `<sniffle>`, `<gasp>`, `<owh>`, `<uhm>`, and `<ah>`. As shown in

TABLE II
EVALUATION RESULTS ON TWO BENCHMARKS: VIVOS (ASR) AND VINONVERBAL (ASR + NONVERBAL PREDICTION).

Model	Paras	VIVOS	ViNonverbal-Benchmark (500 samples)			
		WER(%)	WER(%)	Precision(%)	Recall(%)	F1(%)
PhoWhisper (large) [23]	1.55B	4.67	11.82	0.0	0.0	0.0
wav2vec2-large-vi-vlsp2020	317M	8.61	14.79	0.0	0.0	0.0
Whisper-large-v3 (ft)	1.55B	10.73	19.54	48.71	27.44	27.85
Qwen3-ASR (ft)	1.7B	6.16	13.61	62.08	56.67	57.21

***Note**: "ft" denotes fine-tuning from Whisper-large-v3 and Qwen3-ASR with extended nonverbal tokens.*

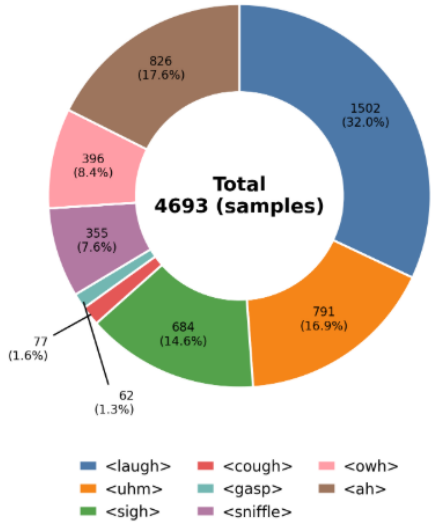

Fig. 2. Distribution of nonverbal vocalizations in the ViNonverbal dataset.

Fig 2, the distribution is inherently imbalanced, with frequent events such as `<laugh>` and `<uhm>` dominating the dataset. To address this, we apply label normalization and data augmentation, including synthetic generation of nonverbal samples, to improve data balance and model robustness.

3) Data Quality: We ensure data quality through a multi-stage preprocessing pipeline, including noise reduction, segmentation, and DNSMOS-based filtering to remove low-quality segments. Additionally, a subset of the data is manually verified to refine transcription and nonverbal annotations. This human-in-the-loop process improves annotation reliability and reduces labeling noise in the final dataset.

V. CONCLUSION

In this paper, we introduced **ViNonverbal**, the first Vietnamese speech dataset specifically designed for nonverbal-aware ASR, comprising approximately **4,693** high-quality samples with 8 distinct nonverbal event types. By employing a scalable, semi-automatic pipeline, we demonstrated that integrating nonverbal annotations into state-of-the-art multilingual models **Qwen3-ASR**

Experimental results underscore the robustness of our approach, with the fine-tuned **Qwen3-ASR** achieving a leading **F1-score of 57.21%** for nonverbal recognition while maintaining a competitive **Word Error Rate (WER) of 6.16%** on the VIVOS benchmark. These findings confirm that modeling expressive nuances does not compromise standard transcription performance, even in low-resource languages like Vietnamese.

The ViNonverbal dataset provides a foundational resource for future research in expressive speech understanding. Future work will focus on expanding the dataset with more diverse conversational sources, refining annotation granularity, and exploring downstream applications in **expressive speech synthesis (TTS)** to enable more natural and human-like machine interactions.

REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 5530–5540.

[3] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5734–5738.

[4] C. Cui, Y. Ren, J. Liu, F. Chen, R. Huang, M. Lei, and Z. Zhao, "Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model," *arXiv preprint arXiv:2106.09317*, 2021.

[5] M. Borisov, E. Spirin, and D. Diatlova, "Nonverbaltts: A public english corpus of text-aligned nonverbal vocalizations with emotion annotations for text-to-speech," *arXiv preprint arXiv:2507.13155*, 2025.

[6] R. Ye, Y. Zhou, R. Yu, Z. Lin, K. Li, X. Li, X. Liu, G. Zeng, and Z. Wu, "A scalable pipeline for enabling non-verbal speech generation and understanding," *arXiv preprint arXiv:2508.05385*, 2025.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2023.

[8] 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 (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.

[9] X. Shi, X. Wang, Z. Guo, Y. Wang, P. Zhang, X. Zhang, Z. Guo, H. Hao, Y. Xi, B. Yang, J. Xu, J. Zhou, and J. Lin, "Qwen3-asr technical report," *arXiv preprint arXiv:2601.21337*, 2026.

[10] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.

[11] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2408–2412, 2010.

[12] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 993–1003.

[13] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[14] B. Schuller and A. Batliner, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[15] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[16] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, "Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 4044–4054, 2025.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[18] 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, M. Zeng, X. Yu, and F. Wei, "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.

[19] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: Neural building blocks for speaker diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7124–7128.

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

[21] M. Liu, C. Zhang, H. Xing, C. Feng, M. Chen, J. Bishop, and G. Ngapo, "Scalable data annotation pipeline for high-quality large speech datasets development," *arXiv preprint arXiv:2109.01164*, 2021.

[22] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, and P. Martínez-Barco, "Applying human-in-the-loop to construct a dataset for determining content reliability to combat fake news," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107152, 2023.

[23] 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.

[24] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2015, pp. 167–176.