

RocknRiver: Multilingual Speaker Anonymization System for VoicePrivacy 2026 Challenge

Gia-Huy Bui, Minh-Xuan Phan, Quoc-Huy Nguyen, Candy Olivia Mawalim, Kertkeidkachorn Natthawut, Masashi Unoki

Japan Advanced Institute of Science and Technology, Japan

{giahuy.bui, xuan.phan, hqnguyen, candylim, natt, unoki}@jaist.ac.jp

Abstract

A multi-objective fine-tuning framework is proposed for speaker anonymization in the VoicePrivacy 2026 Challenge, jointly optimizing the speech synthesis model for privacy protection, speech intelligibility, and emotion preservation. The framework is guided by strong semi-informed automatic speaker verification (ASV), automatic speech recognition (ASR), and speech emotion recognition (SER) models. To further strengthen anonymization, a prosodic normalization strategy is introduced to suppress speaker-discriminative prosodic cues and a multilingual reference-speaker selection strategy to suppress residual speaker identity information in speaker embeddings. Experimental results demonstrate that the proposed system substantially improves privacy protection, with the best configuration increasing Equal Error Rate (EER) by 15% over the strongest baseline while maintaining competitive Word Error Rate (WER) and Unweighted Average Recall (UAR).

Index Terms: speaker anonymization, multi-objective fine-tuning, speech synthesis, prosodic normalization

1. Introduction

Speech signals inherently contain privacy-sensitive information beyond linguistic content, including speaker identity, accent, age, speaking style, and emotional state [1]. As speech data are increasingly collected and shared, protecting speaker privacy has become a critical requirement for trustworthy speech technologies [2].

To advance research in speech privacy protection, the VoicePrivacy Challenge (VPC) [3] has been held since 2020 to provide standardized benchmarks for evaluating speaker anonymization systems, which aim to conceal speaker's identity while preserving speech quality. The challenge evaluates systems using three complementary perspectives: privacy protection, speech intelligibility, and emotion preservation (starting from the 2022 edition). These are measured using Equal Error Rate (EER), where a higher value indicates stronger resistance against Automatic Speaker Verification (ASV); Word Error Rate (WER), where a lower value indicates better speech intelligibility; and Unweighted Average Recall (UAR), where a higher value indicates better preservation of emotional content.

VPC 2026 retains these evaluation metrics while introducing several new challenges [4]. In addition to the original English anonymization task (Track 1), as new multilingual task (Track 2) requires systems to anonymize speech in four language: English, French, German and Spanish. Furthermore, privacy evaluation adopts a substantially stronger semi-informed ASV attacker based on WavLM self-supervised representations, making residual identity cues more difficult to conceal. A cross-gender evaluation is also introduced in Track 1,

further increasing its difficulty.

Among the official baselines, BM3 achieves the strongest overall performance for multilingual privacy preservation. BM3 is a TTS-based voice anonymization system that uses Whisper to extract the linguistic content of the input speech and ECAPA-TDNN [5] to obtain the speaker representation. The original speaker representation is replaced with an anonymized speaker embedding generated using a Generative Adversarial Network (GAN), which is then used by IMS Toucan [6] to synthesize speech with the anonymized identity. Unlike systems that explicitly modify prosodic features such as F0 and energy, BM3 relies on the TTS model to generate these characteristics during synthesis. Nevertheless, the framework has two limitations. First, GAN-based anonymization maps speaker embeddings to a learned distribution whose statistical characteristics may be exploited by a semi-informed attacker trained on anonymized speech [7]. Second, because prosodic features are deterministically predicted from the transcription, they may retain speaker-discriminative information [8].

To address these challenges, this paper propose a multi-objective fine-tuning framework that jointly optimizes the speech synthesis model for privacy protection, speech intelligibility, and emotion preservation. Motivated by the stronger semi-informed attacker adopted in VPC 2026, the anonymization model is optimized against a pretrained semi-informed ASV while jointly preserving speech intelligibility and emotion using ASR and SER supervision. Next, the learned anonymized embedding distribution is also replaced with multilingual reference-speaker sampling to better conceal speaker identity. Finally, predicted prosodic features are normalized to suppress speaker-discriminative prosodic cues.

2. System Description

Fig. 1 illustrates the proposed anonymization framework, which consists of three main components: Speaker Embedding Selection, Prosodic Normalization, and Multi-Objective Fine-tuning. The framework aims to optimize IMS Toucan to generate high-quality synthesized speech that can withstand the stronger semi-informed ASV evaluation. Although VPC 2026 emphasizes multilingual anonymization, the EERs for French, German, and Spanish are already close to perfect, leaving limited room for further improvement [4]. In contrast, English remains considerably more challenging under the stronger ASV evaluation. Therefore, the proposed method focus on improving anonymization performance for English while maintaining the existing performance across other languages.

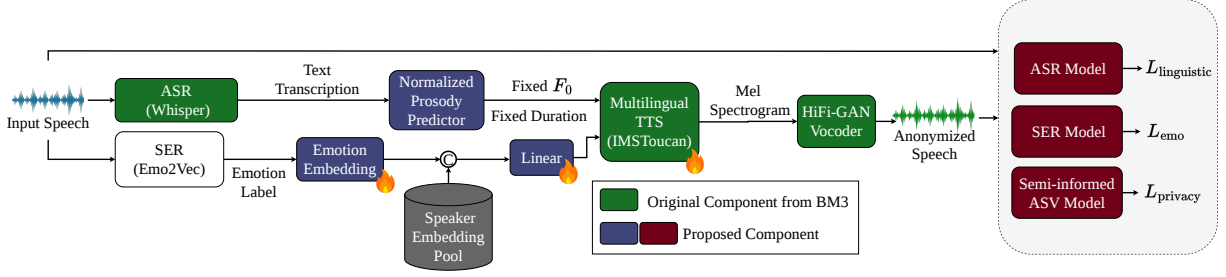


Figure 1: Detailed architecture of the proposed anonymization system. An emotion label is first mapped to an emotion embedding and concatenated with the selected target speaker embedding to control the emotional characteristics of the synthesized speech. During fine-tuning, IMSToucan is guided by three models responsible for privacy preservation, emotion consistency, and linguistic content preservation.

2.1. Speaker Embedding Selection

To replace the GAN-based anonymization module in BM3, speaker embeddings are selected from a multilingual pool of reference speakers rather than generated by a learned mapping. For each input utterance, the extracted speaker embedding is replaced with the selected reference embedding before synthesis.

This study investigate two speaker selection strategies. In the first strategy, a single target reference embedding from different dataset is selected and used for all utterances, resulting in a common anonymized identity across entire data. In the second strategy, each utterance is independently assigned a randomly selected target reference embedding from a pool, where each reference embedding is extracted from a different speaker. This strategy introduces greater variability in the anonymized speaker space and reduces consistency across utterances originating from the same source speaker.

2.2. Prosodic Normalization

To suppress speaker-discriminative prosodic cues, the predicted pitch and duration are replaced with fixed reference sequences before synthesis. The predicted energy is left unchanged because our preliminary experiments showed that modifying it substantially degrades speech quality. The resulting prosodic features are then provided to IMSToucan [6] to generate the anonymized speech. As a result, the proposed method produces speech with a more consistent F_0 range and substantially less pitch variation than BM3, whose outputs exhibit noticeably greater variation. Such speaker-dependent variation in F_0 may provide residual cues that can be exploited by an ASV system to distinguish and associate anonymized speech with the original speaker.

2.3. Multi-Objective Fine-tuning

To improve privacy while preserving utility, the IMSToucan [6] synthesizer is fine-tuned using a combination of emotional speech datasets and synthetic emotional speech generated with CosyVoice 2.0 [9].

To explicitly control emotional content, an extra emotion embedding layer is introduced into IMSToucan. Each emotion is represented by a learnable embedding vector, which is concatenated with the speaker embedding and projected through a linear layer before being fed to the downstream model.

During inference, the input utterance is first classified by an SER model, and the predicted emotion embedding is combined with the anonymized speaker embedding to preserve the origi-

nal emotional characteristics. Since prosody normalization can suppress emotional expression, the fixed pitch and duration described in Section 2.2 are applied only to utterances classified as neutral. For non-neutral speech, the prosodic features predicted by the fine-tuned BM3 model are retained.

To further preserve emotional characteristics, an emotion consistency loss is also employed. For each emotion category, a prototype is computed from pre-extracted SER emotion embeddings. We adopt a prototype-based contrastive loss [10], which encourages the embedding of anonymized speech to remain close to the prototype of its target emotion while separating it from other emotion categories. The loss is defined in Eq (1):

$$L_{\text{emo}} = -\log \frac{\exp(\text{sim}(c_a, c_y)/\tau)}{\sum_{k=1}^K \exp(\text{sim}(c_a, c_k)/\tau)}, \quad (1)$$

$$c_k = \frac{1}{N_k} \sum_{i=1}^{N_k} c_i^{(k)}, \quad (2)$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function. Each c_k represents the centroid of emotion class k , calculated as the average of the N_k embeddings $c_i^{(k)}$. K denotes the number of emotion classes, τ is the temperature parameter of the contrastive loss, and c_a denotes the speaker embedding of the anonymized speech.

To further enhance privacy while preserving linguistic content, two additional objectives are introduced during fine-tuning. First, a semi-informed ASV model is trained on BM3-anonymized speech to simulate a strong attacker. The ASV model is then frozen and used as a differentiable privacy objective. Given original and anonymized speech, speaker embeddings extracted by the frozen ASV are encouraged to be dissimilar by minimizing:

$$L_{\text{privacy}} = \text{sim}(z_o, z_a), \quad (3)$$

where z_o and z_a denote the speaker embeddings of the original and anonymized speech, respectively, extracted by the frozen semi-informed ASV. Second, a frozen Whisper ASR model [11] is further employed to extract hidden representations from both the original and anonymized speech. The linguistic objective encourages these representations to remain similar, thereby preserving the spoken content during anonymization. The linguistic loss is defined as:

$$L_{\text{linguistic}} = 1 - \text{sim}(h_o, h_a), \quad (4)$$

Table 1: *Track 1: Privacy and utility results on anonymized data compared with the original speech. Bolded and underlined text indicate the best and second-best results, respectively.*

Metric	Development			Evaluation		
	Orig.	B3	Ours (T11)	Orig.	B3	Ours (T11)
EER ($\uparrow$)	7.34	16.32	37.79	3.91	14.24	32.42
WER ($\downarrow$)	1.80	4.31	<u>4.45</u>	1.84	4.31	6.58
UAR ($\uparrow$)	69.08	<u>38.08</u>	30.39	71.06	<u>35.24</u>	31.81

where h_o and h_a are the hidden representations extracted by the frozen Whisper ASR model from the original and anonymized speech, respectively.

The overall training objective is a weighted combination of the emotion, privacy, and linguistic objectives:

$$L = \beta_1 L_{\text{emo}} + \beta_2 L_{\text{privacy}} + \beta_3 L_{\text{linguistic}}, \quad (5)$$

where β_1 , β_2 , and β_3 are hyperparameters that control the contribution of each objective during optimization.

3. Experiment

This section describes the experimental setup used to evaluate the proposed framework in protecting speaker privacy while preserving linguistic and emotional information.

3.1. Experimental Settings

3.1.1. Dataset

This paper use the VPC 2026 development and evaluation datasets. Track 1 is evaluated on the English LibriSpeech [12] and IEMOCAP [13] subsets for ASV/ASR and SER, respectively. Track 2 uses the Multilingual LibriSpeech [14] subsets covering English, French, German, and Spanish for ASV/ASR, while emotion preservation is evaluated on EmoTrack2 [4], which comprises four language-specific emotional speech corpora.

For fine-tuning, multiple datasets are combined, including: EMO-DB [15], CREMA-D [16] and RAVDESS [17] datasets with 3000 synthesize English speeches generated by CosyVoice 2.0, where reference embeddings and transcriptions are sampled from the LibriSpeech dataset. Then, samples that don't have their emotion labeled as neutral, happy, sad, or angry are removed.

3.1.2. Metrics

This paper follow the official VPC 2026 evaluation protocol [4]. Speaker privacy is measured by ASV EER, where higher values indicate stronger anonymization. Linguistic utility is evaluated by ASR WER, where lower values indicate better content preservation. Emotional utility is measured by SER UAR, where higher values indicate better emotion preservation. For Track 2, language-wise EER and WER are averaged across French, Spanish, English, and German.

3.1.3. Baseline

The proposed method is compared with two systems from the VPC 2026. For track 1, the B3 model encode speech feature using Global Style Token, anonymized data using GAN, and synthesize data using FastSpeech2 [18] framework. For track 2, the BM3 model is used. Their systems are very close to the

proposed method, which serve as good baselines for comparing speaker privacy, content preservation, and emotion-related utility.

3.1.4. Implementation

One system is evaluated for Track 1 (T11) and two systems are evaluated for Track 2 (T21 and T22). T11 and T21 use a fixed reference speaker embedding for all utterances, whereas T22 selects a different reference speaker embedding for each utterance using the Farthest Point Sampling (FPS) algorithm (Section 2.1). Correspondingly, T11 and T21 use the pitch and duration sequences of the selected reference utterance for prosody normalization, while T22 uses the average pitch and duration across all reference utterances. Finally, prosody normalization is applied to all utterances in T11 but only to neutral utterances in T21 and T22. Owing to the limited challenge timeline, only the fixed-reference strategy was evaluated for Track 1. For our submissions to the Challenge's organizer, system T11, T21 and T22 are mapped to anonymization suffix `_BM3_modify_feats8`, `_BM3_modify_final_6` and `_BM3_modify_final_3`, respectively.

3.1.5. Hyperparameter settings

The number of training steps are 10000, with 500 warm up steps, a learning rate set to 10^{-5} and batch size 24. For loss functions, β_1 , β_2 and β_3 are set to 0.05 and 0.1 and 0.1, respectively.

3.2. Experimental Results

3.2.1. Main Results

Tables 1 and 2 summarize the results for Tracks 1 and 2, respectively. Overall, the proposed systems consistently outperform the corresponding official baseline, BM3, in terms of privacy protection. On Track 1, T11 achieves the highest EER, improving the evaluation-set EER by nearly 13 percentage points over BM3. On Track 2, both T21 and T22 outperform BM3 across all four languages. Since BM3 already achieves high EER for French, German, and Spanish, improvements in these languages are relatively modest, whereas English remains substantially more challenging and benefits the most from the proposed framework, with T21 improving the evaluation-set EER by more than 15 percentage points over BM3. A paired t-test between T21 and T22 against BM3 were conducted the observed EER improvements are statistically significant for both systems ($p < 0.001$).

T21 consistently achieves higher EER than T22. By mapping all utterances to a single target speaker embedding, T21 causes anonymized speech to collapse into a common region of the embedding space, making them more difficult for the semi-informed ASV system to distinguish. In contrast, assigning different target speakers preserves greater inter-speaker separability, resulting in weaker privacy protection.

Regarding speech utility, the ASR-guided loss enables both systems to maintain competitive WER compared with BM3. T21 achieves the lowest WER on several development and evaluation sets, whereas T22 suffers higher WER on the non-English languages, likely because aggressive prosody normalization removes acoustic cues that are beneficial for speech recognition. T11 model apply prosody normalization to all utterances has the lowest UAR, demonstrating that suppressing prosodic information degrades emotion preservation.

Table 2: *Track 2: Privacy and utility results on anonymized data compared with the original speech for different languages. Bolded and underlined text indicates the best and second-best results, respectively.*

Language	Metric	Development				Evaluation			
		Orig.	BM3	Ours (T21)	Ours (T22)	Orig.	BM3	Ours (T21)	Ours (T22)
MLS-fr (French)	EER ($\uparrow$)	6.15	46.11	<u>47.33</u>	47.42	7.51	46.25	<u>47.81</u>	47.91
	WER ($\downarrow$)	5.80	13.77	11.09	20.19	5.13	<u>11.99</u>	13.32	17.36
MLS-es (Spanish)	EER ($\uparrow$)	8.97	46.23	48.31	<u>47.25</u>	4.86	45.92	<u>46.01</u>	47.81
	WER ($\downarrow$)	3.79	10.56	8.42	15.44	3.75	10.12	9.24	16.56
MLS-en (English)	EER ($\uparrow$)	2.24	22.02	32.45	<u>26.78</u>	4.75	20.40	32.60	<u>24.29</u>
	WER ($\downarrow$)	4.57	6.71	6.12	<u>6.69</u>	5.46	7.73	8.84	<u>7.61</u>
MLS-de (German)	EER ($\uparrow$)	3.00	46.54	48.51	<u>48.00</u>	11.39	<u>46.05</u>	47.09	43.64
	WER ($\downarrow$)	5.25	9.74	<u>8.40</u>	15.59	5.38	10.10	8.96	16.37
MLS-Avg.	EER ($\uparrow$)	5.09	40.91	44.15	<u>42.36</u>	7.13	40.57	44.98	<u>42.03</u>
	WER ($\downarrow$)	4.85	10.20	<u>8.50</u>	14.47	4.93	<u>9.98</u>	10.09	14.47
EmoTrack2	UAR ($\uparrow$)	72.09	26.24	<u>39.40</u>	38.57	78.90	24.84	<u>34.47</u>	34.08

Table 3: *Ablation Study. (-E) Replace speaker embedding selection with GAN-based anonymization, (-P) Remove Prosodic Normalization.*

Metric	Development			Evaluation		
	T21	T21-E	T21-P	T21	T21-E	T21-P
EER ($\uparrow$)	32.45	28.17	<u>30.82</u>	37.16	29.42	35.34
WER ($\downarrow$)	<u>6.12</u>	6.53	6.06	8.84	<u>8.62</u>	8.21
UAR ($\uparrow$)	<u>39.40</u>	37.57	39.54	34.47	34.52	<u>34.24</u>

3.2.2. Ablation Study

To investigate the contribution of each component in the proposed framework, an ablation study is conducted by systematically removing individual components and evaluating their impact on speaker privacy and speech utility. All variants are based on the T21 model, and are tested on MLS-en subset and EmoTrack2 dataset. Result are shown in Table 3.

The results show that speaker embedding selection has a clear impact on privacy protection, whereas the effect of Prosody Normalization is more modest. This suggests that the choice of speaker embedding may also influence the generation of speaker-discriminative prosodic cues, which can be exploited by the strong semi-informed ASV system. In addition, removing Prosody Normalization primarily improves WER, as the modification is selectively applied only to neutral speech in the original T21 system (Section 2.3).

3.2.3. Discussions

In this section, several limitations of the proposed systems are discussed. As shown in Table 2, preserving prosody for emotional utterances improves emotion retention but may also preserve speaker-discriminative information, resulting in weaker privacy protection than for neutral speech. This limitation is not reflected in the current challenge because EER and UAR are evaluated on different datasets. Future evaluations should jointly assess privacy, intelligibility, and emotion preservation on the same speech samples to better characterize the privacy-utility trade-off.

The result of system T22 depends on the diversity of the

target speaker pool. Ideally, each utterance would be assigned a unique target speaker, requiring a reference pool with as many distinct speakers as utterances. Such a large pool is usually impractical, a workaround is employed by grouping utterances by source speaker using a pretrained ASV model and select one diverse target reference per source speaker using the greedy algorithm in [19], which maximizes the minimum distance between selected embeddings. Each utterance is then randomly assigned a reference from this pool. Although this increases target-speaker diversity, the limited pool size means that some utterances may share similar target identities, which may reduce the power of anonymization compared with an ideally sized pool containing a unique reference speaker for every utterance.

Additionally, this paper investigate the applicability of the proposed method to both Track 1 (English) and Track 2 (multilingual). Since all components of BM3 are inherently multilingual, the same framework can be seamlessly applied to both tracks without any language-specific modifications. Experimental results show that the performance gains on Track 2 are smaller than those on Track 1, primarily because the baseline EER for non-English languages is already relatively high, leaving less room for improvement. Consequently, the main challenge in Track 2 is to improve the English EER while maintaining comparable privacy protection for the remaining languages.

Compared with Track 1, the evaluation-set EER in Track 2 is higher. A likely explanation is that Track 1 benefits from much more training data, enabling the semi-informed ASV model to generalize better. In terms of utility, our multi-objective fine-tuning framework preserves WER by periodically optimizing only the linguistic preservation objective, thereby minimizing the increase in WER across both tracks.

4. Conclusion

This paper proposed a multi-objective guided fine-tuning framework that jointly optimizes privacy protection, speech intelligibility, and emotion preservation; multilingual reference-speaker sampling to better conceal speaker-identity; and prosody normalization to suppress speaker-discriminative prosodic cues. Experimental results demonstrate that our solution effectively protect speaker identity while maintaining competitive speech intelligibility and emotion preservation.

5. Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers (25H01139 and 25K21245).

6. Generative AI Use Disclosure

Generative AI (ChatGPT) was used solely to check grammar and polishing manuscript. All AI-generated contents were carefully reviewed, and verified. The authors take full responsibility for the accuracy and integrity of the manuscript.

7. References

- [1] N. A. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. W. D. Evans, J. Patino, J. Bonastre, P. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Interspeech*, 2020, pp. 1693–1697.
- [2] T. Bäckström, "Privacy in Speech Technology," *Proc. IEEE*, vol. 113, no. 7, pp. 668–692, 2025.
- [3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The voiceprivacy 2020 challenge evaluation plan," 2022. [Online]. Available: <https://arxiv.org/abs/2205.07123>
- [4] X. Miao, N. Tomashenko, R. Arefeen, S. Meyer, M. Panariello, X. Wang, E. Vincent, N. Evans, J. Yamagishi, and M. Todisco, "The VoicePrivacy 2026 Challenge Evaluation Plan," Tech. Rep. [Online]. Available: <https://hal.science/hal-05561895>
- [5] 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.
- [6] F. Lux, S. Meyer, L. Behringer, F. Zalkow, P. Do, M. Coler, E. A. P. Habets, and N. T. Vu, "Meta Learning Text-to-Speech Synthesis in over 7000 Languages," in *Interspeech 2024*, 2024, pp. 4958–4962.
- [7] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first voiceprivacy attacker challenge," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025*, 2025, pp. 1–2.
- [8] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [9] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.
- [10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.
- [11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.
- [13] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [14] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech*, 2020, pp. 2757–2761.
- [15] F. Burkhardt, "A database of german emotional speech," 2000.
- [16] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [17] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
- [18] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=piLPYqxtWuA>
- [19] C. R. Qi, L. Yi, H. Su, and L. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/d8bf84be3800d12f74d8b05e9b89836f-Paper.pdf