

VoxCPM2 Voice-Design Anonymization for VoicePrivacy 2026: English and Multilingual Systems

Yibo Duan^{1,*}, Xuyang Xing^{1,*}, Zihan Chai^{1,3}, Maquan Yang¹, Shengchen Li³, Sheng Zhao^{1,2,**}

¹VOICECODE TECHNOLOGY PTE. LTD., Singapore

²SHANGHAI LONGYUAN TECHNOLOGY CO., LTD., China

³XI'AN JIAOTONG-LIVERPOOL UNIVERSITY, Suzhou, China

duanyibo@lyxxkj.com.cn, xingxuyang@lyxxkj.com.cn, Zihan.Chai23@student.xjtlu.edu.cn,
yangmaquan@lyxxkj.com.cn, shengchen.li@xjtlu.edu.cn, zhaosheng@lyxxkj.com.cn

Abstract

We present team-Voicecode's systems for VoicePrivacy Challenge 2026 Tracks 1 (English) and 2 (multilingual). Both use an ASR-TTS resynthesis pipeline: Whisper large-v3 transcribes the source, coarse delivery attributes are estimated, and fixed pretrained VoxCPM2 synthesizes speech from the transcript and a natural-language voice-design prompt. No source-speaker acoustic embedding or reference audio is provided to the TTS model. Track 1 uses emotion2vec-guided prompting and ASR/SER-aware candidate selection; Track 2 uses language-aware prompts and an ASR-gated fallback. On the updated official Track 1 test evaluation, the system obtains 43.43% semi-informed EER, 1.29% WER, and 59.35% UAR. On Track 2, it obtains four-language averages of 31.15% semi-informed EER, 10.61% WER, and 41.53% UAR.

Index Terms: voice anonymization, text-to-speech, prompt engineering, multilingual, speech emotion recognition

1. Introduction

The VoicePrivacy Challenge promotes speaker anonymization systems that conceal voice identity while preserving linguistic and paralinguistic attributes [1]. VPC 2026 introduces two independent tracks: Track 1 targets English evaluation on LibriSpeech and IEMOCAP, while Track 2 extends to a multilingual setting covering English, German, Spanish, and French from Multilingual LibriSpeech (MLS) and language-matched emotion corpora. Both tracks adopt a strengthened semi-informed attacker setting, where the attacker retraining an Automatic Speaker Verification (ASV) model on anonymized training data produced by the same anonymization system family as the evaluation sets.

Prior anonymization systems span two paradigms. Voice conversion (VC) approaches modify speaker embeddings or acoustic features but can retain residual speaker-characteristic patterns exploitable by an adapted semi-informed attacker [2, 3]. Cascaded Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems [4, 5] transcribe and resynthesize speech to reduce the direct transfer of source-speaker characteristics, but often struggle to preserve prosody and emotion, while fixed target speaker pools limit voice diversity. Prompt-based controllable TTS models, such as VoxCPM2 [6] and PromptTTS 2 [7], specify voice identity through natural-language text descriptions (voice-design mode), enabling synthesis without reference audio or source-speaker acoustic embeddings.

We extend the ASR-TTS paradigm by replacing speaker embedding conditioning with natural-language voice-design prompts, supported by dedicated emotion (Track 1) and language-fidelity (Track 2) frontends. Our contributions are:

1. A unified VoxCPM2 voice-design anonymization framework operating with fixed pretrained weights and without directly conditioning synthesis on source-speaker acoustic embeddings.
2. An emotion2vec-guided prompt construction and SER-aware candidate selection for Track 1, preserving emotional expressiveness on IEMOCAP evaluation.
3. Language-aware prompt profiles and an ASR-gated conservative fallback for Track 2, reducing cross-lingual transcript drift across German, Spanish, and French.
4. Strong privacy protection against the updated official semi-informed evaluation attack, achieving the highest Track 1 EER among ranked systems and competitive multilingual privacy-utility trade-offs on Track 2.

2. Shared System Architecture

2.1. Core Pipeline

Each utterance is processed in five stages: (1) a Whisper large-v3 ASR frontend [8] (implemented via Hugging Face Transformers [9]) predicts the reference transcript from source audio; (2) delivery analysis and gender classification (using a pretrained ECAPA-TDNN binary gender classifier [10]) estimate speaking rate, energy, and speaker gender; (3) a natural-language VoxCPM2 voice-design prompt is assembled, specifying an anonymous target voice with an explicit gender flip (male source $\rightarrow$ female target voice and vice versa), delivery style, and text-fidelity constraints; (4) VoxCPM2 synthesizes anonymized speech candidates from the combined prompt and reference transcript; and (5) the output waveform is resampled to 16 kHz mono 16-bit PCM WAV. No source-speaker embedding enters the synthesis path. VoxCPM2 is used in inference-only mode with pretrained weights from `openbmb/VoxCPM2`; large-scale batch inference uses `Nano-vLLM`.

2.2. VoxCPM2 Voice-Design Mode

VoxCPM2 [6] is a 2B-parameter tokenizer-free multilingual TTS model built on a MiniCPM backbone [11]. It employs a continuous diffusion-autoregressive architecture that synthesizes 48 kHz waveforms directly from text and speech representations without discrete acoustic tokenization. In voice-design mode, the intended target-voice characteristics are spec-

*These authors contributed equally.

**indicates the corresponding author.

ified through a natural-language prompt; no reference audio is required. Because no acoustic embedding or spectral feature from the source speaker enters the synthesis path, the system avoids direct conditioning on the original acoustic timbre and thereby reduces one potential channel of speaker-information transfer. This design does not imply that all speaker-correlated information is eliminated.

2.3. Base Prompt Template and Attribute Dissemination

The prompt structure is shared across both tracks, while its language and style fields are instantiated according to the track and input attributes. For Track 1, the target-speaker field uses one of two fixed descriptions under the explicit gender-flip rule: a source utterance classified as male is assigned the fixed phrase “An adult female English speaker with a natural clear voice”, whereas a source utterance classified as female is assigned the fixed phrase “An adult male English speaker with a natural clear voice”. Thus, the description is selected from a fixed inventory rather than generated freely for each utterance. One complete Track 1 prompt instantiation is:

An adult female English speaker with a natural clear voice, angry and tense, with firm articulation, speaks at a fast pace, with energetic delivery and natural intonation, clear articulation, studio-quality recording, distinct from the original speaker. Read the transcript exactly as written; do not add, skip, repeat, paraphrase, or translate any words.

The prompt separates identity control from content and prosodic delivery. Speaking rate is computed from the Whisper-predicted input transcript and source duration as

$$r = \frac{N_{\text{words}}}{T_{\text{audio}}}, \quad (1)$$

where N_{words} is the transcript word count and T_{audio} is the utterance duration in seconds. Using the 30th- and 70th-percentile thresholds from the submitted Track 1 prompt records, $r \leq 1.7144$ words/s is labeled *slow*, $r \geq 2.8250$ words/s is labeled *fast*, and intermediate values are labeled *moderate*. Energy is calculated from the root-mean-square (RMS) amplitude of the mono source waveform:

$$e_{\text{dBFS}} = 20 \log_{10}(\max\{\text{RMS}, 10^{-8}\}). \quad (2)$$

Values $e_{\text{dBFS}} \leq -40.5606$ are labeled *soft*, values $e_{\text{dBFS}} \geq -30.7197$ are labeled *energetic*, and intermediate values are labeled *balanced*. For emotion-aware prompting, a sad style deterministically uses *slow* pace and *soft* delivery, while a neutral style uses *balanced* delivery. To limit direct transfer of speaker-correlated acoustic detail, only these discrete *coarse textual categories* are passed to VoxCPM2 rather than continuous *source-derived acoustic features* (such as frame-level F0 contours or bottleneck embeddings). This low-bandwidth mapping restricts the amount of source-derived information supplied to VoxCPM2 while retaining broad prosodic guidance for naturalness and intelligibility; it does not guarantee complete removal of speaker information.

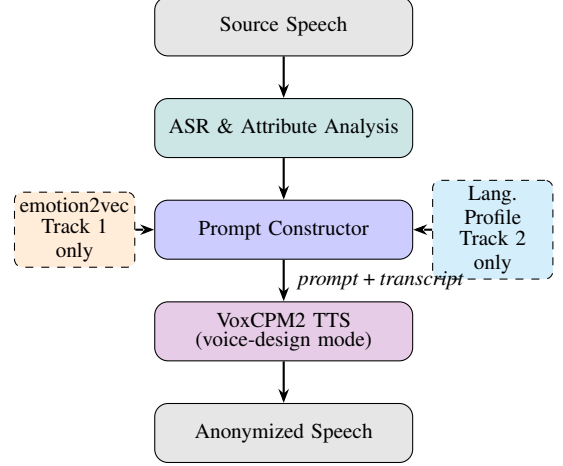


Figure 1: *VoxCPM2 voice-design anonymization pipeline shared across both tracks. No source-speaker embedding is passed to VoxCPM2. Dashed boxes indicate track-specific frontends (emotion2vec for Track 1; language-aware prompt profiles for Track 2).*

2.4. Post-Attacker Data Preparation

Both tracks anonymize the full training corpus with the same VoxCPM2 pipeline used for enrollment and trial sets, so that the semi-informed attacker training distribution more closely matches the evaluation distribution. This reduces the risk of overstating privacy because of cross-family distribution mismatches. The anonymized training data preserves standard directory structures, utterance IDs, and metadata conventions, enabling downstream ASV fine-tuning with standard recipes. Only the audio waveforms are replaced by VoxCPM2 resynthesis; transcript supervision, trial lists, and enrollment definitions remain unchanged.

3. Track 1: English System

3.1. Emotion2vec Frontend

Track 1 includes an emotion prediction frontend using `emotion2vec_plus_large` [12], a 300M-parameter SER model predicting one of nine emotion classes. Utterance-level emotion prediction is performed across all input utterances as a general preprocessing step to guide prompt construction. We apply the deterministic mapping in Table 1 to collapse the predictions into a four-class IEMOCAP-compatible label used solely as a prompt control variable. The mapping is a heuristic designed to match the four-class evaluation protocol and reduce label fragmentation. Among labels without an exact match, *disgusted* is grouped with *angry* because both are represented by a tense, negative delivery prompt, whereas *fearful* is grouped with *sad* because the available prompt inventory represents both using a more subdued negative-affect style rather than an energetic positive style. *Surprised* is valence-ambiguous and has no stable counterpart among the four target classes, so it is assigned to neutral; *other* and *unknown* are also mapped to neutral to avoid imposing an unsupported expressive style when the frontend prediction is indeterminate. These choices provide a deterministic four-class control interface; they neither imply semantic equivalence between the original categories nor establish that this heuristic is an optimal emotion taxonomy. Only

Table 1: *emotion2vec* to IEMOCAP class mapping.

emotion2vec class	IEMOCAP class
angry, disgusted	angry
fearful, sad	sad
happy	happy
neutral, other, surprised, unknown	neutral

Table 2: *Emotion to prompt style mapping (Track 1)*.

Emotion	Prompt style description
neutral	neutral and conversational
sad	subdued, gentle, low-energy, slower-paced
angry	tense, firm, energetic, but not distorted
happy	warm, bright, expressive, positive

the discrete class label is passed to the prompt builder; no continuous hidden-state features are forwarded to VoxCPM2. This low-bandwidth discrete mapping filters out fine-grained classifier variation and avoids forwarding continuous classifier scores or hidden states while aligning with the IEMOCAP evaluation protocol.

The mapped label is translated into the expressive prose style descriptors shown in Table 2. The descriptors were chosen as broad, interpretable summaries of affective delivery rather than fine-grained prosodic targets: neutral uses an unmarked conversational style, sad emphasizes lower energy and slower pace, angry emphasizes firmness and energy, and happy emphasizes warmth and brightness. This wording gives VoxCPM2 coarse and distinguishable style instructions while avoiding source-derived pitch contours, durations, or continuous emotion scores that could convey finer speaker-correlated detail. The deterministic prompt mapping was not learned or tuned from evaluation labels and is presented as an engineering design choice rather than an empirically optimal mapping.

3.2. Candidate Selection

For IEMOCAP utterances, multiple VoxCPM2 candidates are generated per utterance with varied decoding parameters. The reference transcript (predicted by the Whisper ASR frontend from original speech) is stored. During candidate selection, each generated candidate audio is transcribed by the ASR frontend to produce a hypothesis transcript. An **ASR gate** evaluates candidate transcripts against the reference transcript using character/word error rates. An **SER gate** scores candidate audio using an SER classifier against the target predicted emotion label. A weighted selection criterion balancing transcript fidelity and emotion match selects the final output (**SER-aware selection**).

3.3. Post-Attacker ASV Fine-Tuning

The semi-informed post-attacker ASV model is produced by fine-tuning the baseline WavLM+ECAPA-TDNN model [13, 14] on anonymized `train-clean-360` following challenge scripts with standard hyperparameters (Table 3).

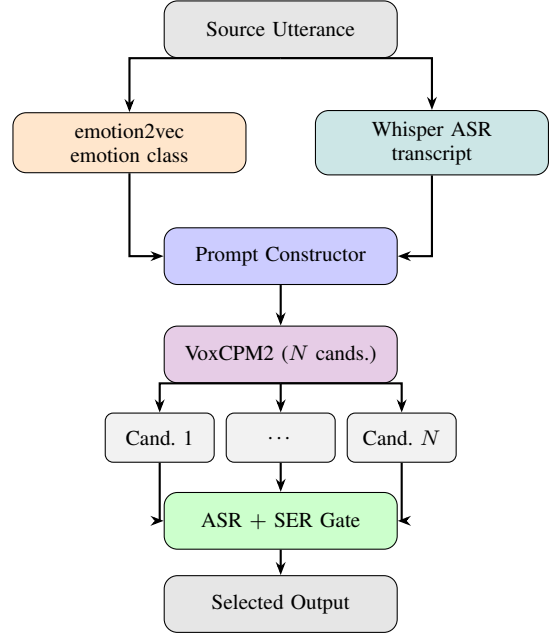


Figure 2: *Track 1 emotion-aware candidate selection pipeline. emotion2vec predicts the emotion class to guide prompt construction. VoxCPM2 generates N candidates; the weighted ASR+SER gate selects the output balancing transcript fidelity and emotion match.*

Table 3: *Track 1 post-attacker ASV fine-tuning hyperparameters.*

Hyperparameter	Value
Epochs	2
Batch size	32
Optimizer	AdamW
Learning rate	0.01
Sample rate	16 kHz

4. Track 2: Multilingual System

4.1. Multilingual Coverage and Submitted Directories

Track 2 anonymizes speech in English, German, Spanish, and French from MLS [15] using the same VoxCPM2 backbone. Submitted directories cover evaluation sets (`{lang}_dev`, `{lang}_test`, `emodata-track2`) and training sets (`train_english`, `train_german`, `train_spanish`, `train_french`).

4.2. Language-Aware Prompt Profiles

Cross-lingual transcript drift is a primary challenge in multilingual synthesis. VoxCPM2 can exhibit elevated WER for non-English languages due to specific language traits: German (compound word structures and complex morphological inflections), Spanish (phoneme-grapheme mapping differences), and French (liaison and elision). To mitigate drift, Track 2 evaluates three language-aware prompt profiles per non-English utterance:

- **asr_lang_simple**: concise language instruction with minimal constraints.

Table 4: Quick-validation WER (%) on 30 utterances per language (Whisper large-v3).

Language	Standard	Lang-prompt	Fallback
en	31.504	31.723	28.540
de	42.356	39.864	37.599
es	29.097	20.903	19.983
fr	28.808	26.632	23.499
avg	32.941	29.781	27.405

Table 5: Track 1 test system comparison (%). EER is the semi-informed EER reported by the organizers after the VPC 2026 evaluation update. EER and UAR are higher-is-better; WER is lower-is-better. IDs: B2=McAdams [3], B3=STTTS, B4=NAC, B5=ASR-BN VQ.

ID	EER ↑	WER ↓	UAR ↑
Original	3.41	1.84	71.06
B2	3.72	9.11	53.52
B3	14.24	4.26	37.70
B4	11.13	5.90	42.78
B5	17.05	4.53	38.09
Ours (T3)	43.43	1.29	59.35

- **asr_lang_news**: formal register in news-reading style.
- **asr_lang_dictation**: dictation-oriented style emphasizing exact target-language phonological fidelity.

4.3. ASR-Gated Selection and Conservative Fallback

For German, Spanish, and French utterances, candidates generated from the prompt profiles are transcribed with Whisper large-v3 in the target language and scored by character error rate. The candidate with the lowest ASR error is compared against the standard single-pass VoxCPM2 output, retaining whichever achieves lower error (**conservative fallback**). Table 4 shows quick-validation WER across 30 utterances per language.

5. Experimental Setup

Track 1 uses LibriSpeech [16] dev/test sets for ASV and ASR evaluation, and IEMOCAP [17] dev/test sets for SER evaluation. Track 2 uses MLS [15] for per-language ASV and ASR, and language-matched emotion corpora (Oreau, MESD, EMNS, EmoDB) for SER. Metrics include semi-informed ASV EER (↑), ASR WER (↓), and SER UAR (↑). For the camera-ready version, EER values follow the updated official VPC 2026 evaluation scripts, using unit L2 normalization along the embedding dimension and the corrected Track 1 mixed trial lists.

6. Results

6.1. Track 1 Results

Table 5 compares original speech and baseline systems against our proposed system on Track 1. On the updated official test evaluation, our system (T3, `voxcpm2_vd`) achieves a semi-informed EER of 43.43%, a WER of 1.29%, and a UAR of 59.35%. Table 6 provides the per-class SER breakdown.

Table 6: Track 1 ASR WER (%) and IEMOCAP SER UAR (%) with per-class breakdown.

Split	Sig.	WER ↓	UAR ↑	sad	angr.	happy	neut.
dev	orig.	1.802	69.080	63.629	79.777	66.948	65.965
dev	anon.	2.835	60.298	68.060	79.798	47.569	45.765
test	orig.	1.844	71.062	72.571	72.825	67.191	71.662
test	anon.	1.290	59.350	66.404	76.558	48.560	45.878

Table 7: Track 2 test system comparison (%). EER is the semi-informed EER reported by the organizers after the VPC 2026 evaluation update. Values are four-language averages for EER and WER, and four-class averages for UAR. BM1: SSL-based anonymization; BM2: Phone+GAN with modified prosody; BM3: Phone+GAN without F0/prosody extractor.

ID	EER ↑	WER ↓	UAR ↑
Original	4.25	4.93	78.90
BM1	2.10	8.53	42.63
BM2	10.62	27.74	29.04
BM3	39.65	9.98	24.84
Ours (T3)	31.15	10.61	41.53

6.2. Track 2 Results

Table 7 compares original speech, the BM1–BM3 baselines, and our proposed system on the updated official Track 2 test evaluation. On the updated official test evaluation, our system (T3, `voxcpm2_vd`) achieves four-language averages of 31.15% semi-informed EER and 10.61% ASR WER, with a four-class average SER UAR of 41.53%. Compared with BM1 and BM2, our system achieves higher EER; compared with BM3, it achieves higher UAR but lower EER and higher WER, reflecting a different multilingual privacy-utility trade-off. Table 8 breaks down semi-informed EER and WER per language on the test split.

Table 8: Track 2 updated official per-language test semi-informed EER and WER (%) for our system (T3).

Lang.	EER ↑	WER ↓
en	17.67	4.51
de	36.03	12.40
es	34.95	10.79
fr	35.93	14.74
avg	31.15	10.61

7. Discussion

7.1. Privacy under the Updated Official Evaluation

The VoxCPM2 voice-design pipeline removes the direct source-acoustic-embedding conditioning path, rather than establishing that speaker identity is completely decoupled from the output. Under the updated official VPC 2026 evaluation, Track 1 obtains 43.43% semi-informed EER, while Track 2 obtains a four-language average of 31.15% semi-informed EER. The Track 1 EER remains substantially above those of the original-speech and baseline systems reported by the organizers in Table 5, but the updated embedding normalization and corrected Track 1 mixed trial lists produce lower privacy scores than the earlier evaluation outputs. The results should therefore be inter-

puted as privacy protection against the specified updated semi-informed evaluation attack, not as evidence that all speaker information is removed under every possible attacker.

7.2. Neutral Class SER Preservation

On Track 1 IEMOCAP evaluation, neutral class accuracy shows a drop relative to expressive classes (angry, happy). This stems from three main factors: (1) VoxCPM2’s neutral prompt style (“neutral and conversational”) can introduce subtle expressive pitch variations that discriminative SER classifiers interpret as non-neutral; (2) emotion2vec predictions on subtle or flat speech can exhibit small classification errors; and (3) distribution shift between synthetic speech and IEMOCAP actor speech disproportionately impacts the neutral class boundary.

7.3. Utility and Cross-Lingual Intelligibility

The Whisper ASR frontend extracts a transcript before waveform resynthesis, limiting the synthesis input to text and coarse delivery controls while achieving strong intelligibility (1.29% Libri WER on Track 1). For Track 2, language-aware prompt profiles and ASR fallback maintain the updated official average test WER at 10.61%. German WER (12.40%) and French WER (14.74%) are higher than Spanish WER (10.79%).

8. Conclusion

We presented team-Voicecode’s VoicePrivacy 2026 submissions for Track 1 and Track 2, built on a unified VoxCPM2 voice-design anonymization framework. The system avoids direct conditioning on source acoustic embeddings, relies on natural-language voice descriptions, and incorporates track-specific emotion and language frontends. Under the updated official VPC 2026 evaluation, the system obtains 43.43% semi-informed EER, 1.29% WER, and 59.35% UAR on Track 1, and four-language averages of 31.15% semi-informed EER, 10.61% WER, and 41.53% UAR on Track 2. These results characterize privacy against the specified updated semi-informed evaluation attacks and do not establish the absence of speaker information under every possible attacker.

9. Acknowledgements

This work was supported by Basic Research Program of JiangSu Province under Grant BG2024027.

10. Generative AI Use Disclosure

OpenAI Codex was used to assist with language editing, consistency checking, and drafting revisions to improve the clarity of the manuscript. It was not used to run experiments or generate evaluation results. All authors reviewed the final manuscript and take full responsibility for its content.

11. References

- [1] N. Tomashenko, X. Miao, M. Arefeen, S. Meyer, M. Panariello, X. Wang, E. Vincent, N. Evans, J. Yamagishi, and M. Todisco, “The VoicePrivacy 2026 challenge evaluation plan,” VoicePrivacy Challenge, Tech. Rep., 2026.
- [2] M. Panariello, W. Simone, and M. Todisco, “Speaker anonymization with phone-level prosody modification,” in *Proc. Interspeech*, 2023.
- [3] P. Aichinger, “McAdams coefficient-based voice anonymization,” in *Proc. Interspeech*, 2019, pp. 2355–2359.
- [4] 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 *Proc. ICASSP*, 2023.
- [5] H. X. Li, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, “HLTCOE JHU submission to the voice privacy challenge 2024,” *arXiv:2409.08913*, 2024.
- [6] VoxCPM Team, “VoxCPM2: Tokenizer-free TTS for multilingual speech generation, creative voice design, and true-to-life cloning,” *arXiv:2606.06928*, 2026.
- [7] Y. Leng *et al.*, “PromptTTS 2: Describing and generating voices with text prompt,” in *Proc. ICLR*, 2024.
- [8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. ICML*, 2023, pp. 28 492–28 518.
- [9] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proc. EMNLP System Demonstrations*, 2020, pp. 38–45.
- [10] J. Huh *et al.*, “Voice gender classification using ECAPA-TDNN architecture,” in *Proc. Interspeech*, 2023.
- [11] S. Hu *et al.*, “MiniCPM: Unveiling the potential of small language models,” *arXiv:2404.06395*, 2024.
- [12] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” in *Findings of the 62nd Annual Meeting of the ACL*, 2024.
- [13] S. Chen *et al.*, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [14] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [15] V. Pratap *et al.*, “MLS: A large-scale multilingual dataset for speech research,” in *Proc. Interspeech*, 2020, pp. 2757–2761.
- [16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [17] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.