

Two-Stage Speaker Anonymization with Emotion-Conditioned Resynthesis and Mild McAdams Post-Processing: A CosyVoice2 System for the VoicePrivacy 2026 Challenge

Jiarui Xu¹, Zhiyu Wan^{1,**}, Sihan Xie^{1,**}

¹ShanghaiTech University, Shanghai, China

xujr2025@shanghaitech.edu.cn, wanzhy@shanghaitech.edu.cn, xiesh2024@shanghaitech.edu.cn

Abstract

Voice anonymization aims to conceal speaker identity while preserving the linguistic and paralinguistic information. The VoicePrivacy 2026 Challenge makes this further demanding by evaluating speaker protection from a stronger, domain-aware semi-informed attack, assessing how well emotional states are preserved apart from linguistic contents. In this paper, we present a two-stage speaker anonymization system for Track 1 of the challenge. The two anonymization stages are preceded by a Stage 0 analysis front-end, which transcribes each utterance and assigns one of four emotion labels. Stage 1 resynthesizes the transcript with CosyVoice2 using a fixed prompt selected according to the predicted emotion. Stage 2 applies a McAdams transformation to the resynthesized waveform, aiming to further perturb residual spectral-envelope cues with minimal degradation to speech utility. Under the official semi-informed attacker, the resynthesis-only configuration achieves a test set equal error rate of 35.41%, a word error rate of 3.96%, and an emotion unweighted average recall of 68.14%, compared with 34.89%, 4.29%, and 65.97%, respectively, for the overall two-stage system, both of which surpass every Track 1 baseline. A complete ablation over the McAdams coefficient range shows that the second stage provides no measurable privacy benefit. Across the tested range, the semi-informed EER varies by approximately half a percentage point, whereas WER increases and emotion UAR decreases monotonically as the perturbation strengthens. Identity protection therefore stems primarily from the resynthesis stage, while the main remaining limitation is reduced emotion recognition accuracy for sad speech.

Index Terms: speaker anonymization, voice privacy, speech synthesis, McAdams coefficient, emotion preservation, VoicePrivacy Challenge

1. Introduction

Speech signals contain personally identifiable information, including speaker identity, and therefore fall within the scope of privacy regulations such as the General Data Protection Regulation (GDPR) [1]. Downstream tasks such as automatic speech recognition and speech emotion recognition require linguistic content and emotional state to be preserved. Voice anonymization addresses this requirement by modifying speech to conceal speaker identity while retaining linguistic content and relevant paralinguistic attributes [2].

VPC 2026 [3], the fourth edition of the VoicePrivacy Challenge, evaluates anonymization systems against a stronger, domain-aware automatic speaker verification (ASV) attacker pretrained on large-scale voice-converted (VC) speech and fine-

tuned on anonymized data. Our work targets Track 1 for English speech. In this track, equal error rate (EER) measures identity protection, word error rate (WER) measures intelligibility, and unweighted average recall (UAR) measures emotion preservation.

An automatic speech recognition and text-to-speech (ASR-TTS) anonymization pipeline first transcribes the source utterance and then generates a new waveform from the recognized text. Replacing the original waveform in this way suppresses many speaker-specific acoustic cues, but the transcript alone does not represent the emotion that should be retained. Before anonymization, the Stage 0 analysis front-end uses Whisper large-v3 [4] and emotion2vec [5] to process the source utterance in parallel. Stage 1 uses CosyVoice2 [6] to resynthesize the transcript with a fixed prompt selected according to the predicted emotion. Stage 2 applies a mild McAdams transformation to the resynthesized waveform to perturb formant-related spectral-envelope cues [7].

We evaluate whether the McAdams stage improves privacy after neural resynthesis by comparing the two-stage system with the resynthesis-only configuration and an additional configuration that uses a lower McAdams coefficient range. The semi-informed EER remains nearly unchanged across the tested ranges, while WER increases and emotion UAR decreases as lower McAdams coefficients are applied. The McAdams stage therefore provides no measurable privacy improvement and reduces speech utility.

2. Background

2.1. Task and Threat Model

VPC 2026 formulates privacy protection as a game between a user who shares anonymized utterances for downstream ASR and speech emotion recognition (SER), and an attacker who attempts to re-identify the speaker of each anonymized trial utterance [3]. A valid anonymization system must (a) output a speech waveform, (b) conceal speaker identity at the utterance level, and (c) distort neither the linguistic content nor the emotional state. The utterance-level requirement mandates that the pseudo-speaker assignment be applied independently per utterance, using an identical process across utterances that does not rely on speaker labels.

The attacker is semi-informed and has access to the anonymization system, anonymizes its own enrollment and training data to reduce mismatch, and fine-tunes a strong ASV model, pretrained on large-scale voice-converted speech, on anonymized English data. This is the strongest threat model considered in the challenge and is used to derive the official rankings. Track 1 additionally introduces cross-gender trials alongside same-gender trials, with the attacker assumed not to know whether the source

^{**}indicates the corresponding author.

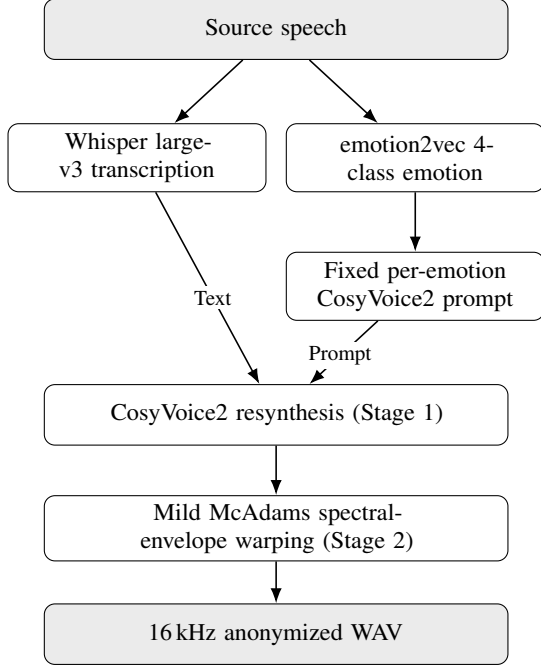


Figure 1: System pipeline. Stage 0 transcribes the source speech and predicts its emotion in parallel. Stage 1 resynthesizes the transcript using the corresponding fixed prompt, and Stage 2 applies mild McAdams post-processing to the synthesized waveform.

gender was preserved.

2.2. Metrics and Ranking

Privacy is measured by the EER of the semi-informed attacker ASV (higher is better for the user). Intelligibility is measured by the WER of an ASR model trained on original speech (lower is better), and emotion preservation by the UAR of an SER model (higher is better). Following the VPC 2024 edition, $N=4$ evaluation conditions are defined by minimum target EERs of 10%, 20%, 30% and 40%, and systems that meet a given privacy floor are ranked by utility at that level [3].

3. System Description

Our system, registered under the team name team-Chameleos, corresponds to the anonymized-data suffix `_cosyemotion.v2_mcadams.mild.v1` and comprises a Stage 0 analysis front-end followed by two anonymization stages, as depicted in Figure 1. The intermediate output of Stage 1 carries the suffix `_cosyemotion.v2`. In Stage 0, transcription and emotion recognition are computed in parallel from the source utterance. In Stage 1, the recognized text supplies the content and the predicted emotion label selects the fixed prompt for synthesis. Stage 2 post-processes the synthesized waveform. The implementation and configuration files are available at <https://github.com/AnDuinwryn/vpc2026-cosyemotion-mcadams>.

3.1. Stage 0: Analysis front-end

Each 16 kHz source utterance is transcribed with Whisper large-v3 [4]. In parallel, emotion2vec (emotion2vec-plus-large) [5]

Table 1: Mapping labels of emotion2vec.

emotion2vec label	class
angry, disgusted	ang
happy	hap
neutral, other, surprised, unknown	neu
sad, fearful	sad

Table 2: The four fixed per-emotion CosyVoice2 prompts, all drawn from a single IEMOCAP speaker (the session-2 male actor).

Class	Source utterance id	Prompt transcript
ang	Ses02F_script03_2_M047	“You’re a evil vampire and I hope to God I never see you again as long as I live.”
hap	Ses02F_impro03_M010	“No. I’m just as happy as you are. This is—”
neu	Ses02F_impro05_M000	“Um, okay, it didn’t—didn’t come out on the carousel—”
sad	Ses02F_script01_3_M023	“It’s all mixed up with so many other things. You remember overseas, I was in command of a company?”

predicts an emotion distribution, whose top label is mapped onto four coarse classes as shown in Table 1. No ground-truth transcripts, ground-truth emotion labels, or target-speaker labels are used at any point.

3.2. Stage 1: Fixed Per-emotion CosyVoice2 Control

Unlike a per-utterance reference scheme, this system binds each emotion class to a single fixed CosyVoice2 [6] prompt. All utterances with the same predicted emotion thus share one control configuration. The four prompts are drawn from IEMOCAP recordings, one for each ang/hap/neu/sad. Table 2 lists their source utterance ids and transcripts. All four prompts originate from the same speaker (the male actor of session 2) from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [8]; the effective target voice of the system is therefore a single speaker realized in four emotional styles. The ang prompt uses the instruction-controlled path (`inference.instruct2`) with the style instruction “Speak in a clearly angry voice, with tense energy, sharp emphasis, and a forceful pace”; the other three use the zero-shot path (`inference.zero_shot`) with their fixed prompt waveform and transcript. Each prompt is registered once via `add_zero_shot_spk` under a fixed cache id (`vpc2026_ang`, `_hap`, `_neu`, `_sad`), which every matching utterance reuses. All randomness is seeded deterministically from the global seed (2026), the dataset name and the utterance id, with deterministic cuDNN flags; a fallback path splits over-long texts into chunks (220 characters or 32 words) that are synthesized separately and concatenated. Stage 1 output is 16 kHz, mono, signed 16-bit PCM WAV.

3.3. Stage 2: Mild McAdams Post-processing

The synthesized waveform is post-processed with a mild McAdams spectral-envelope perturbation [7]. Operating on the linear predictive coding (LPC) spectral envelope, the McAdams transformation warps formant-like structure and thereby alters vocal-tract and timbre speaker cues, without a further ASR, SER

Table 3: *Privacy and utility performance of original speech, Stage 1, and the two McAdams configurations on the LibriSpeech and IEMOCAP development and test sets.*

Condition	Split	EER _{base} (%) $\uparrow$	EER _{semi} (%) $\uparrow$	WER (%) $\downarrow$	UAR (%) $\uparrow$
Original	dev	–	–	1.80	69.15
	test	–	–	1.84	71.06
Stage 1 only (<i>_cosyemotion_v2</i>)	dev	44.58	36.94	4.14	67.73
	test	44.22	35.41	3.96	68.14
Two-stage system (<i>_v2-mcAdams_mild_v1</i> , $\alpha \sim \text{Unif}(0.75, 0.95)$)	dev	44.00	36.57	4.55	66.09
	test	44.63	34.89	4.29	65.97
Additional McAdams $\alpha \sim \text{Unif}(0.55, 0.85)$	dev	44.24	36.79	10.07	61.23
	test	46.13	35.32	9.25	60.04

or TTS pass and without using any speaker-identity information. Within the tested range, smaller values of α produce stronger spectral-envelope warping. We use a deliberately mild setting comprising 20 LPC coefficients, a McAdams coefficient sampled from $[0.75, 0.95]$, a 20 ms analysis window with 10 ms shift, and seed 0. The purpose of this stage is to introduce a complementary, signal-level perturbation of speaker identity in addition to the neural resynthesis, at minimal cost to intelligibility and emotion preservation. Because the McAdams stage processes only the waveform, the full system remains utterance-level.

4. Experimental Setup

Data. Following the Track 1 submission protocol [3], we anonymize the LibriSpeech [9] dev and test sets for the ASV/ASR conditions, the IEMOCAP [8] dev and test sets for the SER (emotion) condition, and train-clean-360, used to fine-tune the attacker ASV. All datasets are processed by the same anonymization system, as required.

Evaluation Models. Privacy is evaluated under mixed same- and cross-gender trials using the challenge-provided WavLM-ECAPA [10, 11] attacker. We report results from the pretrained base model and from the official semi-informed model, $\text{ASV}_{\text{en}}^{\text{anon}}$. The latter is fine-tuned on anonymized train-clean-360 data and determines the official ranking. Utility is evaluated with the challenge-provided ASR and SER models using WER and UAR, respectively. All evaluation scripts are used without modification.

Computational Cost. Anonymizing the 113,800 utterances required approximately 19 h for Stage 1 on the available GPUs. Stage 2 completed in approximately 21 min on CPUs with 32 parallel jobs. The official evaluation, including fine-tuning of the semi-informed attacker, required approximately 4.2 h.

5. Results

Table 3 reports privacy and utility on the LibriSpeech and IEMOCAP development and test sets. EER_{base} is obtained from the pretrained base attacker. EER_{semi} is obtained from the official $\text{ASV}_{\text{en}}^{\text{anon}}$ attacker after fine-tuning on anonymized training data and determines the official ranking. Both EER columns use mixed-gender trials. Higher EER and UAR and lower WER indicate better performance.

Stage 1 is evaluated without McAdams post-processing. For each utterance, the two-stage system samples the McAdams coefficient α uniformly from $[0.75, 0.95]$. Because smaller values produce stronger spectral-envelope warping, we also evaluate $[0.55, 0.85]$ to examine a more pronounced transformation.

Table 4: *Comparison with the official VPC 2026 Track 1 baselines on the development and test sets.*

System	EER _{base} (%) $\uparrow$	EER _{semi} (%) $\uparrow$	WER (%) $\downarrow$	UAR (%) $\uparrow$
Original	–	–	1.80 / 1.84	69.15 / 71.06
ASR-BN + HiFi-GAN	33.15 / 28.87	24.20 / 21.28	4.91 / 4.44	40.08 / 38.25
STTTTS	28.05 / 23.18	19.14 / 16.85	4.31 / 4.31	38.08 / 35.24
McAdams	11.42 / 7.22	6.97 / 4.71	10.44 / 9.96	55.66 / 53.49
Neural audio codec	23.00 / 17.76	17.96 / 14.33	6.16 / 5.90	41.97 / 42.78
Stage 1 only	44.58 / 44.22	36.94 / 35.41	4.14 / 3.96	67.73 / 68.14
Two-stage system	44.00 / 44.63	36.57 / 34.89	4.55 / 4.29	66.09 / 65.97

Table 5: *Per-class SER accuracy and UAR (%) on IEMOCAP across the tested configurations.*

Split	Condition	sad	angry	happy	neutral	UAR
dev	Original	63.63	80.06	66.95	65.97	69.15
	Stage 1	41.61	82.25	72.15	74.92	67.73
	Two-stage, $\alpha \sim \text{Unif}(0.75, 0.95)$	32.11	81.83	74.53	75.88	66.09
	Two-stage, $\alpha \sim \text{Unif}(0.55, 0.85)$	22.59	71.73	77.70	72.89	61.23
test	Original	72.57	72.82	67.19	71.66	71.06
	Stage 1	36.96	85.03	73.04	77.52	68.14
	Two-stage, $\alpha \sim \text{Unif}(0.75, 0.95)$	28.18	84.32	73.38	78.02	65.97
	Two-stage, $\alpha \sim \text{Unif}(0.55, 0.85)$	17.52	74.99	74.53	73.14	60.04

Table 4 compares Stage 1 and the two-stage system with the four official Track 1 baselines across the privacy and utility metrics.

Privacy. Under the official semi-informed attacker, Stage 1 obtains EERs of 36.94% on the development set and 35.41% on the test set. The two-stage system obtains 36.57% and 34.89%, respectively. McAdams post-processing therefore produces no EER improvement on the development or test set.

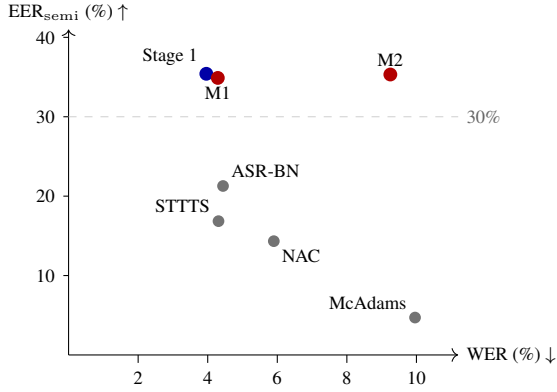
Intelligibility. WER increases from 1.80%/1.84% (original, dev/test) to 4.14%/3.96% after Stage 1 and to 4.55%/4.29% after McAdams post-processing. At a test-set WER of 4.29%, the anonymized speech remains highly intelligible, and the residual errors are attributable primarily to the Whisper front-end and its propagation through synthesis.

Emotion Preservation. Stage 1 achieves UARs of 67.73% and 68.14% on the development and test sets, respectively. With McAdams post-processing, UAR decreases to 66.09% and 65.97% for $\alpha \sim \text{Unif}(0.75, 0.95)$, and to 61.23% and 60.04% for $\alpha \sim \text{Unif}(0.55, 0.85)$.

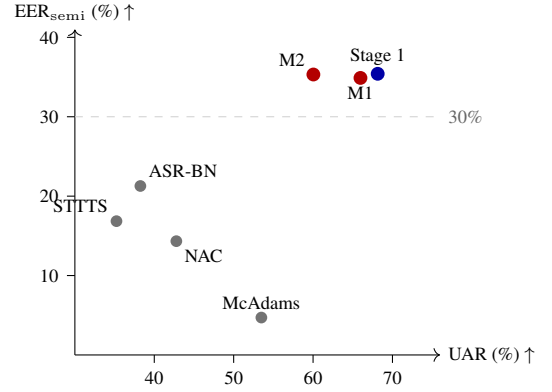
Table 5 gives the per-class SER results. The decline is concentrated in sad speech, whose accuracy decreases at each stage and continues to decrease when the McAdams coefficient range is lowered. The angry, happy and neutral classes vary less across the configurations. Table 6 shows that emotion2vec routes 94.2% of the true-sad utterances to the sad prompt. Under the same official SER, sad recall decreases from 71.6% on original speech to 37.2% after Stage 1, while neutral predictions increase from 20.0% to 55.0%. This indicates that most of the degradation in sad speech occurs during resynthesis. McAdams post-processing further increases the neutral prediction rate to 64.9% for $\text{Unif}(0.75, 0.95)$ and 70.1% for $\text{Unif}(0.55, 0.85)$.

Privacy-Utility Trade-off. The two-stage system obtains a test semi-informed EER of 34.89%, a WER of 4.29%, and a UAR of 65.97%. It therefore satisfies the 10%, 20%, and 30% minimum-EER conditions but not the 40% condition.

Figure 2 compares the three configurations with the four official baselines. Stage 1 applies no McAdams post-processing. The two post-processed configurations use coefficient ranges with means of 0.85 and 0.70. Their test semi-informed EERs



(a) privacy vs. intelligibility



(b) privacy vs. emotion preservation

Figure 2: Privacy–utility results on the LibriSpeech and IEMOCAP test sets. Gray markers denote the official Track 1 baselines, the blue marker denotes Stage 1, and red markers denote the two McAdams configurations. M1 uses $\alpha \sim \text{Unif}(0.75, 0.95)$, and M2 uses $\alpha \sim \text{Unif}(0.55, 0.85)$.

Table 6: Emotion assignments for true-sad IEMOCAP test utterances ($n=556$, %). Emotion2vec selects the synthesis prompt from the source speech. The other rows show official SER predictions for the corresponding waveforms.

Condition	→ ang	→ hap	→ neu	→ sad
Emotion2vec prompt selection	0.2	4.9	0.7	94.2
Original	1.4	7.0	20.0	71.6
Stage 1	1.8	5.9	55.0	37.2
Two-stage, $\alpha \sim \text{Unif}(0.75, 0.95)$	2.2	5.4	64.9	27.5
Two-stage, $\alpha \sim \text{Unif}(0.55, 0.85)$	5.8	7.2	70.1	16.9

are 35.41%, 34.89%, and 35.32%, with no monotonic change across the three configurations. WER rises from 3.96% to 4.29% and then to 9.25%, while UAR falls from 68.14% to 65.97% and then to 60.04%. The lower-coefficient configuration also has the highest base-attacker EER, 46.13%, but does not improve the semi-informed EER. This discrepancy confirms the importance of evaluating privacy with the adapted attacker.

6. Discussion

The ablation indicates that the privacy protection of the full system derives entirely from Stage 1 resynthesis. Although Stage 1 concentrates its target voices on four fixed prompts from a single speaker, the non-adapted EERs of both stages remain near 44% and the full system retains a semi-informed EER close to 35%. The resynthesis discards the acoustic identity of the source speaker.

With the Stage 1 figure now available, the marginal effect of the McAdams stage can be quantified, and it is not distinguishable from zero. A plausible reading is that the two mechanisms are not complementary but redundant. The McAdams transform removes speaker information carried by the vocal-tract spectral envelope; applied to natural speech it is effective, and the corresponding baseline drives the semi-informed EER from near-perfect identification down to 4.71%. Applied to CosyVoice2 output, however, there is little of the source speaker’s envelope left to remove because the resynthesis has already replaced it with that of the prompt speaker. The perturbation then acts only on the synthetic voice, degrading the signal without fur-

ther concealing the source identity. If this reading is correct, signal-level post-processing should be expected to add little to any recognition–resynthesis pipeline, and its cost is best spent elsewhere.

The utility cost is not negligible. Relative to Stage 1, the mild setting increases WER by 0.3–0.4 percentage points and reduces UAR by 1.6–2.2 points. For the lower-range configuration, the corresponding changes reach 5.3 and 8.1 points. The principal weakness, apparent in Table 5, concerns the sad class, whose accuracy decreases markedly relative to the original while the remaining classes are preserved or improved. Table 6 shows that these utterances are indeed re-labelled as neutral, and that both stages contribute. The fixed sad prompt accounts for most of the shift and the McAdams perturbation for a further ten points. Possible remedies, both straightforward under the fixed-prompt design, include replacing the sad prompt and synthesizing the sad class through the instruction-controlled path. Finally, the cross-gender condition and the VC-pretrained attacker remain the most demanding aspects of the privacy evaluation.

7. Conclusion

We presented a two-stage, utterance-level speaker anonymization system for VPC 2026 Track 1. The system applies mild McAdams spectral-envelope post-processing to emotion-conditioned CosyVoice2 resynthesis with fixed per-emotion prompts. Under the official semi-informed attacker, the two-stage system obtains a test EER of 34.89%, a WER of 4.29%, and an emotion UAR of 65.97%. It satisfies the 10%, 20%, and 30% minimum-EER conditions.

Across the tested McAdams coefficient ranges, semi-informed EER changes by approximately half a percentage point, while WER increases and UAR decreases as the coefficients are lowered. The measured privacy protection is therefore provided by the resynthesis stage, and the McAdams post-processing does not improve the official privacy metric. The main limitation is the reduced emotion recognition accuracy for sad speech, which is caused primarily by the fixed sad prompt and is further reduced by McAdams post-processing. Future work will investigate more diverse emotion prompts and prompt-selection strategies.

8. Generative AI Use Disclosure

Anthropic Claude Opus 4.8 and OpenAI gpt-5.6-sol were used for writing refinements only.

9. References

- [1] A. Nautsch *et al.*, “Preserving privacy in speaker and speech characterisation,” *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.
- [2] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. No  , and M. Todisco, “Introducing the VoicePrivacy initiative,” in *Proc. Interspeech*, 2020, pp. 1693–1697.
- [3] The VoicePrivacy Challenge organizers, “The VoicePrivacy 2026 challenge evaluation plan,” 2026, evaluation plan v1.2, https://www.voiceprivacychallenge.org/vp2026/docs/VPC_2026_march17_v2.pdf.
- [4] 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*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [5] 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 Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024, pp. 15 747–15 760.
- [6] Z. Du *et al.*, “CosyVoice 2: Scalable streaming speech synthesis with large language models,” *arXiv preprint arXiv:2412.10117*, 2024. [Online]. Available: <https://arxiv.org/abs/2412.10117>
- [7] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, “Speaker anonymisation using the McAdams coefficient,” in *Proc. Interspeech*, 2021, pp. 1099–1103.
- [8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [9] 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.
- [10] 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, J. Wu, 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.
- [11] 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.