

Enhanced Selection-Based Voice Anonymization via Soft-HuBERT and Statistical Pitch Transformation

Yilin Han¹, Ruixi Jiang¹, Tianyu Song¹, Lijie Liu¹, Qihang Lu¹, Wenxin Fu¹, Yingming Gao¹, Ya Li^{1,**}

¹ Beijing University of Posts and Telecommunications, Beijing, China

hyl0927@bupt.edu.cn, jrx725058@bupt.edu.cn, 2097278296@qq.com, 920769497@qq.com,
luqihang@bupt.edu.cn, fuwenxin2003@bupt.edu.cn, yingming.gao@bupt.edu.cn,
yli01@bupt.edu.cn

Abstract

Voice anonymization systems must conceal speaker identity while preserving linguistic content and paralinguistic attributes. We present an enhanced selection-based voice-anonymization pipeline built on the VoicePrivacy Challenge BM1 framework. Soft-HuBERT extracts linguistic-content features, and ECAPA-TDNN extracts speaker embeddings. We introduce a log-domain Z-score F0 transformation toward a target register and speaker pooling with Gaussian perturbation, comparing weighted farthest and uniform random selection. We evaluate the system on four Multilingual LibriSpeech languages under the VoicePrivacy 2026 protocol. We conduct speaker verification experiments to assess privacy protection, speech recognition experiments to measure content preservation, and emotion recognition experiments on the EmoTrack2 benchmark. Results show substantial privacy gains over BM1, with a clear trade-off in ASR and emotion-recognition performance.

Index Terms: voice anonymization, speaker privacy, Soft-HuBERT, ECAPA-TDNN, pitch transformation

1. Introduction

Speech is a natural and information-rich form of human communication. Beyond linguistic content, speech carries paralinguistic attributes—emotion, intonation, speaking style—and speaker-specific traits such as age, gender, accent and vocal tract physiology. These cues form a voiceprint for biometric identification, making speech data a privacy concern under GDPR. Voice anonymization aims to suppress speaker-identifiable information while preserving linguistic and paralinguistic content for downstream applications.

The VoicePrivacy Challenge, abbreviated here as VPC, has driven progress since 2020 [1, 2, 3, 4]. VPC 2026 introduces two tracks: Track 1 continues the English-only task with stronger attackers, while Track 2 extends to a multilingual setting covering four languages. This work targets Track 2 exclusively; the English-only Track 1 is outside the scope of this work. Systems are evaluated along three axes: privacy via the equal error rate, or EER, of an automatic speaker verification, or ASV, attacker; content preservation via the word error rate, or WER, of an automatic speech recognition, or ASR, model; and emotion preservation via the unweighted average recall, or UAR, of a speech emotion recognition, or SER, model.

The privacy metric assumes a semi-informed adversary, the strongest attacker defined by the challenge: it knows the anonymization system, uses it to anonymize the designated training data, and fine-tunes the ASV model on that anonymized

training speech. The fine-tuned attacker is then evaluated on anonymized enrollment and trial utterances. A system that displaces speaker identities into a fixed region of the embedding space offers little protection, because the retrained attacker learns that displacement. Track 2 makes this challenge more difficult. The anonymizer must remain language-independent, whereas the attacker is retrained separately for each language and can exploit language-specific regularities. Thus, privacy protection must rely on properties that remain effective after adaptation.

Existing approaches fall into several categories. Signal-processing methods such as the McAdams coefficient baseline [5] require no training data but offer limited protection against modern attackers. Neural voice conversion approaches [6, 7, 8] disentangle speech into speaker, content and prosodic representations, anonymize the speaker component and resynthesize the waveform. The text-to-speech pipeline [9] reduces linguistic content to phonetic transcriptions, and the neural audio codec approach [10] uses semantic and acoustic tokens with a language model. TTS-based pipelines suffer from information loss through textual bottlenecks, whereas the SSL-based BM1 [8] offers a promising alternative: it replaces the language-dependent ASR acoustic model with a self-supervised HuBERT soft content encoder [11, 12], achieving language independence and strong utility.

Yet BM1’s privacy protection remains weak. Its semi-informed Anon-vs-Anon EER averages 1.64% across the four development languages—below 2% for English, German and Spanish, and 2.10% for French. We hypothesize that this weakness is related to two properties: the fundamental frequency, denoted F0, is a strong speaker-identifiable cue that is passed to the vocoder unchanged; and the anonymous x-vector is generated from a similarity-defined subset of the speaker pool, so its construction remains explicitly conditioned on the source x-vector.

To address these limitations, we develop an enhanced selection-based pipeline based on BM1. Our contributions are fourfold. First, we introduce a log-domain Z-score transformation for F0 that remaps pitch distributions toward a target register under an explicit strength factor, thereby disrupting a key speaker-identifiable cue. Second, we introduce a speaker-pooling stage with additive Gaussian perturbation, instantiated with two selection rules—temperature-scaled weighted selection from the farthest candidates and query-independent uniform random selection—to compare source-conditioned and source-independent pooling policies. Third, we conduct a complete Track 2 evaluation for four MLS languages, reporting *both* development and test partitions with comparisons against the official BM1, BM2, and BM3 baselines [4]. Finally, we con-

**indicates the corresponding author.

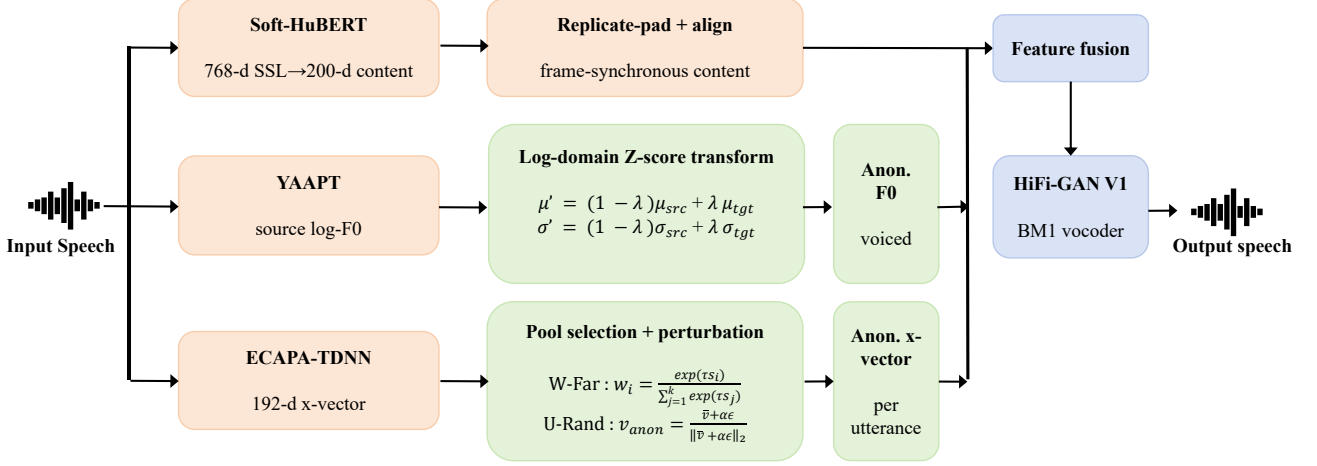


Figure 1: Overall framework of the proposed selection-based voice-anonymization system. The input waveform is decomposed into content, F0, and speaker-embedding streams. The content stream is aligned and retained, whereas the F0 and speaker-embedding streams are anonymized before concatenation and HiFi-GAN synthesis.

duct a controlled comparison of the two selection rules to examine whether farthest-neighbor selection provides a benefit.

2. Methodology

2.1. The BM1 baseline

BM1 [4, 8] is the SSL-based baseline of the VoicePrivacy Challenge. Unlike TTS-based pipelines that convert speech to discrete text and back, BM1 operates directly on continuous SSL representations, avoiding the information loss inherent in textual bottlenecks. The input waveform is decomposed into three parallel streams. A HuBERT-based soft content encoder [11, 12] extracts 200-dimensional linguistic features at a frame rate of 50 Hz. The YAAPT algorithm [13] extracts an F0 contour. An ECAPA-TDNN speaker encoder [14] extracts a 192-dimensional utterance-level x-vector. The source x-vector is replaced with an anonymous x-vector drawn from a pool constructed from LibriTTS-train-other-500 [15]. The three streams are then concatenated into a 393-dimensional representation and fed to a HiFi-GAN vocoder [16].

We choose BM1 as our foundation for two reasons. First, its SSL-based content encoding is language-independent, making it directly applicable to multilingual Track 2 without language-specific ASR models. Second, its direct feature-to-waveform synthesis preserves finer acoustic detail than cascaded TTS pipelines, which matters for emotion preservation. We retain the content encoder, pitch extractor, and vocoder unchanged, modifying only the F0 and speaker-embedding streams. Figure 1 distinguishes the inherited components from the components introduced in this work.

2.2. Feature extraction

We retain BM1’s three-stream extraction. The input audio is resampled to 16 kHz. The Soft-HuBERT predictor extracts 768-dimensional SSL features, which are projected to a 200-dimensional continuous space by cosine similarity against 200 learned prototype vectors. The ECAPA-TDNN model produces a 192-dimensional x-vector. YAAPT extracts F0 with a 10 ms frame step; only voiced frames, defined as frames with F0 above 50 Hz, are used for the statistical computation. For temporal

consistency, content features are replicate-padded, and all three streams are explicitly aligned to the same number of frames before concatenation.

2.3. Statistical F0 transformation

Fundamental frequency is one of the strongest speaker-identifiable cues, and BM1 passes it through unchanged, allowing an ASV attacker to exploit pitch patterns for re-identification. Our transformation operates in the logarithmic domain, where log-F0 is often approximately Gaussian. For each utterance we compute the mean μ_{src} and standard deviation σ_{src} of the log-F0 over voiced frames.

Rather than mapping each utterance directly to the target moments, we interpolate between the source and target moments using a strength factor λ , which controls the extent of the displacement:

$$\begin{aligned}\mu' &= (1 - \lambda) \mu_{src} + \lambda \mu_{tgt}, \\ \sigma' &= (1 - \lambda) \sigma_{src} + \lambda \sigma_{tgt},\end{aligned}\quad (1)$$

We then apply the Z-score remapping:

$$\log f_{anon} = \frac{\log f_{src} - \mu_{src}}{\sigma_{src}} \sigma' + \mu'. \quad (2)$$

Unvoiced frames are left unchanged. After exponentiating the transformed log-F0 values, we clip the resulting F0 to the range 0 to 600 Hz to keep it within a plausible human-speech range. This largely preserves the relative pitch contour within an utterance while shifting its absolute register.

Two design choices deserve comment. Operating in the log domain rather than in the linear-frequency domain matches the approximately logarithmic relation between fundamental frequency and perceived pitch: a constant offset in $\log f$ corresponds to a constant interval shift regardless of the speaker’s register, whereas an equivalent linear shift would distort low-register voices much more than high-register voices. The interpolation in Eq. (1) makes the transformation partial rather than forcing a complete remapping. At $\lambda = 1$, every utterance would be assigned the same target mean and variance, removing speaker-dependent register differences and suppressing

moment-level variation that may carry prosodic and emotional information; at $\lambda < 1$, a fraction of the original moments is retained, so λ provides a single continuous control over how much pitch identity is altered.

The target moments are fixed constants rather than corpus-estimated quantities. We take two reference registers of 120 Hz and 210 Hz for typical male and female voices, respectively; with F0 measured in hertz, their natural logarithms are $\ln(120) \approx 4.787$ and $\ln(210) \approx 5.347$. We set a single gender-neutral target near their midpoint: $\mu_{\text{tgt}} = 5.06$, corresponding to a median F0 of roughly 158 Hz, with the target standard deviation fixed at $\sigma_{\text{tgt}} = 0.20$. The same gender-neutral target is applied to every utterance without gender detection or gender-conditioned target pools. The target moments and the strength factor $\lambda = 0.8$ were fixed *a priori*: the former from the two reference registers above and the latter to retain a partial transformation. No parameter sweep was performed.

2.4. Speaker pooling with perturbation

BM1 [8] generates an anonymous x-vector by identifying the 200 pool vectors farthest from the source x-vector, randomly selecting 100 of them, and computing their unweighted average. We add a stochastic perturbation, and instantiate the selection stage with two alternative rules, compared in Section 3.4.

Weighted farthest selection (W-Far). For a source x-vector $\mathbf{v}_q$ we rank the pool by cosine similarity, keep the $N = 50$ *least* similar candidates, and randomly sample $k = 10$ of them. Let s_i denote the cosine similarity between $\mathbf{v}_q$ and the i -th sample. We combine the samples using temperature-scaled softmax weights:

$$w_i = \frac{\exp(\tau s_i)}{\sum_{j=1}^k \exp(\tau s_j)}, \quad \tau = 5, \quad (3)$$

which emphasizes the less extreme members of the farthest-selected set.

Uniform random selection (U-Rand). We discard the similarity ranking entirely and draw $k = 10$ entries uniformly at random from the whole pool, independently of the query, with uniform weights $w_i = 1/k$. Under this rule, the candidate-set size N and temperature τ have no effect.

In both cases, we define the weighted mean as $\bar{\mathbf{v}} = \sum_{i=1}^k w_i \mathbf{v}_i$, perturb it with zero-mean Gaussian noise with per-dimension scaling, and then renormalize it:

$$\mathbf{v}_{\text{anon}} = \frac{\bar{\mathbf{v}} + \alpha \boldsymbol{\epsilon}}{\|\bar{\mathbf{v}} + \alpha \boldsymbol{\epsilon}\|_2}, \quad \epsilon_d \sim \mathcal{N}(0, \sigma_{\text{local},d}^2), \quad (4)$$

with $\alpha = 0.01$. Here $\sigma_{\text{local},d}$ is the standard deviation of dimension d computed over the $k=10$ *selected* vectors—not over the $N=50$ candidate set—so the perturbation adapts to how tightly that particular draw clusters. The pool vectors enter the average and this standard deviation in their raw form; L2-normalized copies are used only for the similarity ranking, and unit norm is restored by Eq. (4). The result is layer-normalized before entering the vocoder, as in BM1. We set $\tau = 5$ and $\alpha = 0.01$.

Following the official protocol, anonymization is applied at the *utterance level*: the pooled vector is resampled independently for every utterance, including enrollment and trial segments. This utterance-level configuration is also used by BM1, BM2 and BM3.

2.5. Speech synthesis

The three processed streams—200-dimensional content features, a 1-dimensional transformed F0, and a 192-dimensional

anonymous x-vector—are concatenated into a 393-dimensional representation. We use the same HiFi-GAN V1 generator [16] as BM1, comprising a multi-receptive-field fusion module and five transposed-convolution upsampling stages with rates 5, 4, 2, 2, 2.

3. Experiments and results

3.1. Experimental setup

We evaluate on the MLS [17] development and test partitions for English, German, French, and Spanish. These partitions contain 42, 30, 18, and 20 speakers per language, respectively, with enrollment and trial lists defined by the VPC 2026 protocol [4]. For emotion recognition, we use the EmoTrack2 benchmark, which comprises the categories angry, happy, neutral, and sad. The benchmark samples French utterances from Oreau, Spanish utterances from MESD, English utterances from EMNS, and German utterances from EmoDB, totaling 1,047 development and 994 evaluation utterances. Our baselines are the three official Track 2 systems [4]: BM1 [8] and the TTS-based BM2 and BM3 [18]. BM2 transplants prosody from the source utterance, whereas BM3 does not.

Three metrics follow the official protocol [4]. We measure privacy by the EER of an ECAPA-TDNN ASV system using WavLM-large and filter-bank, or Fbank, features [19, 4] under the semi-informed Anon-vs-Anon scenario. The ranking metric $\text{ASV}_{\text{eval}}^{\text{anon}}$ is computed *after* the attacker has been fine-tuned on language-specific anonymized training data. We measure content preservation by the WER of the official Whisper-large-v3 model [20, 4] and emotion preservation by the UAR of the official `emotion2vec_plus_large` checkpoint [21, 4], whose model ID is `iic/emotion2vec_plus_large`. Anonymization runs at approximately 1.2 s per utterance on a single NVIDIA RTX 3090. Of the two selection rules, U-Rand is our primary configuration.

3.2. Privacy

Table 1 reports Anon-vs-Anon EER. Both variants improve privacy over BM1 in every language on both partitions. W-Far increases the four-language development average from 1.64% to 5.68% and the test average from 2.49% to 6.21%; U-Rand reaches 6.56% on development and 7.05% on test. For U-Rand, the per-language development gains range from an increase of 1.3 percentage points for English to an increase of 9.0 percentage points for Spanish. The smaller gain for English may reflect the larger amount of attacker fine-tuning data available in that language. Both variants have lower EER than BM3 across all languages. BM3 approaches chance-level EER in German, French, and Spanish, whereas its English EER is substantially lower.

3.3. Utility and emotion preservation

Table 2 reports WER. U-Rand achieves an average development WER of 12.81%, and W-Far achieves 13.26%, compared with 8.44% for BM1, 26.52% for BM2, and 10.20% for BM3. The test-set averages follow the same ordering. For U-Rand, degradation relative to BM1 is largest for French on the development partition, where it increases by 7.2%, and for German on the test partition, where it increases by 6.9%. French nevertheless remains the language on which every system, including BM1, performs comparatively poorly. Both variants are markedly more intelligible than BM2, particularly on French and Spanish.

Table 1: *Privacy: semi-informed Anon-vs-Anon EER (%) under ASV^{anon}_{eval}, following fine-tuning of the attacker. Higher is better.*

System	en	de	fr	es	Avg.
<i>Development</i>					
BM1	1.40	1.43	2.10	1.63	1.64
BM2	6.34	11.38	16.93	11.91	11.64
BM3	23.49	46.84	47.09	46.23	40.91
Ours (W-Far)	2.95	4.74	5.92	9.12	5.68
Ours (U-Rand)	2.68	5.79	7.19	10.59	6.56
<i>Test</i>					
BM1	3.73	1.75	2.63	1.83	2.49
BM2	8.82	11.88	16.11	10.23	11.76
BM3	22.25	47.02	46.54	46.46	40.57
Ours (W-Far)	5.46	5.23	7.41	6.74	6.21
Ours (U-Rand)	5.08	6.24	9.54	7.34	7.05

Table 2: *Content preservation: ASR WER (%) from Whisper-large-v3. Lower is better.*

System	en	de	fr	es	Avg.
<i>Development</i>					
BM1	7.40	7.81	12.42	6.15	8.44
BM2	14.29	19.28	52.17	20.37	26.52
BM3	6.71	9.74	13.77	10.56	10.20
Ours (W-Far)	11.75	11.55	19.75	10.01	13.26
Ours (U-Rand)	11.30	11.31	19.59	9.04	12.81
<i>Test</i>					
BM1	8.54	9.38	9.93	6.21	8.51
BM2	15.41	24.40	42.24	28.92	27.74
BM3	7.73	10.10	11.99	10.12	9.98
Ours (W-Far)	12.93	16.66	16.41	12.15	14.54
Ours (U-Rand)	12.93	16.25	15.97	11.67	14.21

Table 3 reports emotion-recognition performance. Both variants achieve substantially lower UAR than BM1, but higher UAR than BM3 and comparable UAR to BM2. For U-Rand, the per-class development recalls are 43.7% for neutral, 34.3% for angry, 20.3% for happy, and 18.2% for sad.

Table 3: *Emotion preservation: UAR (%) on EmoTrack2. Higher is better.*

System	Dev	Eval.
BM1	41.30	42.63
BM2	27.13	29.04
BM3	26.24	24.84
Ours (W-Far)	28.93	28.21
Ours (U-Rand)	29.13	29.57

3.4. Comparison of the two selection rules

Because W-Far and U-Rand share the same F0 transformation, perturbation, and utterance-level anonymization, their differences provide a controlled comparison of the candidate-selection and weighting policies.

In the reported evaluation, U-Rand achieves higher EER, lower WER, and higher UAR than W-Far, although the margins are modest. Its EER is higher by 0.88 points on the development partition and by 0.84 points on the test partition. It also achieves slightly lower WER on both partitions and slightly higher UAR. The per-language results are not uniform. On the development

partition, U-Rand achieves an EER of 5.79% for German, compared with 4.74% for W-Far; 7.19% for French, compared with 5.92%; and 10.59% for Spanish, compared with 9.12%. English marginally favors W-Far, with an EER of 2.95% compared with 2.68% for U-Rand.

Selecting candidates according to their distance from the source makes the pooling policy explicitly source-dependent, whereas U-Rand is independent of the query. Under the semi-informed attack setting, U-Rand achieves higher privacy and lower WER than W-Far in our comparison. Because the two policies differ in both candidate selection and weighting, the experiment measures their combined effect and does not isolate the effect of similarity ranking. We therefore use U-Rand as the primary configuration.

4. Discussion and conclusion

We presented an enhanced selection-based voice-anonymization pipeline for VPC 2026 Track 2, adding a strength-controlled statistical F0 transformation and a perturbed speaker-pooling stage to the BM1 framework. Both selection variants consistently improve privacy over BM1 across four languages and both partitions, at the cost of higher WER and lower emotion UAR. Relative to BM2, the proposed variants are substantially more intelligible and provide intermediate privacy performance; relative to BM3, they preserve more emotion but provide lower privacy.

Because the two enhancements are not evaluated separately, their individual contributions cannot be quantified. The comparison in Section 3.4 keeps the F0 transformation active in both conditions and therefore evaluates the combined effect of the two pooling policies. Component-wise ablations are left to future work.

Future work should evaluate component-wise ablations and a broader range of the strength factor λ to characterize the privacy-utility trade-off. Further extensions include gender-conditioned targets and emotion-adaptive transformation strengths.

5. Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2024YFB2808802), the National Natural Science Foundation of China (NSFC) (No. 62271083), and the Key Project of the National Language Commission (No. ZDI145-81).

6. Generative AI Use Disclosure

The authors used large language models to assist with structural drafting and language editing. All methodological decisions, code implementation, experimental design, analyses, and conclusions were developed and verified by the authors.

7. References

- [1] 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, “Introducing the VoicePrivacy initiative,” in *Proc. Interspeech*, 2020, pp. 1693–1697.
- [2] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, “The VoicePrivacy 2022 challenge evaluation plan,” 2022.
- [3] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and

- M. Todisco, “The VoicePrivacy 2024 challenge evaluation plan,” 2024.
- [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,” VoicePrivacy Challenge, 2026, version 1.2. [Online]. Available: https://www.voiceprivacychallenge.org/vp2026/docs/VPC_2026_june14.pdf
 - [5] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, “Speaker anonymisation using the McAdams coefficient,” in *Proc. Interspeech*, 2021, pp. 1099–1103.
 - [6] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, “Speaker anonymization using X-vector and neural waveform models,” in *Proc. ISCA SSW*, 2019, pp. 155–160.
 - [7] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, “Privacy and utility of X-vector based speaker anonymization,” *IEEE/ACM TASLP*, vol. 30, pp. 2383–2395, 2022.
 - [8] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, “Language-independent speaker anonymization approach using self-supervised pre-trained models,” in *Proc. Odyssey*, 2022, pp. 279–286.
 - [9] S. Meyer, F. Lux, J. Koch, P. Denisov, C. Tilli, and N. T. Vu, “Prosody is not identity: A speaker anonymization approach using prosody cloning,” in *Proc. ICASSP*, 2023, pp. 1–5.
 - [10] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, “Speaker anonymization using neural audio codec language models,” in *Proc. ICASSP*, 2024, pp. 4725–4729.
 - [11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, 2021.
 - [12] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, “A comparison of discrete and soft speech units for improved voice conversion,” in *Proc. ICASSP*, 2022, pp. 6562–6566.
 - [13] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking,” in *Proc. ICASSP*, 2002, pp. I–361–I–364.
 - [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] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Proc. Interspeech*, 2019, pp. 1526–1530.
 - [16] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, 2020, pp. 17 022–17 033.
 - [17] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, “MLS: A large-scale multilingual dataset for speech research,” in *Proc. Interspeech*, 2020, pp. 2757–2761.
 - [18] S. Meyer, F. Lux, and N. T. Vu, “Probing the feasibility of multilingual speaker anonymization,” in *Proc. Interspeech*, 2024, pp. 4448–4452.
 - [19] 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 J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
 - [20] 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.
 - [21] 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*, 2024, pp. 15 747–15 760.