Mitigating Language Mismatch in SSL-Based Speaker Anonymization

Interspeech 2025

Zhe Zhang[1], Wen-Chin Huang[2], Xin Wang[1], Xiaoxiao Miao[3], Junichi Yamagishi[1]
[1]National Institute of Informatics, Japan
[2]Nagoya University, Japan
[3]Duke Kunshan University, China

Check out our paper, samples, and codes!

Paper Samples Codes

Motivation

- Existing speaker anonymization systems (SASs) primarily developed and evaluated using English speech, leading to degraded performance when applied to other languages.
- Preliminary experiments reveal significantly degraded intelligibility (increased CERs in ASR evaluation) for Japanese and Mandarin speech inputs.

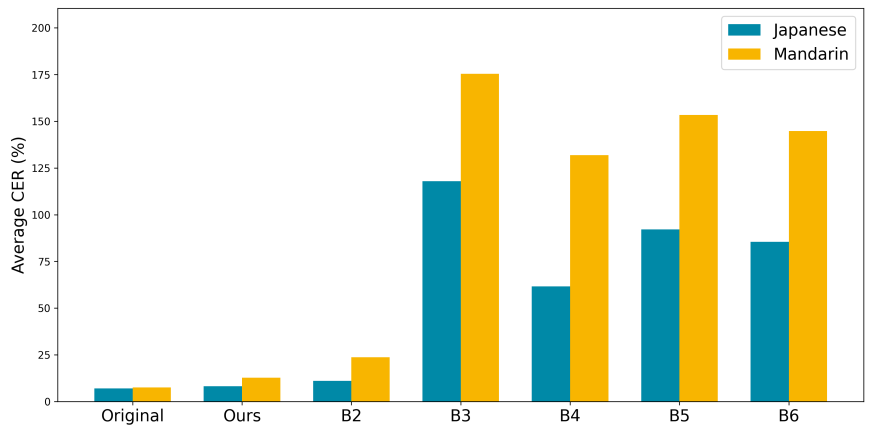

Figure 1: Utility evaluation via ASR on anonymized Japanese and Mandarin speech using VPC baselines.

Contributions

- Analyzed the impact of language mismatch on SSL-based speaker anonymization systems.
- Fine-tuned a multilingual SSL model (mHuBERT) on Japanese speech data to improve multilingual generalization.
- Demonstrated enhanced speech intelligibility in Japanese and Mandarin while preserving speaker privacy.

SSL-based Speaker Anonymization System

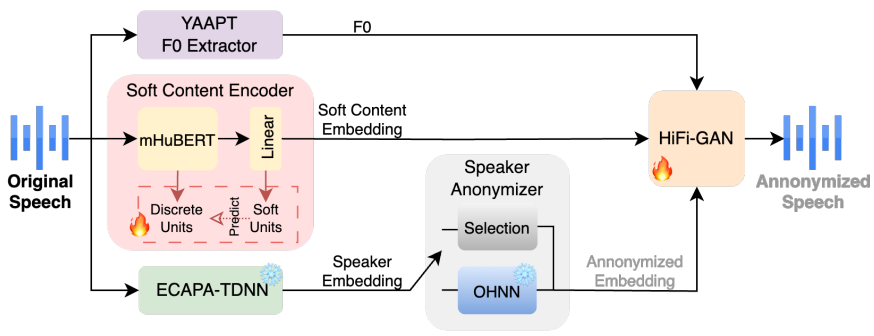

Figure 2: Framework of the SSL-based multilingual SAS.

- Soft Content Encoder: pre-trained mHuBERT fine-tuned on 100h of Japanese speech (CSJ) to predict soft content units.
- Speaker Encoder: ECAPA-TDNN providing speaker vectors.
- Speaker Anonymizer: selection-based (averaged vectors from an external pool) or OHNN-based (orthogonal Householder neural network) to hide speaker identity.
- Vocoder: HiFi-GAN synthesizing audio from extracted F0, content embeddings, and anonymized speaker embeddings.

Evalaution

- Language-Adapted Scenario (Japanese):
  - ASR (`Whisper-large-v3`): JVS corpus (9,976 utterances)
  - ASV (`Ignorant`): JTubeSpeech (76 enrollments, 276 tests)
- Language-Expanded Scenario (Mandarin):
  - ASR (`Whisper-large-v3`): AISHELL-3 (4,246 utterances)
  - ASV (`Ignorant`): AISHELL-3 (4,179 enrollments, 88 tests)

Table 1: Evaluation of privacy and utility of SASs in Japanese and Mandarin.

Metrics(%)	Ori.	Resynthesis			Selection			OHNN		
		HU-EN	HU-JA	mHU-JA	HU-EN	HU-JA	mHU-JA	HU-EN	HU-JA	mHU-JA
EER_{ja} ↑	14.91	29.41	27.19	24.90	47.87	48.15	39.91	43.33	44.70	39.44
CER_{ja}^{kata} ↓	3.03	5.27	4.35	**4.04**	5.57	4.88	**4.18**	6.18	4.78	**4.68**
CER_{ja}^{kanji} ↓	6.94	9.58	8.37	**8.06**	9.97	9.03	**8.18**	10.93	8.95	**8.90**
EER_{cn} ↑	5.56	18.76	16.51	14.09	44.33	41.62	35.21	42.55	33.28	31.20
CER_{cn} ↓	7.50	23.10	15.95	**10.39**	25.97	21.24	**12.67**	25.74	22.76	**14.13**

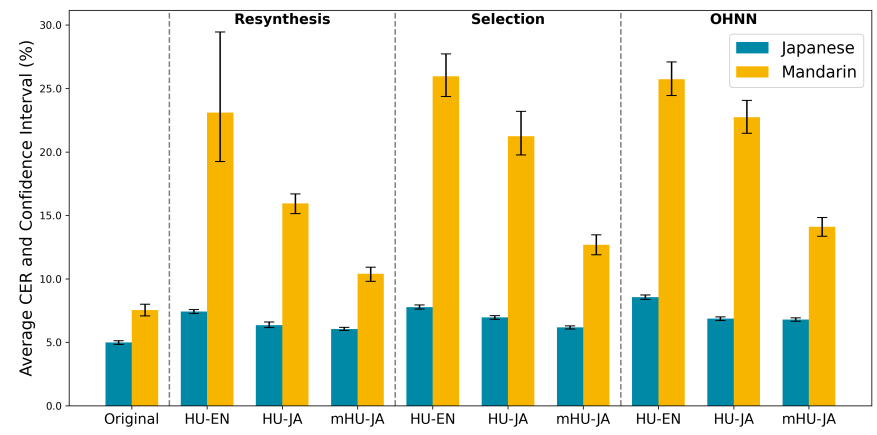

Figure 3: CER scores and confidence intervals.

- Multilingual SSL fine-tuning markedly enhances intelligibility of anonymized Japanese speech.
- The same model generalizes to Mandarin in a zero-shot manner, boosting ASR utility without extra tuning.
- Speaker privacy remains robust, achieving strong anonymization alongside improved utility.

Phonetic Analysis

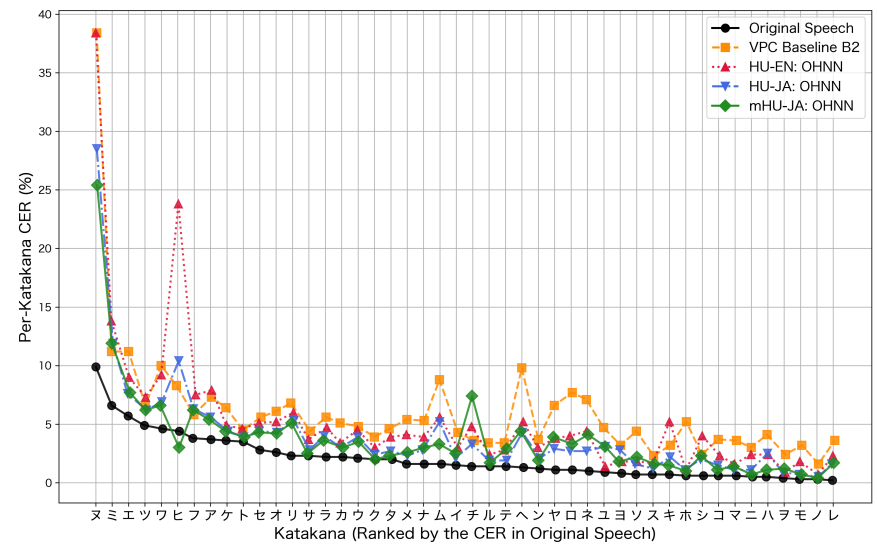

Figure 4: Per-Katakana CER of synthesized Japanese Speech.

- Per-Katakana CER of Japanese speech for fine-grained diagnosis of syllabary phonetic blocks in annonymzied utterances.
- Multilingual SSL effectively improves intelligibility of specific syllables problematic for English-only SAS models.

Conclusion and Future Work

- English-only SASs exhibit clear utility degradation on Japanese and Mandarin; addressing language mismatch is essential.
- Multilingual SSL fine-tuning (e.g., mHuBERT) improves intelligibility while maintaining privacy in both adapted and zero-shot settings.
- Next: extend to more (including low-resource) languages and dialects, deepen phonetic and listening evaluations, and evaluate the SASs under stronger attacker scenarios.

INTERSPEECH 2025 17-21 AUGUST—ROTTERDAM
NII National Institute of Informatics
名古屋大学
昆山杜克大学 DUKE KUNSHAN UNIVERSITY