

RocknRiver: Multilingual Speaker Anonymization System for VoicePrivacy 2026 Challenge

4th Voice Privacy Challenge

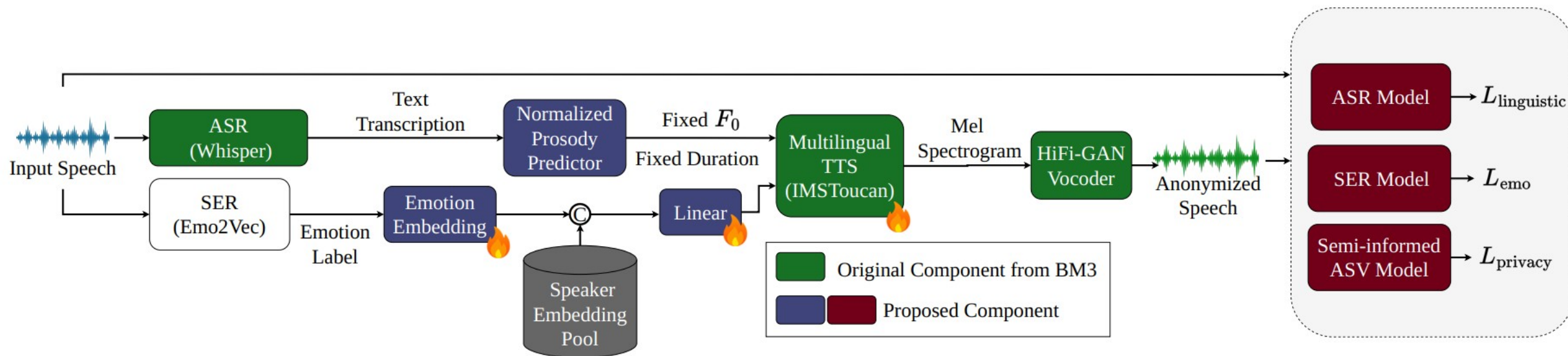
September, 2026

Japan Advanced Institute of Science and Technology

Main Approach

Modify BM3 baseline to minimize residual identity information from original speaker.

Use semi-informed ASV, SER and ASR as training targets to control properties of anonymized speech.



Multi-Objective Finetuning

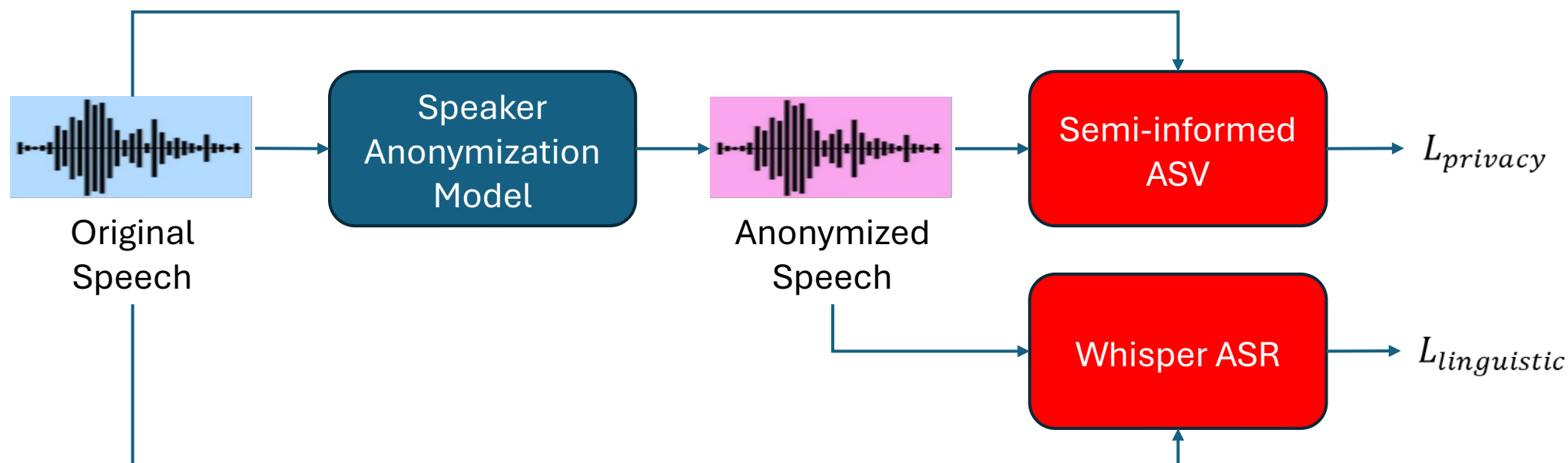
Privacy and Linguistic Loss are calculated by comparing similarity of original and anonymized speech.

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

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

z_o, z_a are embedding of original and anonymized speech extracted using semi-informed ASB, respectively.

h_o, h_a are final hidden layer of original and anonymized speech extracted using Whisper model, respectively.

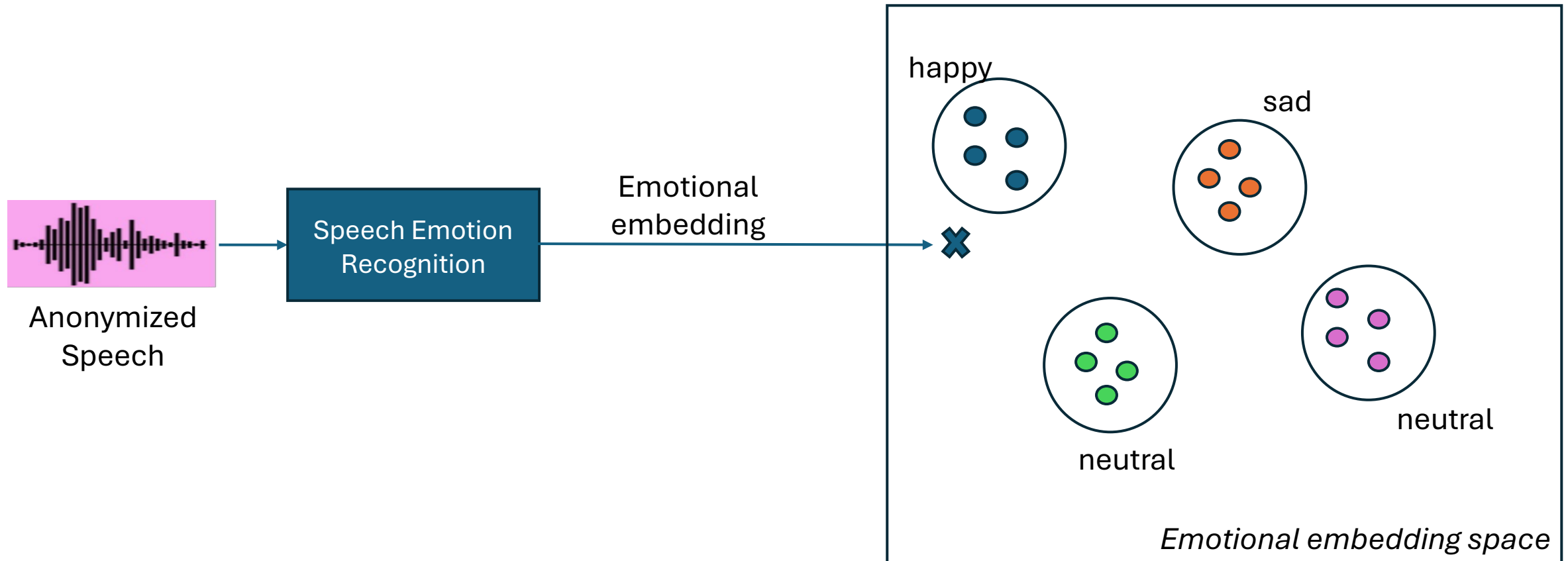


Multi-Objective Finetuning

For emotional loss, we employ a prototype-based contrastive loss.

A set of emotional embedding prototypes are prepared for each emotion.

- Anonymized speech's embedding is optimized to be close to target emotion.

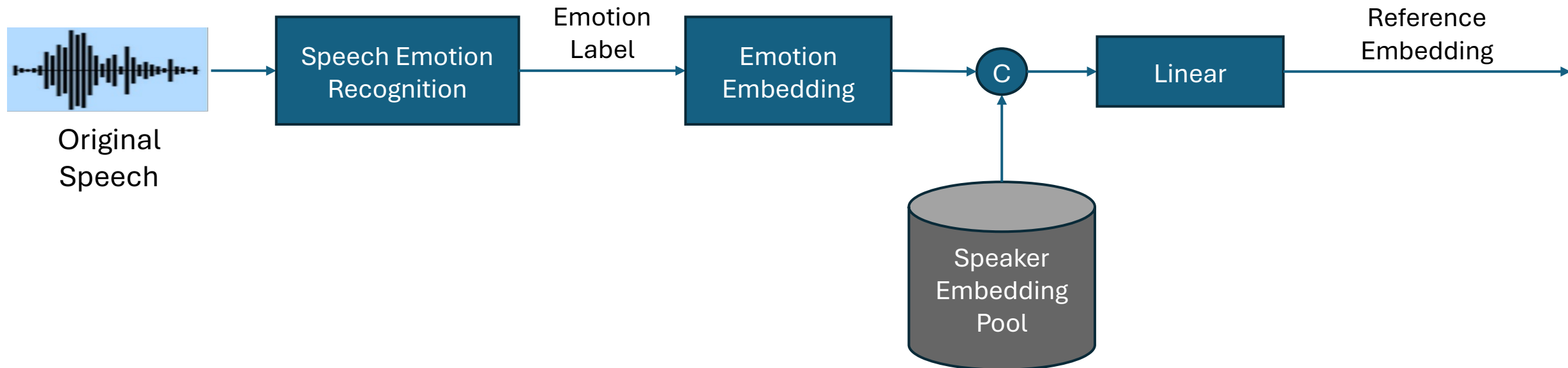


Speaker Embedding Selection

Replace GAN in BM3 with a pool of random speaker embedding since ASV could potentially reverse the mapping of GAN model.

The pool can be two options:

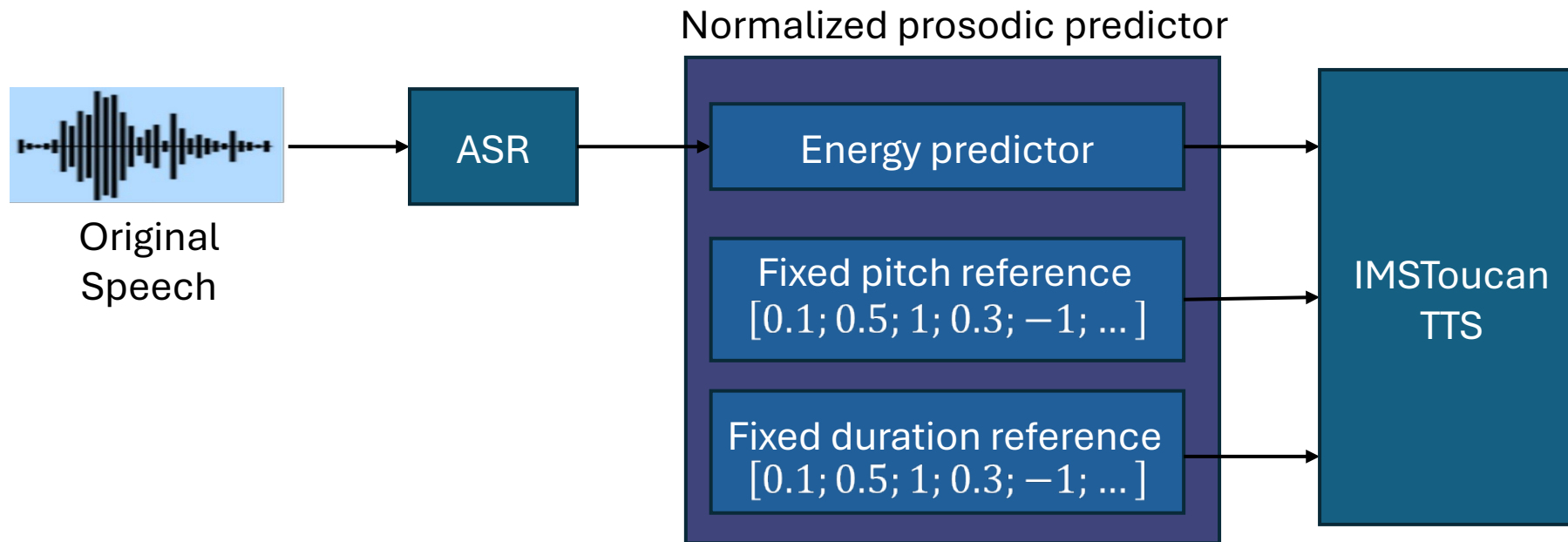
- Same reference embedding for every utterance (T21)
- Use one reference from a difference speaker for every utterance (T22)



Prosodic Normalization

Replace predicted prosodic feature (pitch, duration) with a fix reference sequence to suppress speaker-specific prosodic cues.

Energy feature is not replaced because it significantly degrade speech quality.



Fine-tuning Data

We combined multiple popular emotional dataset for fine-tuning our model, including:

- EMO-DB
- CREMA-D
- RAVDESS

Additionally, 3000 English utterances are synthesized using CosyVoice 2.0 model using reference and transcriptions from the LibriSpeech dataset.

Experiment Result

T21 consistently outperform baseline BM3 and T22.

T22 violate challenge’s rule due to not being utterance-level anonymization.

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	<u>13.77</u>	11.09	20.19	5.13	11.99	<u>13.32</u>	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	<u>10.56</u>	8.42	15.44	3.75	<u>10.12</u>	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	7.61
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	<u>9.74</u>	8.40	15.59	5.38	<u>10.10</u>	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	<u>10.20</u>	8.50	14.47	4.93	9.98	<u>10.09</u>	14.47
EmoTrack2	UAR ($\uparrow$)	72.09	26.24	39.40	<u>38.57</u>	78.90	24.84	34.47	<u>34.08</u>

Ablation Study

T21-E shows that anonymization using GAN (BM3) did not completely conceal original speaker identity
T21-P shows minimal improvement compare to T21, suggest that TTS rely heavily on speaker embedding instead of predicted prosodic features.

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	<u>35.34</u>
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>

Conclusion

- This study propose multi-objective guided fine-tuning using semi-informed ASV and modify components in BM3 baseline to better conceal original speaker identity.
- Experiment demonstrate that our solution effectively protect speaker identity while maintaining competitive speech intelligibility and emotion preservation.