



北京邮电大学
Beijing University of Posts and Telecommunications



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

Yilin Han, Ruixi Jiang, Tianyu Song, Lijie Liu,
Qihang Lu, Wenxin Fu, Yingming Gao, Ya Li*

Beijing University of Posts and Telecommunications, China

Motivation and contributions

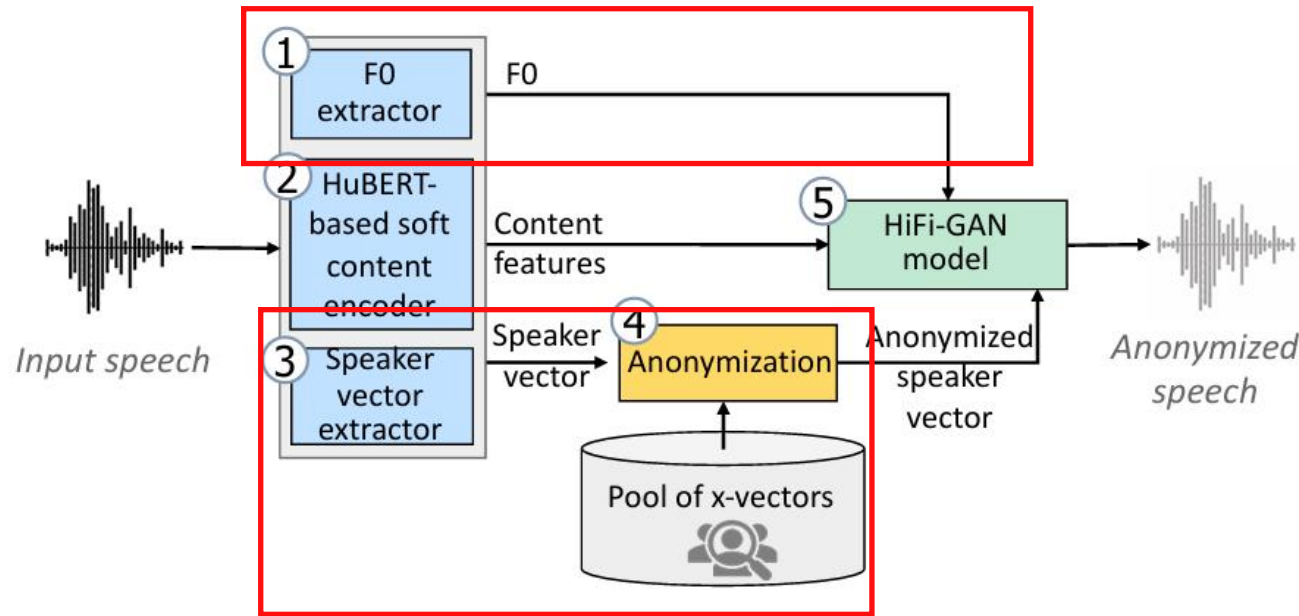


Figure 3: Baseline anonymization system BM1.

BM1 provides strong content preservation, but its privacy protection remains limited.

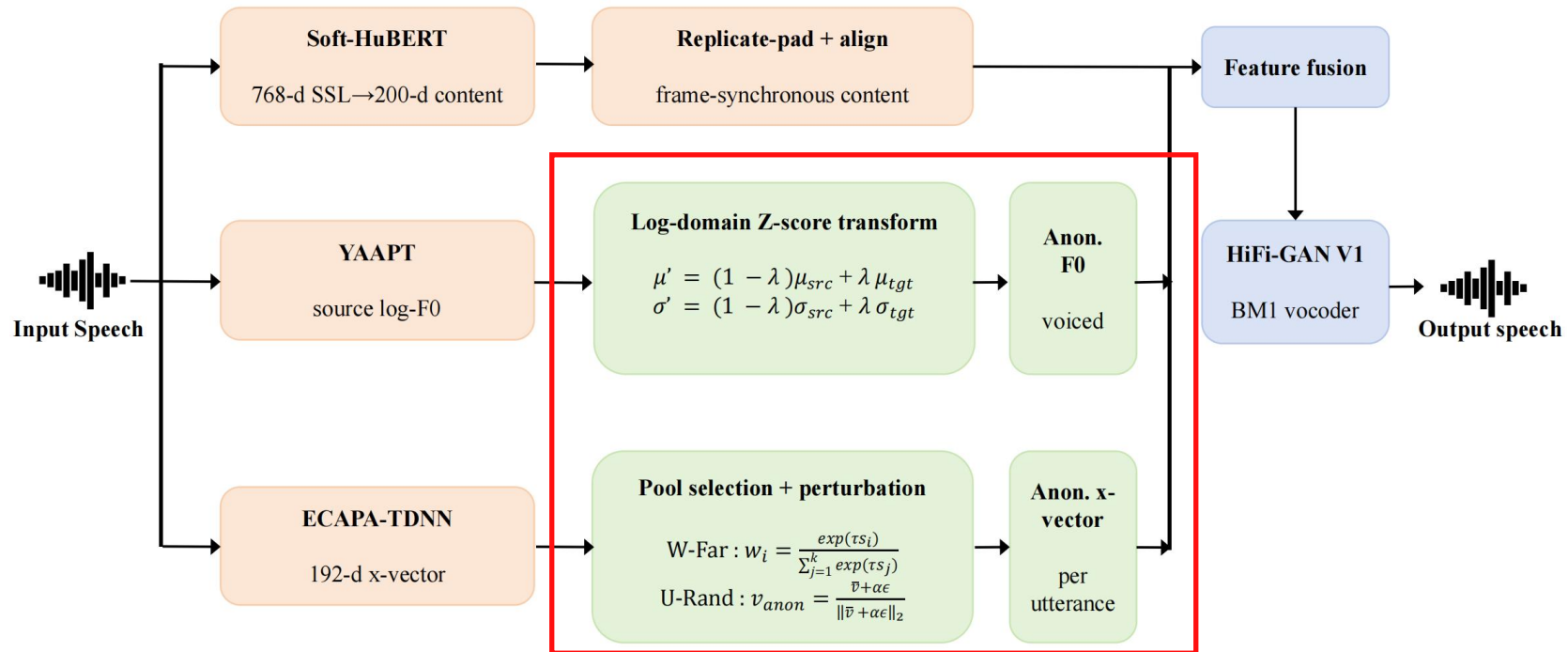
➤ **Unchanged F0**

- BM1 passes the original F0 contour directly to the vocoder.
- Pitch remains a strong speaker-identifiable cue.

➤ **Source-conditioned speaker pooling**

- BM1 selects anonymous speaker vectors according to their distance from the source x-vector.
- The anonymization process therefore remains explicitly conditioned on the source speaker.

Motivation and contributions



We modify the two speaker-related streams while retaining the BM1 content path and vocoder.

➤ **Statistical F0 transformation**

Remap log-F0 statistics toward a gender-neutral target register.

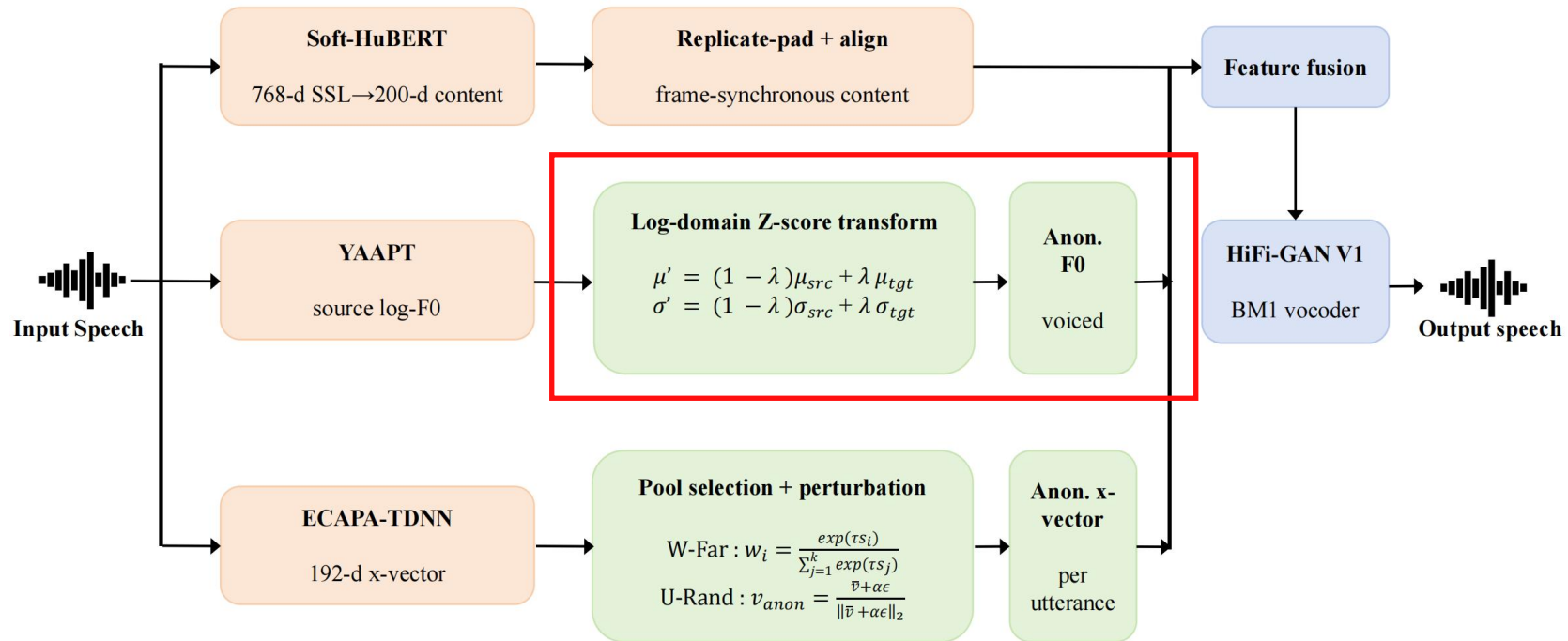
➤ **Alternative speaker-pool selection**

Compare weighted farthest selection with query-independent uniform random selection.

➤ **Stochastic embedding perturbation**

Add locally scaled Gaussian noise and resample the anonymous x-vector for each utterance.

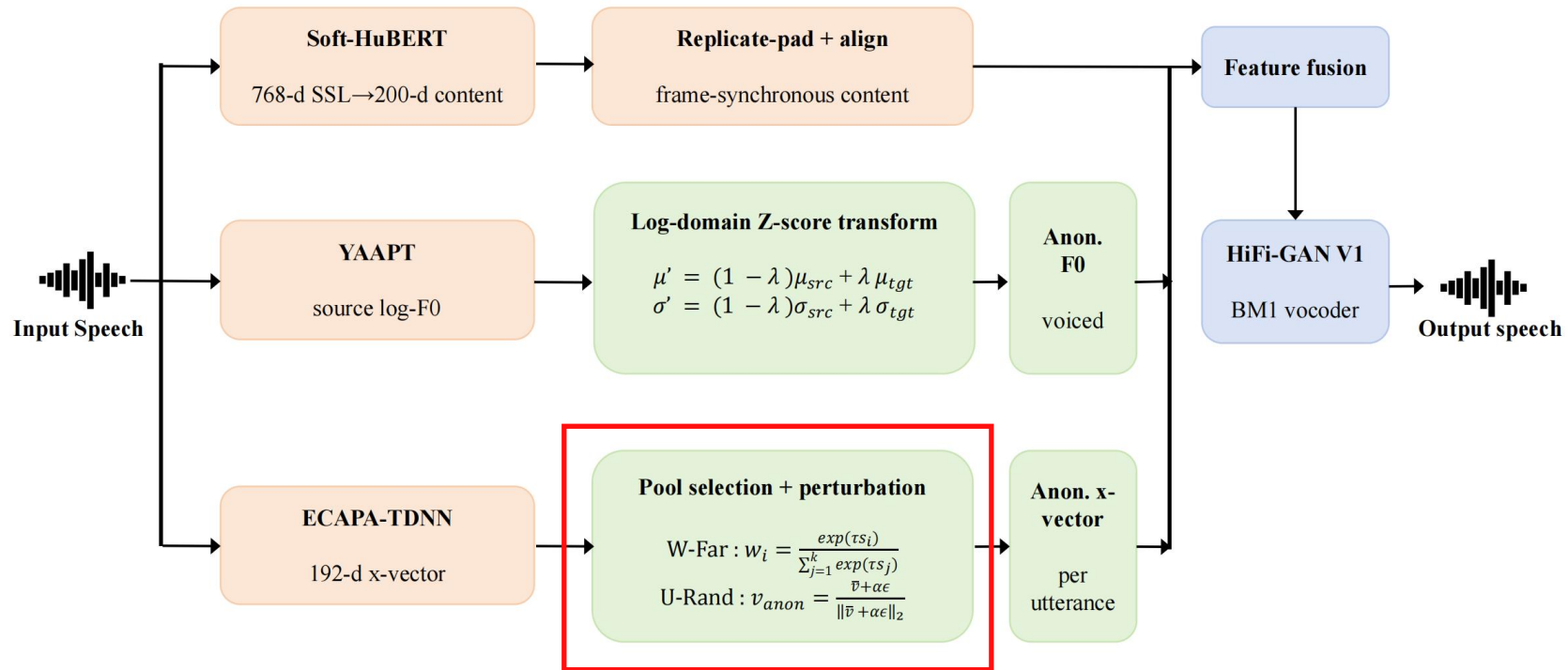
Contribution 1: Statistical F0 transformation



Speaker-dependent pitch statistics are shifted while the relative F0 contour is largely preserved.

Source log-F0 mean: μ_{src} std: σ_{src} $\mu' = (1 - \lambda) \mu_{src} + \lambda \mu_{tgt},$ $\sigma' = (1 - \lambda) \sigma_{src} + \lambda \sigma_{tgt},$	→	Z-score remapping in log domain $\log f_{anon} = \frac{\log f_{src} - \mu_{src}}{\sigma_{src}} \sigma' + \mu'.$	→	Target register median F0 approx: 158 Hz sigma_tgt = 0.20
---	---	---	---	--

Contribution 2: Speaker-pool selection



We compare a source-conditioned selection rule with a query-independent alternative.

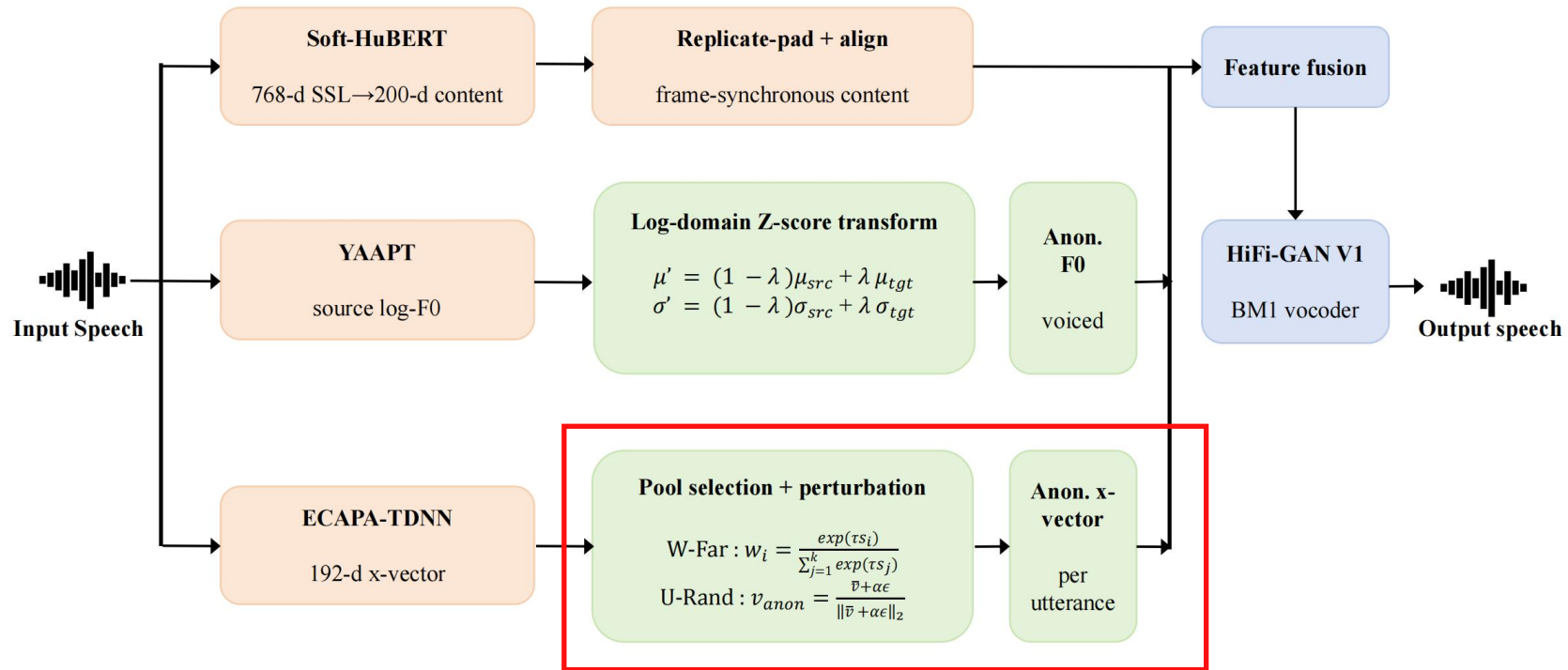
A. W-Far

Keep 50 least-similar candidates
Sample 10 vectors
Temperature-scaled weighting
Source-conditioned

B. U-Rand

Sample 10 vectors from full pool
Uniform weighting
Query-independent
Primary configuration

Contribution 3: Stochastic embedding perturbation



The pooled speaker embedding is perturbed and independently regenerated for every utterance.

- Pool the selected vectors $\bar{v} = \sum_{i=1}^k w_i v_i$
- Estimate local variation $\epsilon_d \sim \mathcal{N}(0, \sigma_{local,d}^2)$,
- Perturb and renormalize $v_{anon} = \frac{\bar{v} + \alpha \epsilon}{\|\bar{v} + \alpha \epsilon\|_2}$,

Main results

	EER(%)					WER(%)					UAR(%)
system	en	de	fr	es	Avg.	en	de	fr	es	Avg.	Eval.
BM1	3.40	1.36	2.08	1.58	2.10	8.54	9.38	9.93	6.21	8.51	42.63
BM2	7.53	10.90	15.30	8.74	10.62	15.41	24.40	42.24	28.92	27.74	29.04
BM3	20.40	46.05	46.25	45.92	39.65	7.73	10.10	11.99	10.12	9.98	24.84
Ours(W-Far)	4.55	4.29	6.62	5.80	5.32	12.93	16.66	16.41	12.15	14.54	28.21
Ours(U-Rand)	4.64	5.79	7.45	5.90	5.95	12.93	16.25	15.97	11.67	14.21	29.57

➤ Privacy

U-Rand increases EER compared with BM1.

Both proposed variants improve EER across all four languages.

➤ Content

WER increases from 8.51% to 14.21%.

U-Rand remains substantially better than BM2 at 27.74%.

➤ Emotion

UAR decreases compared with BM1.

U-Rand retains more emotion information than BM3.

Summary

➤ Conclusions

➤ System

- We enhanced the BM1 pipeline by anonymizing both pitch statistics and speaker embeddings. The Soft-HuBERT content path and HiFi-GAN vocoder remain unchanged.
- Both W-Far and U-Rand consistently improve privacy over BM1. U-Rand achieves the strongest overall performance among our two variants. The privacy gain comes with higher WER and lower emotion UAR.

➤ Key observation

- Query-independent random pooling performs slightly better than source-conditioned farthest selection in the reported evaluation.

➤ Limitations and future work

➤ Limitations

- The individual effects of F0 transformation and embedding perturbation remain unknown.
- W-Far and U-Rand differ in both selection and weighting.

➤ Future work

- A broader sweep of the transformation strength
- Gender-conditioned target pitch distributions
- Emotion-adaptive F0 transformation
- Improved privacy–utility optimization



北京邮电大学
Beijing University of Posts and Telecommunications



Thank You & Feedback Welcome!

Yilin Han, Ruixi Jiang, Tianyu Song, Lijie Liu,
Qihang Lu, Wenxin Fu, Yingming Gao, Ya Li*

e-mail: hyl0927@bupt.edu.cn