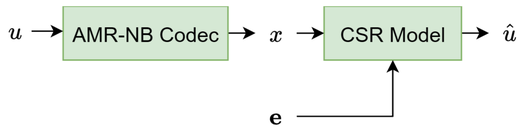
PERSONALIZED LOW-BITRATE CODED SPEECH RESTORATION VIA EMBEDDING CONDITIONING AND CLUSTER-BASED MODEL SELECTION

Eray Özgünay[1,2,3], Christian Rollwage[1], Jan Rennies[1,2], Xavier Domont[3], Mattes Ohlenbusch[1], Simon Doclo[1,2]

[1]Fraunhofer IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany
[2]Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
[3]CEOTRONICS AG, Germany

Introduction

- **Problem:** Legacy low-bitrate codecs (e.g., AMR-NB) still widely used → narrowband speech, coding artifacts, listener fatigue
- **Coded speech restoration (CSR):** restore coded speech (8 kHz, AMR-NB @ 4.75 kbps) to wideband (16 kHz)
- **Baseline: HiFi-GAN+ [1]**: A GAN-based bandwidth extension model with a non-causal WaveNet generator and multi-scale discriminators (1.061 M params, 30.54 GMACs/s), **retrained for CSR.**
- **Question:** Can personalization improve CSR beyond non-personalized models?



Speaker Embeddings

- Pretrained frozen ECAPA-TDNN [2]: 192-dimensional embedding vector **e**
- **Training time:** one embedding per speaker, averaged over all training utterances:

$$\mathbf{e}_s^{\text{train}} = \frac{1}{|\,U_s^{\text{train}}\,|} \sum_{u \in U_s^{\text{train}}} f\,(u)$$

- **Test time:** one enrollment embedding per test utterance ($j \neq i$, same speaker):

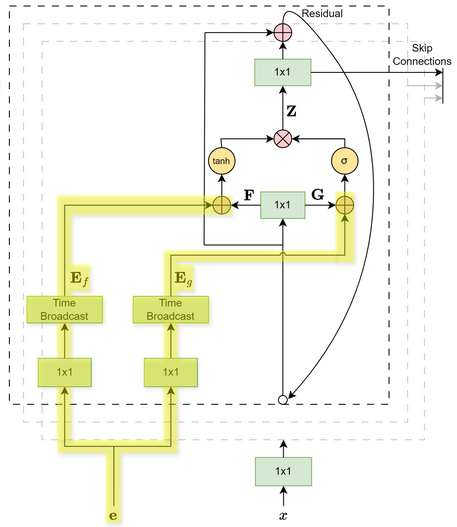
$$\mathbf{e}_{s,j}^{\text{test}} = f\big(u_{s,j}^{\text{test}}\big)$$

Embedding Conditioning

- **Single model** conditioned on a speaker embedding
- In each gated residual block: **e** projected via two learned 1×1 convolutions → added to filter and gate branches before nonlinearities:

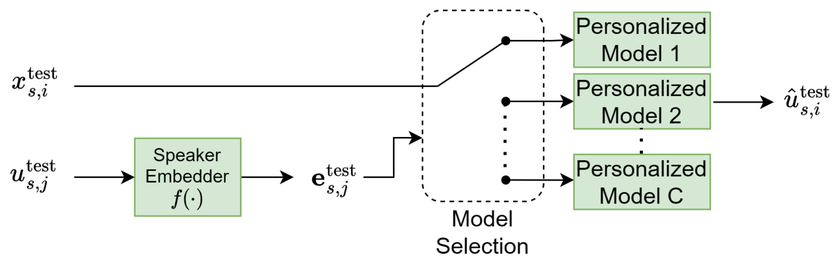
$$\boldsymbol{Z} = \tanh(\boldsymbol{F} + \boldsymbol{E}_f) \odot \sigma(\boldsymbol{G} + \boldsymbol{E}_g)$$

- Runtime complexity = baseline (+0.001% GMACs); storage = 1.4 × baseline



Cluster-Based Model Selection

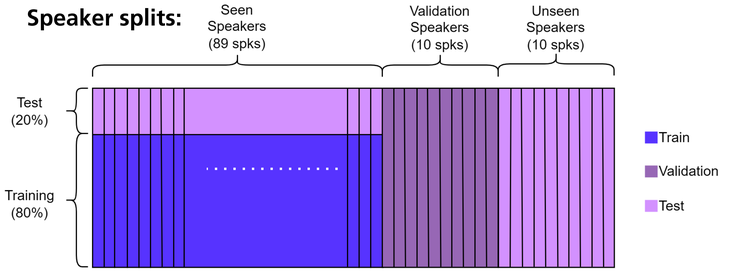
- Inspired by cluster-specific neural speech codecs [3]: Partition training-time speaker embeddings into $C = 4$ clusters via k-means → train **one model per cluster**
- Each model specializes on acoustically similar speakers; identical architecture to baseline
- **Test time:** Select a model based on enrollment embedding (nearest centroid)
- Runtime complexity = baseline (one model active); storage = 4 × baseline



Experimental Setup

- **Dataset:** VCTK: 109 English speakers, ~400 utterances each
- **Codec:** AMR-NB @ 4.75 kbps (8 kHz) → target: 16 kHz clean
- **Metrics:** LSD ↓ (spectral distortion), WB-PESQ ↑, ViSQOL ↑ (perceptual quality)

Speaker splits:



Results

Seen speakers

Method	LSD	ViSQOL	WB-PESQ
CSR baseline	0.95 ± 0.03	2.76 ± 0.49	2.21 ± 0.32
Embedding conditioning	0.95 ± 0.03	2.95 ± 0.51	2.22 ± 0.31
Cluster-based selection	**0.92 ± 0.03**	**3.00 ± 0.48**	**2.39 ± 0.33**

Unseen speakers

Method	LSD	ViSQOL	WB-PESQ
CSR baseline	0.94 ± 0.03	2.90 ± 0.53	2.18 ± 0.34
Embedding conditioning	0.95 ± 0.03	3.01 ± 0.52	2.15 ± 0.32
Cluster-based selection	**0.92 ± 0.03**	**3.06 ± 0.47**	**2.34 ± 0.33**

Conclusions

- Investigated **two personalization methods** for low-bitrate CSR with minimal inference overhead
- Cluster-based model selection consistently **improves** all metrics over the baseline
- **Outlook:** Personalization for causal low-complexity CSR models

1. J. Su, Y. Wang, A. Finkelstein, and Z. Jin. "Bandwidth Extension Is All You Need". In: *Proc. IEEE ICASSP*. 2021.
2. B. Desplanques, J. Thienpondt, and K. Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-Based Speaker Verification". In: *Proc. Interspeech*. 2020.
3. I. Jang, H. Yang, W. Lim, S. Beack, and M. Kim."Personalized Neural Speech Codec". In: *Proc. IEEE ICASSP*. 2024.

Funded by the European Union