

One Embedding Doesn't Fit All

Rethinking Speaker Modeling for Various Speech Applications

Shuai Wang

Nanjing University

October 1, 2026



Background on Speaker Modeling

Observations and Analysis Across Applications

Limitations of Current Speaker-related Tasks

Discussion on Future Directions & Guiding Principles

Conclusion

Speaker Modeling

The voiceprint is an important biometric trait: each speaker's voice carries distinctive characteristics shaped by vocal anatomy and speaking habits.

Speaker modeling learns representations of these characteristics to capture **who is speaking**, independent of what is said.



Speaker modeling in different tasks

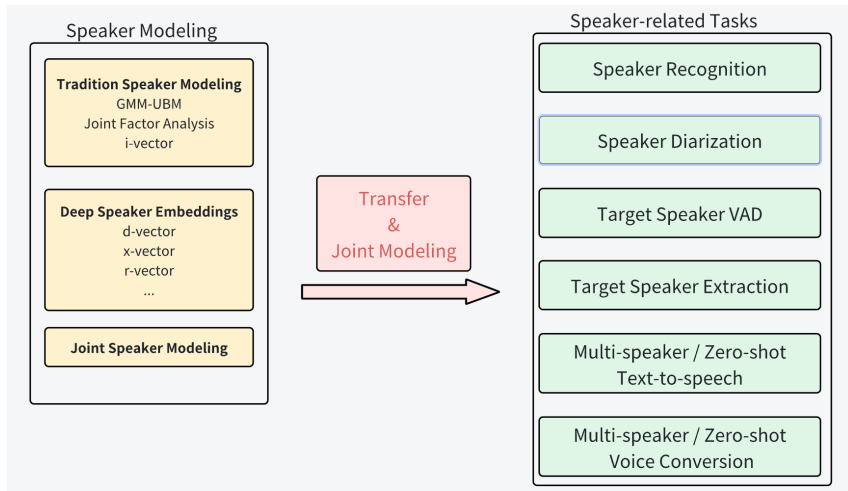


Figure from Wang et al. (2024)¹

¹Wang et al., "Advancing speaker embedding learning: Wespeaker toolkit for research and production".

Applications of Speaker Modeling

Embedding Extraction for speaker verification

Speaker Verification: My voice is my password

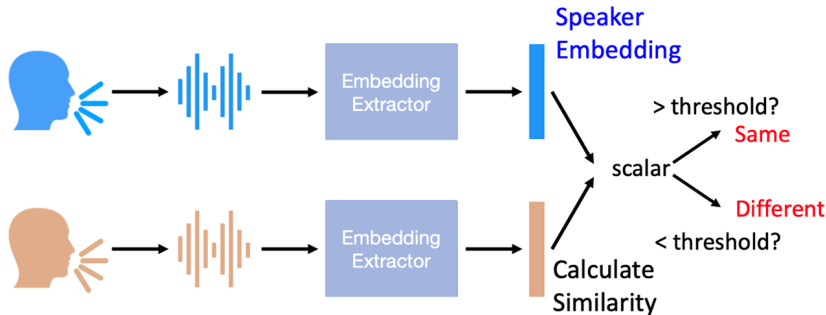


Figure from Prof. Hung-yi Lee's DLHLP20 slides²

²<https://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/>

Applications of Speaker Modeling

Reference/Cue modeling for Target speech extraction

Target Speech Extraction: I only hear your voice

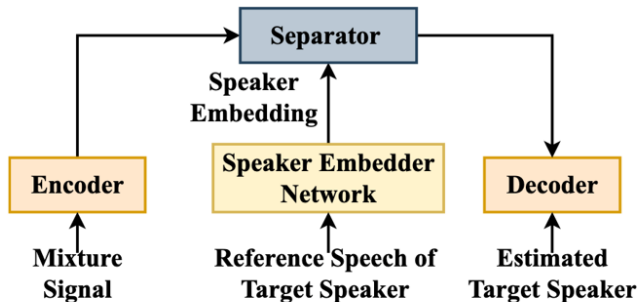


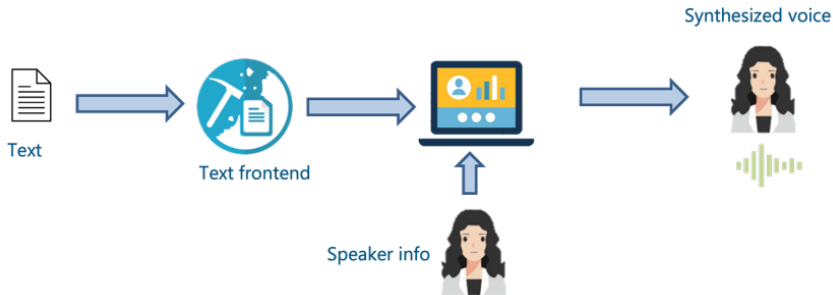
Figure from Sinha et al. (2021)³

³Sinha et al., "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures".

Applications of Speaker Modeling

Target voice identifier for Text-to-speech

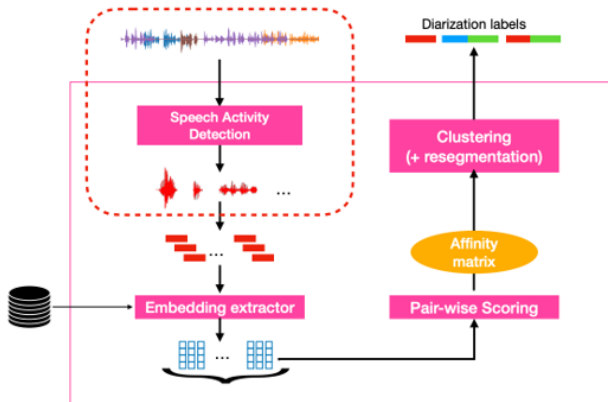
Speech generation: generate voice for the target identity.



Applications of Speaker Modeling

Embedding clustering based Speaker diarization

Speaker diarization: Who Spoke When?



4

Speaker Recognition: The Traditional Stronghold

Established recipe

- ▶ **Encoders:** TDNN, ECAPA-TDNN, ResNet
- ▶ **Training data:** VoxCeleb, VoxBlink, and other large speaker corpora
- ▶ **Losses:** AM/AAM-softmax, GE2E
- ▶ **Evaluation:** EER and minDCF

Strengths

- ▶ Discriminative identity cues
- ▶ Robustness to nuisance variation in tested conditions
- ▶ Compact, reusable embeddings

What it may miss

- ▶ Within-speaker prosody, emotion, and style
- ▶ Mixture-specific cues for target extraction
- ▶ Robustness across deployment scenarios

OpenSVBench: Verification Across Scenarios

26 trial sets from 16 public sources cover **11 deployment scenarios** and 9.15 million verification trials.

SCENARIO RANKINGS

Evaluation coverage

Open a scenario to compare systems using the corresponding group-level EER, minDCF, and trial-set evidence.

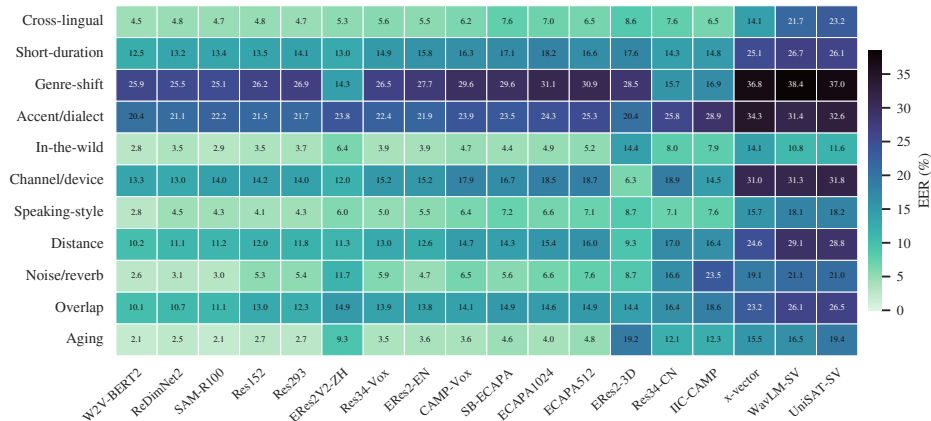
SCENARIO	EVALUATION FOCUS	TRIAL SETS	INCLUDED TRIAL SETS	CURRENT LEADER
Accent / Dialect Robustness Accent/dialect	Accent and dialect variation	2	GLOBE 3D-Speaker Dialect	wespeaker/w2vbert2
Aging Robustness Aging	Speaker aging and time gaps	2	VoxKneset VoxPopuli Aging	wespeaker/samresnet100
Channel / Device Robustness Channel/device	Device and channel variation	2	3D-Speaker Device FFSVC 2022 Cross-Channel	iic/eres2net-large-3dspeaker
Cross-Lingual Robustness Cross-lingual	Language mismatch	1	TidyVoiceX2-ASV	wespeaker/w2vbert2
Distance Robustness Distance	Distance mismatch	4	3D-Speaker Distance AliMeeting Near/Far CHiME-6 Domestic Far-Field +1	iic/eres2net-large-3dspeaker

Online scenario leaderboard (snapshot supplied by the author); 18 ranked systems.⁴

⁴Shangguan, Xu, and Wang, *OpenSVBench: Diagnosing Speaker Verification Generalization Across Deployment Scenarios*.

No Single SV System Leads Everywhere

Scenario EER (%) (lower is better) across 18 released speaker verification systems.



Even within verification, no single system wins everywhere.

What happens when the task is no longer verification?

The "One Embedding Fits All" Fallacy

The Pervasive Assumption

Speaker embeddings optimized for recognition are universally effective

Current Reality:

- ▶ x-vectors, ECAPA-TDNN, ResNets widely adopted
- ▶ Assumed universal applicability
- ▶ Direct transfer to other tasks

The Problem

$SR \neq TTS \neq VC \neq TSP$

Critical Question

What makes an "ideal" speaker representation?

The Core Dilemma

Speaker Recognition

- ▶ Maximize discrimination
- ▶ Minimize intra-speaker variance
- ▶ Robustness to noise/channel
- ▶ Compact representation

Generative Tasks

- ▶ Capture nuanced details
- ▶ Preserve prosody/emotion
- ▶ Enable natural synthesis
- ▶ Rich, expressive representation

Target Speaker Processing

- ▶ Relative Discrimination
- ▶ Maximize the co-relation with target speech in **mixture**
- ▶ Compact representation

The Conflict

Discriminative optimization vs. Generative richness
Absolute discrimination vs. Relative discrimination

- ▶ SID/SV: Suppress intra-speaker variability
- ▶ TTS/VC: Preserve and utilize this variability
- ▶ TSE, TS-ASR, SD: Relative discrimination within a small set of speakers

Background on Speaker Modeling
Observations and Analysis Across Applications
Limitations of Current Speaker-related Tasks
Discussion on Future Directions & Guiding Principles
Conclusion

Text-to-Speech: Beyond Simple Embeddings

The challenge

Generate natural, expressive speech that sounds like the target speaker

What TTS needs

- ▶ Recognizable timbre
- ▶ Natural prosody
- ▶ Emotion and speaking-style control

Two ways to supply voice cues

- ▶ A speaker vector gives a compact identity cue
- ▶ An acoustic prompt can also supply style and timing

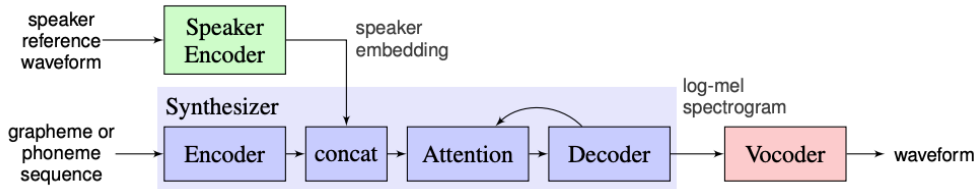
The gap

One pooled speaker vector:

- ▶ Has no explicit time axis
- ▶ May miss changing prosody and emotion
- ▶ Does not by itself control accent or style

Text-to-Speech: Explicit Speaker Conditioning

A reference utterance is mapped to a speaker embedding⁵

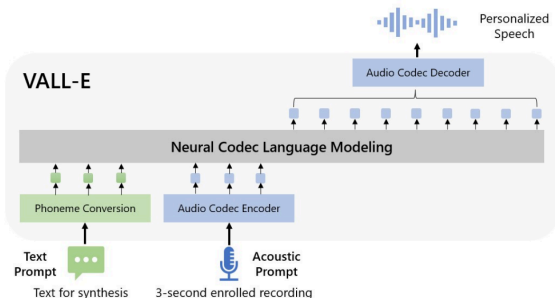


- ▶ A speaker encoder compresses the reference waveform into a fixed vector.
- ▶ The synthesizer conditions on that vector alongside the text representation.
- ▶ This gives a compact identity cue, but does not directly retain the reference's time-varying prosody.

⁵Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis".

Text-to-Speech: Implicit Speaker Modeling

An acoustic prompt supplies the voice context⁶



- ▶ Text specifies what to say; an acoustic prompt demonstrates how the target voice sounds.
- ▶ A codec language model predicts speech tokens from both inputs; no separate speaker vector is extracted.

⁶Wang et al., "Neural codec language models are zero-shot text to speech synthesizers".

Text-to-Speech: Does Embedding Choice Really Matter?

Stan and O'Mahony (2023)⁷

1. The tested embedding sets produced similar synthesis quality in the same FastPitch system.
2. Speaker information remained in the network core under the standard training setup.
3. Zero speaker conditioning did not yield a stable, speaker-neutral output.

What this establishes

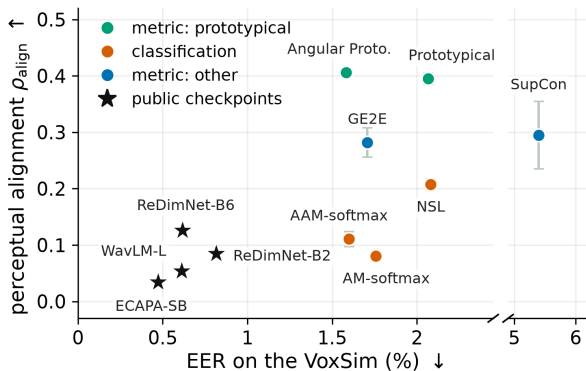
A speaker condition does not guarantee that the rest of a TTS model is independent of speaker identity.

Scope: one non-autoregressive architecture and 18 Romanian speakers; this study does not compare all zero-shot TTS systems.

⁷Stan and O'Mahony, "An analysis on the effects of speaker embedding choice in non auto-regressive TTS".

Voice Similarity: Verification Is Not Perception

VoxSim and a controlled embedding study⁸⁹



- ▶ **VoxSim:** 69,409 listener ratings on 41,578 pairs from 1,251 speakers.
- ▶ Better verification EER does not consistently track graded human voice similarity.
- ▶ For TTS and VC, check listener agreement; assess accent separately.

⁸Ahn et al., "VoxSim: A perceptual voice similarity dataset".

⁹Chen et al., *Rethinking Automated Voice Similarity by Shifting from EER to Embedding Geometry*.

TTS Evaluation: What Speaker Similarity Misses

Accent-emotion case study¹⁰

Key limitation

Speaker-embedding similarity alone does not establish accent fidelity.

- ▶ Published CosyVoice 2 / MaskGCT speaker-similarity rankings differ across WavLM and ERes2Net measures.
- ▶ Accent SMOS for selected utterances: **1.17–4.13** (CosyVoice 2) versus **3.33–4.33** (MaskGCT).

Evaluation implication

Add accent-focused listening and accent-specific analysis.

Scope: 30 listeners; one reference speaker; samples selected for low accent-embedding similarity. Speaker-similarity scores came from another protocol.

¹⁰Hayden, Zhong, and Richmond, “Accent-Emotion Entanglement in LM-Based Text-to-Speech Systems”.

Speaker Similarity Across Languages

Natural Cantonese–English speech¹¹

Human–machine alignment

Across 18 encoders and 40 listeners, machine–human voice-similarity correlations were $\rho = .34\text{--}.57$ (mean .44).

Speech-pair language	Same–different similarity gap
Cantonese only	.19
English only	.21
Mixed language	.16

Implication for accent evaluation

Language changes identity-similarity scores. Evaluate synthesized accent separately.

¹¹Xu et al., “Human-like cross-language generalisation in deep neural speaker embeddings and its acoustic foundations”.

Voice Conversion: The Disentanglement Challenge

The fundamental problem

Speech carries identity, content, prosody, accent, emotion, ...

Current approaches

- ▶ Adversarial training
- ▶ Gradient reversal layers
- ▶ Multi-encoder architectures
- ▶ Self-supervised representations

Challenges

- ▶ Complete separation is hard to verify
- ▶ Residual speaker information
- ▶ Content–speaker entanglement
- ▶ Prosody and accent may drift

Open question

Which factors must be separated for useful voice conversion?

Voice Conversion: Frame Matching Without a Speaker Vector

kNN-VC on LibriSpeech test-clean¹²

Method

Match each source WavLM frame to nearby target-reference frames, then synthesize with a trained vocoder.

System	WER ↓	Naturalness MOS ↑	Speaker SIM ↑
FreeVC	7.61	4.07	2.38
kNN-VC	7.36	4.03	2.91

Implication

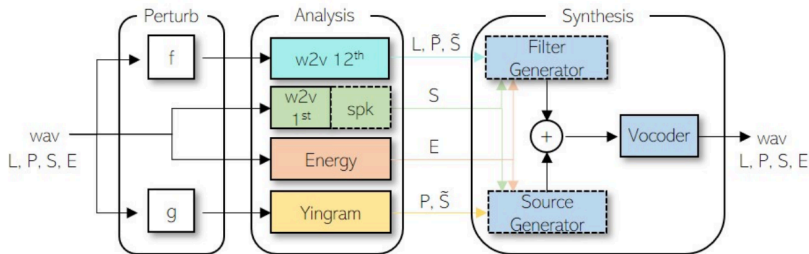
A single speaker-identity vector is not required for successful voice conversion in this setting.

Scope: about eight minutes of target speech per speaker. The paper reports weaker results with less than 30 seconds of reference speech.

¹²Baas, Niekirk, and Kamper, "Voice Conversion With Just Nearest Neighbors".

Voice Conversion: Factorized Conditioning

NANSY analyzes several aspects of the source signal¹³



- ▶ The analysis stage represents linguistic content, pitch, speaker characteristics, and energy separately.
- ▶ The synthesis stage recombines the cues; changing the speaker cue enables voice conversion.

¹³Choi et al., "Neural analysis and synthesis: Reconstructing speech from self-supervised representations".

Target Speaker Extraction: Matching the Reference

Task

Recover one voice from a mixture using a short reference utterance.

Global speaker cue

- ▶ Pool the reference into one vector
- ▶ Condition the separator with that vector
- ▶ Convenient for unseen speakers

Reference interaction

- ▶ Keep the reference as a sequence
- ▶ Align its features with the mixture
- ▶ USEF-TSE uses cross-attention^a

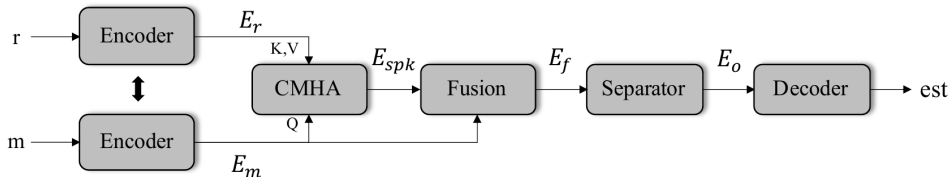
^aZeng and Li, "Usef-tse: Universal speaker embedding free target speaker extraction".

Design question

How much reference detail does the separator need for this mixture?

Target Speaker Extraction: USEF-TSE

The reference and mixture interact before separation¹⁴

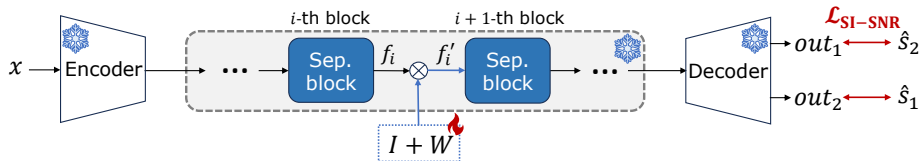


- ▶ The mixture encoding supplies queries; the reference encoding supplies keys and values to cross-attention.
- ▶ Cross-attention produces a frame-level target cue aligned with the mixture.
- ▶ The cue is fused with mixture features before separation, without a precomputed speaker-recognition embedding.

¹⁴Zeng and Li, "Usef-tse: Universal speaker embedding free target speaker extraction".

Target Speaker Extraction: Relative Selection

Two-speaker AV-TSE with a frozen separator¹⁵



- ▶ A small learned matrix swaps two output channels through one intermediate layer; the audio backbone stays frozen.
- ▶ At the final separator block, steering preserves **96.22–99.91%** of the original SI-SDR improvement across four backbones.

What this suggests

The cue may only act as a **local, relative selector** among candidates in the mixture.

¹⁵Kwak, Lee, and Chung, *Plug-and-Steer: Decoupling Separation and Selection in Audio-Visual Target Speaker Extraction*.

Target-Speaker Tasks: ASV Is Not Enough

A unified TS-ASR, TSE, and p-VAD comparison¹⁶

Speaker encoder	ASV EER ↓	TS-ASR WER ↓	TSE SI-SDR ↑	p-VAD mAP ↑
ECAPA-TDNN-c1024	1.32	20.64	9.64	98.28
ECAPA-TDNN-DINO	2.67	18.84	10.59	98.82

What the comparison shows

The better verification encoder gives worse results on all three target-speaker tasks in this comparison.

ASV EER follows the paper's SUPERB protocol; target-speaker results use the clean Libri2Mix subset. Lower WER/EER and higher SI-SDR/mAP are better.

¹⁶Ashihara et al., "Investigation of speaker representation for target-speaker speech processing".

Target-Speaker Tasks: Cues Depend on the Mixture

Closed-set speaker codes and mixture-specific optimization¹⁷

Speaker cue	TS-ASR WER ↓	TSE SI-SDR ↑	p-VAD mAP ↑
WavLM BASE+ enrollment	21.79	11.31	98.87
Speaker identity code	18.56	11.70	99.82

Oracle analysis

Optimizing only the code embedding for each mixture reduced clean TS-ASR WER from 18.56% to 12.60%. The preferred embedding shifted with the competing speaker, even for the same target.

Scope: clean, speaker-closed evaluation; identity codes require known speakers. The optimization used target labels and is an analysis, not an inference-time method.

¹⁷Ashihara et al., “Investigation of speaker representation for target-speaker speech processing”.

Speaker Diarization: Cues for One Recording

1. **Session adaptation:** Kwon et al. combine compression, attention aggregation, and non-speech clustering. Compression alone is inconsistent.¹⁸
2. **Task-matched evaluation:** VoxCeleb verification EER did not reliably rank diarization DER. Same-session trials were more informative.¹⁹
3. **Time resolution:** Heo et al. produce 40 frame-level embeddings from a 3.2-second segment to handle speaker changes inside a window.²⁰

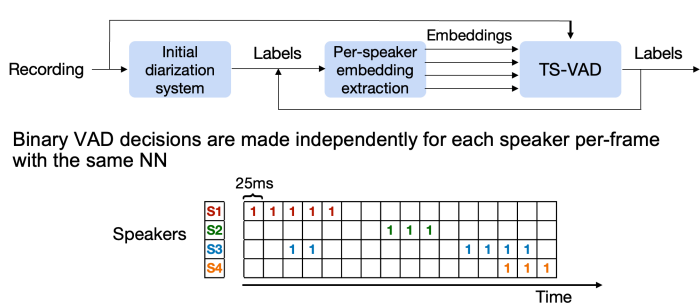
¹⁸Kwon et al., "Adapting speaker embeddings for speaker diarisation".

¹⁹Jung et al., "In search of strong embedding extractors for speaker diarisation".

²⁰Heo et al., "High-Resolution Embedding Extractor for Speaker Diarisation".

Speaker Diarization: Speaker-Conditioned Activity

TS-VAD predicts activity for each speaker over time²¹



Binary VAD decisions are made independently for each speaker per-frame with the same NN

- ▶ An initial diarization pass supplies speaker cues to a shared activity network.
- ▶ The network predicts a timeline for each speaker, including overlapping speech.

Figure courtesy of Mireia Diez Sánchez, reused from the tutorial.

²¹Medennikov et al., "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario".

Background on Speaker Modeling
Observations and Analysis Across Applications
Limitations of Current Speaker-related Tasks
Discussion on Future Directions & Guiding Principles
Conclusion

Limitation 1: Over-reliance on SR-Optimized Embeddings

The Convenience Trap

Easy to use, but often suboptimal

Why This Happens:

- ▶ Availability of pre-trained models
- ▶ Initial success in SR
- ▶ Convenience factor

The Cost:

- ▶ Suboptimal performance
- ▶ Information bottleneck
- ▶ Limited innovation
- ▶ Task mismatch

Evidence

**Embedding-free outperforms
embedding-based methods**

**Prompt based TTS > Embedding
based TTS**

Limitation 2: Inadequate Dynamic Feature Capture

The "Averaging" Problem

SR embeddings collapse diverse acoustic manifestations into single points

What's Lost:

- ▶ Emotional variations
- ▶ Prosodic patterns
- ▶ Speaking rate changes
- ▶ Stylistic nuances
- ▶ Context-dependent features

Impact on Applications:

- ▶ Monotonous TTS output
- ▶ Limited VC expressiveness
- ▶ Poor emotion control
- ▶ Unnatural prosody

Critical Question

How can we preserve intra-speaker variability while maintaining discriminability?

Limitation 3: Disentanglement Challenges

The Fundamental Complexity

Speech factors are inherently intertwined, not independently encoded

The Entanglement:

- ▶ Pitch contour: emotion + linguistic structure
- ▶ Spectral features: timbre + phonetic content
- ▶ Prosody: speaker + emotion + content
- ▶ No clear boundaries

Current Solutions:

- ▶ Adversarial training
- ▶ Mutual information minimization
- ▶ Multi-encoder architectures
- ▶ Specialized loss functions

The Reality

Perfect disentanglement remains an open research problem

Background on Speaker Modeling
Observations and Analysis Across Applications
Limitations of Current Speaker-related Tasks
Discussion on Future Directions & Guiding Principles
Conclusion

Direction 1: Task-Specific Representations

The Paradigm Shift

From universal to specialized speaker modeling

For TTS:

- ▶ Hierarchical representations
- ▶ Multi-level encoding
- ▶ Prosody-aware modeling
- ▶ Emotion-sensitive features

For VC:

- ▶ Disentanglement-focused
- ▶ Content-speaker separation
- ▶ Style transfer capabilities
- ▶ Flexible control mechanisms

For Target Speech Processing:

- ▶ Context-aware modeling
- ▶ Adaptive representations
- ▶ Raw acoustic features

Direction 2: Disentangled and Interpretable Features

The Goal

Move from opaque vectors to interpretable, controllable representations

Desired Properties:

- ▶ Independent control of attributes
- ▶ Interpretable components
- ▶ Semantic meaning
- ▶ Manipulable features

Applications:

- ▶ Voice design
- ▶ Personalized interfaces
- ▶ Assistive technologies
- ▶ Creative applications

Research Questions

- ▶ What constitutes "pure" speaker identity?
- ▶ How can we measure disentanglement quality?
- ▶ What are the fundamental limits?

Direction 3: Enhanced Robustness

The Challenge

Building models that work reliably in real-world conditions

Architectural Advances:

- ▶ Transformer-based models
- ▶ Advanced attention mechanisms
- ▶ Raw waveform processing
- ▶ Multi-scale representations

Training Strategies:

- ▶ Self-supervised pre-training
- ▶ Leveraging
linguistic/semantic/knowledge
constraints/guidance

Data Requirements

- ▶ Larger, more diverse datasets
- ▶ Real-world conditions and Challenging scenarios
- ▶ Effective, scalable data management and simulation

Direction 4: Unified vs. Specialized Frameworks

The Fundamental Debate

One model to rule them all, or specialized solutions?

Unified Approach:

- ▶ Single foundational model
- ▶ Large-scale pre-training
- ▶ Efficient fine-tuning
- ▶ Knowledge transfer

Specialized Approach:

- ▶ Task-specific models
- ▶ Optimized objectives
- ▶ Custom architectures
- ▶ Targeted solutions

Hybrid Solution

Common foundation + task-specific adapters

Principle 1: Align Objectives with Application Needs

The Foundation

Clearly define what constitutes an "ideal" representation for each task

For SR:

- ▶ Maximize discrimination
- ▶ Minimize intra-speaker variance
- ▶ Robustness to nuisance factors
- ▶ Compact representation

For TTS/VC:

- ▶ Capture rich vocal characteristics
- ▶ Preserve prosody/emotion
- ▶ Enable natural synthesis
- ▶ Support style control

Key Question

What aspects of speaker identity are most important for your application?

Principle 2: Balance Discriminative Power with Generative Flexibility

The Trade-off

Discriminative optimization vs. Generative richness

The Challenge:

- ▶ SR: Perfect discrimination
- ▶ TTS: Rich, expressive output
- ▶ VC: Flexible style transfer
- ▶ TSE: Robust extraction

Potential Solutions:

- ▶ Sub-center modeling
- ▶ Time-varying embeddings
- ▶ Multi-level representations
- ▶ Selection rather than conditioning

The Goal

Sufficiently discriminative + Generatively flexible

Principle 3: Holistic Evaluation

Beyond Standard Metrics

Comprehensive assessment of performance, robustness, and user experience

Traditional Metrics:

- ▶ EER, minDCF (SR)
- ▶ MOS, similarity (TTS/VC)
- ▶ SI-SDR (TSE)
- ▶ WER (ASR)

Holistic Assessment:

- ▶ Robustness to unseen conditions
- ▶ Fairness across demographics
- ▶ User experience quality
- ▶ Real-world performance

New Evaluation Protocols

- ▶ Challenging benchmark datasets
- ▶ Human-in-the-loop evaluations
- ▶ Multi-dimensional assessment

Background on Speaker Modeling
Observations and Analysis Across Applications
Limitations of Current Speaker-related Tasks
Discussion on Future Directions & Guiding Principles
Conclusion

The Fundamental Truth

"One Embedding Doesn't Fit All"

1. **Current Limitations:**

- ▶ Over-reliance on SR-optimized embeddings
- ▶ Inadequate dynamic & contextual feature capture
- ▶ Disentanglement challenges
- ▶ Robustness issues

2. **Future Directions:**

- ▶ Task-specific representations
- ▶ Disentangled features
- ▶ Configurable representations

3. **Strategic Approach:**

- ▶ Align objectives with applications
- ▶ Balance competing requirements
- ▶ Consider practical constraints

Theoretical Challenges

Questions that drive future research directions

1. **What constitutes "pure" speaker identity?**
 - ▶ Is perfect disentanglement possible?
 - ▶ What are the fundamental limits?
 - ▶ How do we measure disentanglement quality?
2. **Can we achieve universal speaker understanding?**
 - ▶ One model for all tasks?
 - ▶ Optimal trade-offs?
 - ▶ Knowledge transfer mechanisms?
3. **How do we evaluate speaker modeling?**
 - ▶ Beyond standard metrics?
 - ▶ Human-centric evaluation?
 - ▶ Task-specific evaluation metric?

Email: shuaiwang@nju.edu.cn

Resources

- ▶ Speaker Embedding Learning Toolkit:
<https://github.com/wenet-e2e/wespeaker>
- ▶ Target Speech Extraction Toolkit:
<https://github.com/wenet-e2e/wesep>
- ▶ Wang, Shuai, et al. "Overview of speaker modeling and its applications: From the lens of deep speaker representation learning." T-ASLP 2024.

References I

- ▶ Ahn, Junseok et al. “VoxSim: A perceptual voice similarity dataset”. In: *Interspeech 2024*. 2024.
- ▶ Ashihara, Takanori et al. “Investigation of speaker representation for target-speaker speech processing”. In: *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2024, pp. 423–430.
- ▶ Baas, Matthew, Benjamin van Niekerk, and Herman Kamper. “Voice Conversion With Just Nearest Neighbors”. In: *Interspeech 2023*. 2023, pp. 2053–2057. DOI: [10.21437/Interspeech.2023-419](https://doi.org/10.21437/Interspeech.2023-419).
- ▶ Chen, Szu-Chi et al. *Rethinking Automated Voice Similarity by Shifting from EER to Embedding Geometry*. Embedding geometry; author-provided manuscript. 2026.
- ▶ Choi, Hyeong-Seok et al. “Neural analysis and synthesis: Reconstructing speech from self-supervised representations”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 16251–16265.

References II

- ▶ Hayden, Matthew, Jinzuomu Zhong, and Korin Richmond. “Accent-Emotion Entanglement in LM-Based Text-to-Speech Systems”. In: *Interspeech 2026*. 2026. DOI: 10.21437/Interspeech.2026-2749.
- ▶ Heo, Hee-Soo et al. “High-Resolution Embedding Extractor for Speaker Diarisation”. In: *ICASSP 2023*. 2023. DOI: 10.1109/ICASSP49357.2023.10097190.
- ▶ Jia, Ye et al. “Transfer learning from speaker verification to multispeaker text-to-speech synthesis”. In: *Advances in neural information processing systems* 31 (2018).
- ▶ Jung, Jee-weon et al. “In search of strong embedding extractors for speaker diarisation”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.
- ▶ Kwak, Doyeop, Suyeon Lee, and Joon Son Chung. *Plug-and-Steer: Decoupling Separation and Selection in Audio-Visual Target Speaker Extraction*. Audio-Visual Target Speaker Extraction; author-provided manuscript. 2026.

References III

- ▶ Kwon, Youngki et al. “Adapting speaker embeddings for speaker diarisation”. In: *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. International Speech Communication Association. 2021, pp. 2493–2497.
- ▶ Medennikov, Ivan et al. “Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario”. In: *Interspeech 2020*. 2020. DOI: 10.21437/Interspeech.2020-1602.
- ▶ Shangguan, Qituan, Bosen Xu, and Shuai Wang. *OpenSVBench: Diagnosing Speaker Verification Generalization Across Deployment Scenarios*. NCMMSC 2026 manuscript. 2026.
- ▶ Sinha, Ragini et al. “Speaker-conditioning single-channel target speaker extraction using conformer-based architectures”. In: *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE. 2022, pp. 1–5.

References IV

- ▶ Stan, Adriana and Johannah O'Mahony. "An analysis on the effects of speaker embedding choice in non auto-regressive TTS". In: *12th ISCA Speech Synthesis Workshop (SSW2023)*. 2023, pp. 134–138. DOI: [10.21437/SSW.2023-21](https://doi.org/10.21437/SSW.2023-21).
- ▶ Wang, Chengyi et al. "Neural codec language models are zero-shot text to speech synthesizers". In: *arXiv preprint arXiv:2301.02111* (2023).
- ▶ Wang, Shuai et al. "Advancing speaker embedding learning: Wespeaker toolkit for research and production". In: *Speech Communication* 162 (2024), p. 103104.
- ▶ Xu, Tianze et al. "Human-like cross-language generalisation in deep neural speaker embeddings and its acoustic foundations". In: *Interspeech 2026*. 2026, pp. 2313–2318. DOI: [10.21437/Interspeech.2026-858](https://doi.org/10.21437/Interspeech.2026-858).
- ▶ Zeng, Bang and Ming Li. "Usef-tse: Universal speaker embedding free target speaker extraction". In: *IEEE Transactions on Audio, Speech and Language Processing* (2025).