One Embedding Doesn't Fit All

Rethinking Speaker Modeling for Various Speech Applications

Shuai Wang

Nanjing University

August 22, 2025

Outlines

Speaker Modeling

Speaker modeling aims to capture information related to the identity of the speaker while neglecting other attributes.

Speaker modeling in different tasks

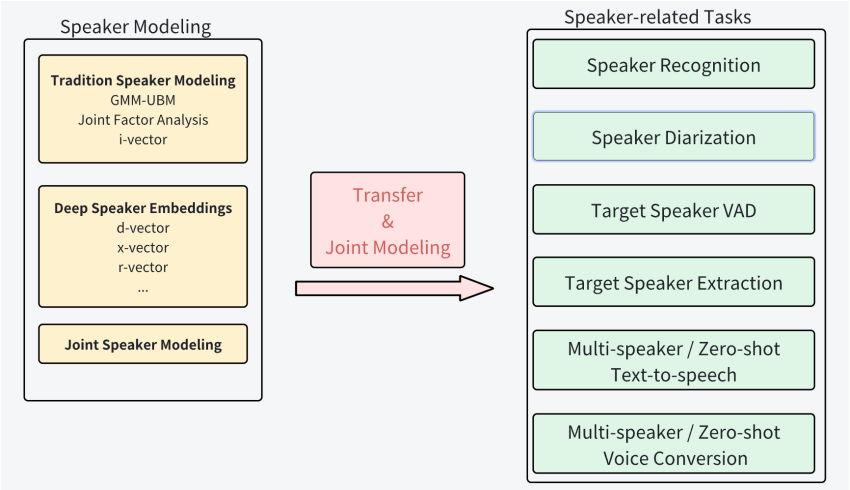


Figure from Wang et al. (2024)[1]

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

Speaker Verification: My voice is my password

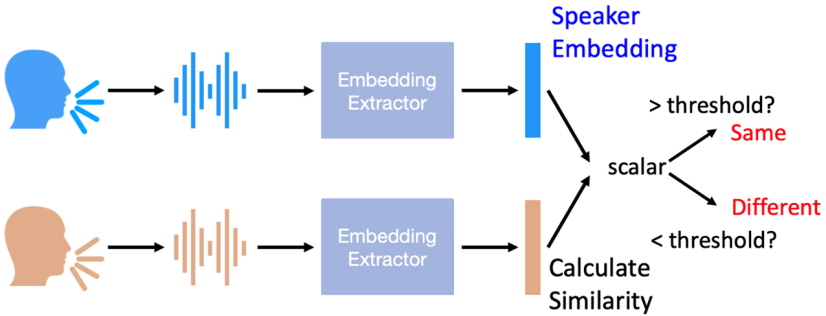


Figure from Prof. Hung-yi Lee's DLHLP20 slides[2]

[2] https://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/

Target Speech Extraction: I only hear your voice

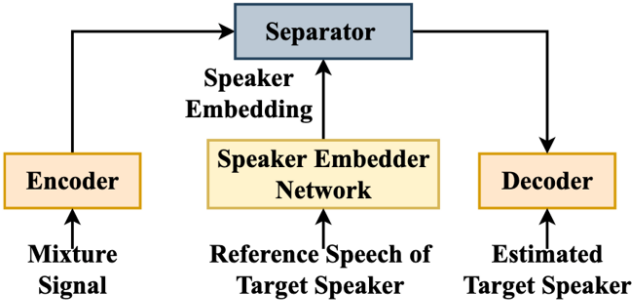
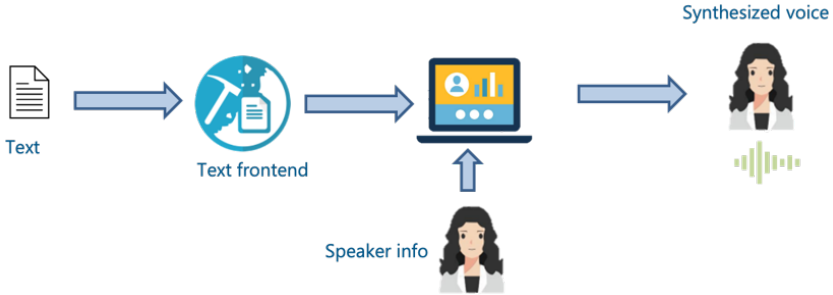


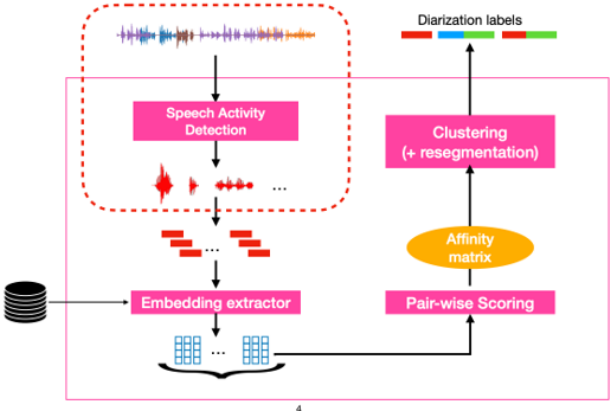
Figure from Sinha et al. (2021)[3]

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

Speech generation: generate voice for the target identity.



Speaker diarization: Who Spoke When?



Speaker Recognition: The Traditional Stronghold

Success Story

- **Architecture**: TDNNs, ECAPA-TDNN, ResNets
- **Training**: Large-scale datasets (VoxCeleb, VoxBlink)
- **Loss**: AM-softmax, AAM-softmax, GE2E, $\cdots$
- **Evaluation**: EER, minDCF

Strengths:

- Highly discriminative
- Robust to environmental factors
- Well-structured embedding space
- Strong generalization

Limitations:

- Oversmoothing of details
- Loss of prosodic information
- Emotion suppression
- Limited expressiveness

Speaker Modeling for Other Tasks

But, what about tasks except speaker recognitions?

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

$\mathbf{SR \neq TTS \neq VC \neq TSP}$

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

Outlines

Text-to-Speech: Beyond Simple Embeddings

The Challenge

Generate natural, expressive speech that sounds like the target speaker

Key Requirements:

- Timbre accuracy
- Prosodic naturalness
- Emotional expressiveness
- Speaking style preservation

Evidence:

- Custom encoders outperform SR embeddings
- Prompt based methods dominates zero-shot TTS

The Gap

SR embeddings lack:

- Dynamic features
- Prosodic details
- Emotional cues
- Style variations

Text-to-Speech: Embedding Comparison

Main Findings from Moya et al. (2023)[4]

1. **Embedding choice does not affect learning process**
   - Network adapts to speaker conditioning regardless of embedding choice
   - Produces high-quality synthesis results
2. **Speaker leakage is inevitable**
   - Core modules contain speaker information in standard training
   - Simple conditioning cannot ensure complete disentanglement
3. **Zero conditioning output is inconsistent**
   - Core network learns similar representations
   - Speaker identity unstable under zero conditioning

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

Voice Conversion: The Disentanglement Challenge

The Fundamental Problem

$$\text{Speech} = \text{Speaker Identity} + \text{Content} + \text{Prosody} + \text{Accent} + \text{Emotion} + \cdots$$

Current Approaches:

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

Challenges:

- Perfect disentanglement impossible
- Residual speaker information
- Content-speaker entanglement
- Prosody control complexity

Open Question

Can we ever achieve perfect disentanglement?

Target Speaker Extraction: The Rise of Embedding-Free

Paradigm Shift

From pre-computed embeddings to adaptive, context-aware modeling

Traditional Approach:

- Pre-trained speaker embeddings
- Fixed representation
- Limited context awareness
- Performance bottlenecks

Modern Approaches:

- USEF-TSE (embedding-free)[a]
- Attention mechanisms
- Multi-level representations[b]
- SSL-based features

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

[b] Zhang et al., "Multi-level speaker representation for target speaker extraction".

Key Insight

Direct acoustic matching can outperform abstract embeddings

Target-Speaker Speech Processing: The Effectiveness of SSL Models

Main Findings from Ashihara et al. (2024)[5]

1. **SSL Models Superiority**: SSL models (especially WavLM and ECAPA-TDNN-DINO) outperform supervised speaker models across all TS tasks
2. **ASV-TS Disconnect**: Speaker verification performance is uncorrelated with TS task performance
3. **Speaker Code Effectiveness**: One-hot speaker codes provide optimal representation (speaker-closed setting)
4. **Optimization Potential**: Gradient-based optimization reveals significant room for improvement (32% WER reduction)

[5] Ashihara et al., "Investigation of speaker representation for target-speaker speech processing".

Speaker Diarization: Towards Customized Speaker Representations

Main Findings from Kwon et al. (2021)[6] and Jung et al. (2023)[7]

1. **Dimention Reduction Helps**: Speaker-verification embeddings are over-parameterized for diarisation; 256→20-D cut cleans noise
2. **ASV-Diarisation Gap**: Conventional EER on VoxCeleb1-O negatively correlates with DER; proposed intra-session verification protocol realigns EE selection with diarisation need.
3. **Frame-level Helps**: Replacing global pooling with conformer-based enhancer → 40 frame-level embeddings per 3.2 s segment, beating 1-embedding baselines by 10 % DER on four datasets

[6] Kwon et al., "Adapting speaker embeddings for speaker diarisation".

[7] Jung et al., "In search of strong embedding extractors for speaker diarisation".

Outlines

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

USEF-TSE 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

Outlines

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 linguistc/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
- Conditional generation

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

Outlines

Key Takeaways

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

Fundamental Open Questions

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?

Contact

Email: shuaiwang@nju.edu.cn

Resources

- Speaker Embedding Learning Tookit: `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

- 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.
- 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 Proces IEEE. 2023, pp. 1–5.
- Kwon, Youngki et al. "Adapting speaker embeddings for speaker diarisation". In: 22nd Annual Conference of the International Speech Communication Association, INTERS International Speech Communication Association. 2021, pp. 2493–2497.
- 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 II

- 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.
- Wang, Shuai et al. "Advancing speaker embedding learning: Wespeaker toolkit for research and production". In: Speech Communication 162 (2024), p. 103104.
- Zeng, Bang and Ming Li. "Usef-tse: Universal speaker embedding free target speaker extraction". In: IEEE Transactions on Audio, Speech and Language Processing (2025).
- Zhang, Ke et al. "Multi-level speaker representation for target speaker extraction". In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Proces IEEE. 2025, pp. 1–5.