Training and Data Strategies for Target Speaker Extraction

LIU YUN

Doctor of Philosophy



Department of Informatics

School of Multidisciplinary Sciences

The Graduate University for Advanced Studies, SOKENDAI

March 2026

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Junichi Yamagishi from National Institute of Informatics (NII) and SOKENDAI, for his continuous guidance, insightful feedback, and generous support throughout my doctoral study. His encouragement and scientific rigor have been invaluable in shaping both my research and my way of thinking as a scholar.

I would also like to sincerely thank the examiners of this thesis: Prof. Tomoki Toda (Nagoya University), Prof. Yu Tsao (Academia Sinica), Prof. Imari Sato (NII), and Prof. Satoshi Ikehata (NII), for their thoughtful reviews, valuable comments, and constructive suggestions, which have greatly helped improve the quality of this work.

My heartfelt thanks go to my co-authors, Xuechen Liu (NII), Haoyu Li (Indeed), and Xiaoxiao Miao (Duke Kunshan University), for their collaboration and inspiring discussions. Their ideas and advice have contributed significantly to the development of this research.

I am also grateful to NII and SOKENDAI for providing the SOKENDAI Special Researcher Fellowship (JPMJFS2136), which offered important financial support during my doctoral study.

Finally, I would like to thank all my friends and colleagues who have accompanied me during this journey. My Ph.D. life has been a truly meaningful experience, filled with learning, growth, and unforgettable memories. Most importantly, I would like to express my deepest gratitude to my parents for their love, support, and encouragement throughout my life.

Abstract

Target speaker extraction (TSE) aims to isolate the voice from a specific speaker from complex acoustic mixtures containing multiple speakers and background noise. Despite recent advances in deep neural network approaches, current TSE systems face significant challenges under real-world environments, primarily due to limited speaker diversity in training data and the gap between artificially constructed training conditions and authentic acoustic scenarios. This thesis addresses these fundamental limitations through a comprehensive framework combining curriculum learning strategies with synthetic speaker generation.

The thesis makes four key contributions. First, we identify and evaluate multiple difficulty measures for TSE training data, demonstrating that speaker similarity based on cosine distance between embeddings provides the most reliable indicator of extraction difficulty. Second, we develop a multi-stage curriculum learning framework that progressively trains models from easier to more challenging examples, achieving up to 0.97 dB improvement in signal-to-distortion ratio compared to conventional random sampling. Third, we leverage speech generative models to create valuable synthetic speakers through voice conversion techniques as additional training data, and the optimal performance was achieved using a 50%-50% balance between real and synthetic speakers. Last but not least, we introduce Libri2Vox, a novel dataset that bridges the gap between controlled experimental conditions and real-world complexity by combining clean speech from LibriTTS with recordings from VoxCeleb2, which contains diverse acoustic conditions and encompasses totally over 7,000 speakers.

Experimental evaluations across multiple modern and state-of-the-art TSE architectures (Conformer, BLSTM, SpeakerBeam, and VoiceFilter) demonstrate consistent improvements ranging from 0.5 to 2.2 dB in extraction performance. Notably, models

trained with our approach show substantial improvements on real-world recordings, achieving positive quality gains where conventional methods yield degraded performance. Throughout this thesis, the integration of curriculum learning, synthetic speaker augmentation, and realistic acoustic conditions significantly enhances model robustness and generalization capability.

This thesis establishes that strategic training data utilization through curriculum design and synthetic augmentation can substantially improve TSE systems. The proposed methods are architecture-agnostic and can be readily integrated into existing TSE training pipelines. These findings have broader implications for speech processing tasks facing similar challenges of limited training data and domain mismatch, contributing to the development of more robust speech technologies for real-world deployment.

Contents

List of Figures xiii

List of Tables xv

1 Introduction 5

1.1 Background . . . 5

1.2 Research Questions and Contributions . . . 6

1.2.1 RQ1: How can we identify and select useful training data for more effective TSE training? . . . 7

1.2.2 RQ2: How do we use the selected data at different training stages to improve TSE performance? . . . 7

1.2.3 RQ3: Can we use speech generative models to generate useful data for training TSE? . . . 8

1.2.4 RQ4: How can we develop training strategies that coordinate multiple complexity factors in challenging multi-speaker scenarios? . . . 8

1.3 Societal Impact of Improved TSE Systems . . . 9

1.4 Thesis Organization . . . 9

2 Overview of Related Technologies 13

2.1 Speech Front-end Foundations . . . 13

2.1.1 Definitions of Speech Front-end Task . . . 13

2.1.2 Historical Development of Speech Enhancement . . . 15

2.1.3 Traditional Speech Enhancement Methods . . . 16

2.1.4 Neural Network-based Methods . . . 17
2.1.5 Limitations of Conventional Speech Enhancement and Speech Separation . . . 19
2.2 Target Speech Extraction . . . 20
2.2.1 TSE Framework . . . 20
2.2.2 Traditional TSE methods . . . 20
2.2.3 Early Neural Network Methods for TSE . . . 21
2.3 Neural Network for Target Speaker Extraction . . . 23
2.3.1 Speaker Encoder Based on ECAPA-TDNN . . . 24
2.4 Evaluation Metrics . . . 25
2.4.1 Common Evaluation Metrics . . . 25
2.5 Standard Training Approaches . . . 27
2.6 Data Augmentation Techniques . . . 29
2.6.1 Source-preserving Augmentation Approaches . . . 29
2.6.2 Non-source Preserving Approaches and Performance Analysis 30
2.7 Datasets and Benchmark . . . 31
2.7.1 State-of-the-art Performance . . . 36
2.8 Limitations of Current Approaches . . . 36
2.8.1 Acoustic and Speaker Diversities . . . 36
2.8.2 Training Strategy Limitations . . . 37

3 Curriculum Learning for Target Speaker Extraction 39
3.1 Curriculum Learning . . . 39
3.1.1 Motivations of CL for TSE . . . 42
3.2 Error analysis of the neural network-based TSE . . . 43
3.2.1 Speaker-Dependent Factors . . . 45
3.2.2 Mixture-Dependent Factors . . . 47
3.2.3 Dynamic Difficulty Measure . . . 48
3.3 Experimental Setup . . . 48
3.3.1 Model Configurations . . . 48
3.3.2 Model Configuration . . . 49
3.3.3 Curriculum Strategy Implementation . . . 49
3.4 Experiments and Results . . . 50

3.4.1 Dataset . . . 50
3.4.2 Experimental Conditions . . . 50
3.4.3 Results . . . 52
3.4.4 Discussion: knowledge-driven vs. data-driven curriculum design. 56
3.5 Conclusion and Future Work . . . 57

4 Curriculum Learning with Synthetic Speakers for Target Speaker Extraction 59
4.1 Introduction . . . 59
4.2 Background and Motivation . . . 60
4.2.1 Limitations of Existing Datasets . . . 60
4.2.2 Synthetic Speaker Generation for Speech Processing . . . 60
4.2.3 Motivations . . . 63
4.3 Synthetic Speaker Augmentation for CL Based TSE Training . . . 64
4.3.1 Main Concept . . . 64
4.3.2 Synthetic Speaker Generation Based on Voice Conversion . . . 65
4.3.3 Generation progress of synthetic interference speakers . . . 66
4.4 Experimental Setup . . . 67
4.4.1 Dataset . . . 67
4.4.2 Feature Extraction and Model Configuration . . . 67
4.4.3 Voice Conversion Setup . . . 68
4.4.4 Training Protocol of TSE . . . 68
4.5 Results and Analysis . . . 69
4.5.1 Improvement from Synthetic Speakers . . . 69
4.5.2 Impact of Real-Synthetic Speaker Ratio . . . 69
4.5.3 Effect of Voice Conversion Hyperparameters . . . 70
4.5.4 Impact of Synthetic Data Quantity . . . 71
4.5.5 Generalization Across TSE Architectures . . . 71
4.5.6 Visualization for Speaker Embedding Space . . . 72
4.6 Discussion . . . 73
4.7 Conclusion . . . 74

5 Libri2Vox : A Novel Dataset for Target Speaker Extraction 77
5.1 Introduction . . . 77
5.2 Background and Motivation . . . 79
5.2.1 Limitations of Existing Datasets . . . 79
5.3 Libri2Vox Dataset . . . 80
5.3.1 Data construction . . . 81
5.3.2 Statistics of Libri2Vox . . . 84
5.4 Synthetic Libri2Vox Dataset . . . 85
5.4.1 Why synthetic speakers? . . . 85
5.4.2 Synthetic interference speaker generation . . . 85
5.4.3 Synthetic data and statistics . . . 87
5.5 Experiments . . . 87
5.5.1 Experimental setup and model configurations . . . 88
5.5.2 Training strategy . . . 88
5.6 Results and Discussion . . . 89
5.6.1 Impact of Libri2Vox . . . 89
5.6.2 Downstream ASR evaluation. . . . 92
5.6.3 Evaluation on real-world recordings . . . 93
5.7 Ablation study regarding synthetic data . . . 95
5.8 Ablation study regarding datasets . . . 97
5.8.1 Cross-dataset evaluation . . . 97
5.8.2 Two-stage CL with multiple datasets . . . 99
5.9 Conclusion . . . 100

6 Training Dynamics-Aware Multi-Factor Curriculum Learning for Target Speaker Extraction **103**
6.1 Introduction . . . 103
6.2 Multi-Factor Complexity in TSE . . . 105
6.2.1 TSE-Datamap: A Training Dynamics Framework . . . 106
6.3 Experimental Evaluation . . . 108
6.3.1 Dataset and Training Configuration . . . 108
6.4 Results and Analysis . . . 110
6.4.1 Analysis of Single-Factor vs. Multi-Factor . . . 110

6.4.2 TSE-Datamap Analysis . . . 113
6.4.3 Fixed-Quantity Ablation of Datamap Regions . . . 114
6.5 Conclusion . . . 115

7 Conclusion and Future Work 117
7.1 Summary of Research Contributions . . . 117
7.1.1 Addressing the Research Questions . . . 117
7.1.2 Chapter-wise Contributions . . . 119
7.2 Limitations and Future Work . . . 120
7.2.1 Enhanced Difficulty Measures . . . 120
7.2.2 Advanced Synthetic Speaker Generation . . . 121
7.2.3 Real-world Deployment Considerations . . . 121
7.2.4 Theoretical Understanding . . . 122
7.2.5 Extension to Complex Scenarios . . . 122
7.2.6 Emotional speech and other expressive speaking styles. . . . 123

Bibliography 125

A List of Publications 147
A.1 Publications For This Thesis . . . 147

B Online Resources 149

List of Figures

1.1 Details of target speaker extraction. 6

1.2 Thesis outline. 10

2.1 Difference between speech enhancement, speech separation, and target speaker extraction, in terms of source and target objectives. 14

2.2 Difference between mapping modeling, masking modeling, and waveform modeling. 18

2.3 An unified illustration of the details of the TSE networks applied for this thesis: Conformer, BLSTM, and SpeakerBeam. Reference speech is processed through ECAPA-TDNN to extract speaker embeddings, which condition both Conformer and BLSTM models. The SpeakerBeam module processes raw waveforms directly. The system generates real and imaginary masks that are applied to the mixture representation before inverse transformation to produce the estimated target speech, which is optimized against ground truth targets during training. 22

3.1 Scatter plot of cosine similarity vs SNR with normal distributions. If the TSE system successfully extracts the target speaker, the SNR should be positive. 43

3.2 Cumulative density function of prediction results from the TSE model based on speaker similarity. Higher CDF values indicate a greater proportion of samples achieving the corresponding iSDR (dB). "Threshold" corresponds to τ_{spk}, and "Random Select" is the baseline. 52

3.3 iSDR changes on the dev set with varying τ_{SDR} in two training stages, along with the data usage percentage in Stage 1. 53

4.1 Illustration of the proposed three-stage CL approach for TSE covered by Chapter 4. Stage 1 focuses on examples with low speaker similarity, Stage 2 introduces high similarity pairs, and Stage 3 incorporates synthetic interference speakers. 64
4.2 Synthetic speaker generation using the k-NN VC / SALT system. The framework consists of an WavLM encoder that extracts latent representations, a k-NN regression module that performs neighbor selection and weighted averaging, and a HiFi-GAN vocoder that generates the synthetic speech. 65
4.3 iSDR performance as a function of the proportion of synthetic speakers within a mini-batch. An x-axis value of 1.0 indicates the exclusive use of synthetic interference speakers. 70
4.4 Speaker Embedding Space Visualization. 73

5.1 Data generation framework of Libri2Vox and its synthetic version. . . . 82
5.2 Impact of synthetic speaker ratio within one mini-batch at Stage 3 of configuration “w/ 3-stage CL (Real + SALT)" on Conformer. Red dashed line corresponds to performance (7.17 dB) where training started from scratch using synthetic data only. 95

6.1 TSE-Datamap computed. Each point denotes a sample, with x-axis the standard deviation of ΔSNR and y-axis the mean ΔSNR. Number of training epochs in this study is 50. 108

List of Tables

2.1 Comparison of Libri2Mix, Libri2Talker, and LibriTTS datasets. 34
2.2 Comparison of VoxCeleb and CN-Celeb datasets. 35

3.1 Dataset and its partitions for this study. No overlap exists between the test partition and the other two partitions at either the utterance or speaker level. 50
3.2 iSDR(dB) results on the test set using the CL-based TSE. “Used data" shows the percentage of the training data utilized in the 1st phase. The 1st phase results were obtained from systems trained using only “easy" samples. The 2nd phase results were from systems additionally trained on all samples based on the model weights from the 1st phase. 51
3.3 iSDR(dB) results on the test set using the self-paced CL approach with 5 training phases, as described in Section 4.2. 55
3.4 iSDR(dB) of CL methods with different network architectures. 55

4.1 Experimental results before and after adding Stage 3, which uses synthetic speakers, to the TSE model based on Conformer in [1]. Hyperparameters are k=4 and p=0.5. 69
4.2 iSDR (dB) results of methods with varying parameters k and p. 70
4.3 Results with different numbers of synthetic speakers, in terms of its resulting size of the data and iSDR. 71
4.4 iSDR results using different TSE architectures. Hyperparameters for speaker generation are k=4 and p=0.5. 72

5.1 Comparison of WSJ0-2mix-extr, Libri2mix, and Libri2talker Datasets. . 81

5.2 Libri2Vox statistics. "Total" column shows the total number of utterances combined with corresponding LibriTTS and VoxCeleb2 sets. . . . 84

5.3 Table 5.3: iSDR(0.07 dB), iPESQ(1.514), iSTOI(0.734), iDNSMOS(2.356), results of three stage methods for Libri2Vox test set. Values are reported as mean ± 95% confidence interval half-width. values in '()' represent baseline performance metrics of original mixture signals compared **WITH** the clean target. "Real only" means use of only real data with cosine similarity less than 0.5 in 1st stage (about 71% of all data). 90

5.4 Downstream ASR performance on the Libri2Vox test set using faster-whisper large-v3 (beam size 5), reported as mean WER (%). Lower is better. We compare the TSE-baseline trained w/o CL (Real only) and the best 3-stage curriculum model from Table 5.3, w/ 3-stage CL (Real + SynVox2). 92

5.5 DNSMOS results on real-world recordings from ICASSP 2021 DNS Challenge test set. Higher values indicate better performance. (Mixture DNSMOS:2.270) 94

5.6 Comparison of different numbers of synthetic speakers. "Stage 3" and "Stage 4" columns indicate which synthetic data types were used in each stage, while "-" indicates no additional data were used in that stage. Each dataset maintains same size to ensure consistent experimental conditions. 97

5.7 Experimental results of cross-database comparisons in iSDR(dB). 98

5.8 Performance comparison (iSDR in dB) in two-stage CL with different test set using multiple training datasets. TSE model was BLSTM. 99

6.1 Impact of single-factor and multi-factor CL. Each row represents a different curriculum configuration with varying factors. The curriculum stages column shows the progression of training difficulty across three stages, with parameters in the format [*num_speakers*, *snr_low*,*snr_high*, *overlap_ratio*, *inter_source*]. *num_speakers*: number of interfering speakers, [*snr_low*, *snr_high*]: SNR range in dB, *overlap_ratio*: fraction of temporal overlap, ranged from 0.0 to 0.4, *inter_source*: whether interference comes from real or synthetic (syn) speakers. 111

6.2 Performance comparison of different stage orderings in CL. Each stage takes one third of total epochs. Baseline uses random sampling without curriculum. 113

6.3 iSDR (dB) under fixed data quantity drawn from Datamap regions. All settings use 70% of the training set for equal-quantity comparison. . . . 114

List of Abbreviations and Mathematical Symbols

List of Abbreviations

Speech Processing Terminology

- **ASR** - Automatic Speech Recognition
- **BLSTM** - Bidirectional Long Short-Term Memory
- **CNN** - Convolutional Neural Network
- **CRM** - Complex Ratio Mask
- **DNN** - Deep Neural Network
- **DNSMOS** - Deep Noise Suppression Mean Opinion Score
- **GMM** - Gaussian Mixture Model
- **HMM** - Hidden Markov Model
- **IBM** - Ideal Binary Mask
- **IRM** - Ideal Ratio Mask
- **LSTM** - Long Short-Term Memory
- **PESQ** - Perceptual Evaluation of Speech Quality

- **PIT** - Permutation Invariant Training
- **RNN** - Recurrent Neural Network
- **SDR** - Signal-to-Distortion Ratio
- **SE** - Speech Enhancement
- **SI-SDR** - Scale-Invariant Signal-to-Distortion Ratio
- **SNR** - Signal-to-Noise Ratio
- **SS** - Speech Separation
- **STFT** - Short-Time Fourier Transform
- **STOI** - Short-Time Objective Intelligibility
- **TSE** - Target Speaker Extraction
- **VC** - Voice Conversion

Research method

- **CL** - Curriculum learning
- **MSE** - Mean Squared Error
- **RQ** - Research Question
- **SALT** - Speaker Anonymization via Latent space Transformation
- **SynVox2** - Synthetic Voice system using orthogonal Householder neural networks
- **Conv-TasNet** - Convolutional Time-domain Audio Separation Network
- **DNS** - Deep Noise Suppression (Challenge Dataset)
- **LibriMix** - Speech separation dataset based on LibriSpeech
- **VoxCeleb** - Large-scale speaker identification dataset
- **WSJ0-2mix** - Two-speaker mixture dataset based on Wall Street Journal corpus

Mathematical Symbols

Time Domain Representation

n - n-th sample of the time-domain speech signal

$y(n)$ - mixture speech signal

$s_0(n)$ - Target speech signal

$s_i(n)$ - Interference speech signal

$u(n)$ - Environmental noise

$\hat{s}(n)$ - Estimated target speech signal

Frequency Domain Representation

$Y(t,f)$ - Time-frequency representation at time t and frequency f

$S(t,f)$ - Clean speech spectrogram

$\hat{S}(t,f)$ - Estimated clean speech spectrogram

$M(t,f)$ - Time-frequency mask

$\xi_{t,f}$ - A priori SNR at time t and frequency f

Loss Functions and Metrics

$\mathcal{L}(\hat{s}, s)$ - General loss function

$\mathcal{L}_{\text{PIT}}$ - Permutation Invariant Training loss

$\mathcal{L}_{\text{SNR}}$ - Signal-to-Noise Ratio loss

$\mathcal{L}_{\text{SI-SNR}}$ - Scale-Invariant Signal-to-Noise Ratio loss

Mathematical Operations

$\|\cdot\|$ - Euclidean norm (L2 norm)

$\langle \cdot, \cdot \rangle$ - Inner product

$\min(\cdot)$ - Minimum function

$\log_{10}(\cdot)$ - Base-10 logarithm

π - Permutation of speaker indices

$\mathcal{P}$ - Set of all possible permutations

Neural Network Notation

$f_\theta(\cdot)$ - Neural network function with parameters θ

V - Embedding matrix in deep clustering

$E(\cdot)$ - Speaker encoder function

$\mathcal{D}$ - Dataset

$\mathcal{B}$ - Batch of training examples

$\mathcal{S}_{\text{target}}$ - Set of target speakers

$\mathcal{S}_{\text{int}}$ - Set of interference speakers

$\mathcal{N}$ - Set of background noise samples

1 Introduction

1.1 Background

Target speaker extraction (TSE) is a foundational technology in speech processing that addresses the challenge of isolating a specific speaker's voice from complex acoustic environments containing multiple competing voices and background noise. Unlike conventional speech separation approaches that attempt to segregate all audio sources simultaneously, TSE uses auxiliary information about the target speaker, typically represented by a reference speech sample, to selectively extract only the desired voice of the target speaker. This technology has become increasingly important in telecommunications, hearing assistive devices, and as a front-end processor for speech recognition systems operating in multi-speaker environments.

Despite recent advancements through deep neural network (DNN) approaches, TSE systems continue to face challenges in real-world conditions. While DNN-based models have shown promising performance on controlled datasets with heavily-controlled, laboratory-like acoustic conditions (e.g., clean recordings, minimal background noise,

limited reverberation, and a deliberately selected single speaker), their effectiveness often deteriorates when confronted with unseen speakers, diverse acoustic environments, or complex overlapping speech patterns. This gap between laboratory performance and real-world robustness stems from several limitations such as restricted speaker diversity in standard training datasets, suboptimal training strategies that fail to address the inherent complexity of the task, and insufficient exploration of synthetic data utilization to expand training distributions.

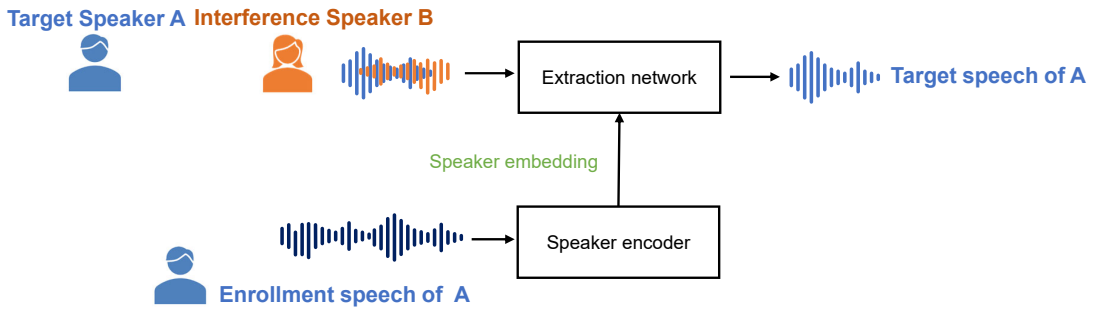


Figure 1.1: Details of target speaker extraction.

Meanwhile, speech signal in real-world environments is degraded by numerous factors beyond simple additive noise, including reverberation, nonlinear distortions, and varying degrees of speaker similarity that make extraction particularly challenging. As teleconferencing, voice-controlled systems, and speech-based interfaces become increasingly common in daily life, developing more robust and generalizable TSE systems has become a practical necessity as an essential component of these applications.

1.2 Research Questions and Contributions

This thesis systematically addresses the core limitations in current TSE approaches by exploring the following research questions and their corresponding contributions:

1.2.1 RQ1: How can we identify and select useful training data for more effective TSE training?

Challenge: Current TSE training approaches predominantly rely on random sampling from the entire dataset, without considering the varying difficulty levels of different training examples. This results in models that struggle with challenging cases, such as when target and interference speakers share similar vocal characteristics.

Contributions: I propose and evaluate a novel curriculum learning (CL) framework for TSE that strategically schedules training examples based on multiple difficulty measures including gender, speaker similarity, signal-to-distortion ratio (SDR), and signal-to-noise ratio (SNR). The proposed method also includes a novel self-paced learning mechanism that dynamically selects training samples based on the current model capabilities. This approach to difficulty assessment enables more structured learning progressions, improving model convergence and performance on complex scenarios across multiple TSE architectures compared to conventional training approaches (Chapter 3).

1.2.2 RQ2: How do we use the selected data at different training stages to improve TSE performance?

Challenge: The order in which training examples are presented to the model greatly affects learning outcomes, but current TSE methods randomly mix easy and hard examples throughout training instead of using a structured learning progression.

Contributions: I develop a progressive learning approach that divides the training process into multiple stages with increasing complexity. The model first establishes foundational skills with "easier" examples (e.g., different-gender speakers or low speaker similarity pairs), then gradually incorporates more challenging scenarios (e.g., same-gender speakers or high speaker similarity pairs). This staged training method ensures the model builds robust speaker separation capabilities before tackling more difficult cases, leading to improved generalization across diverse extraction scenarios. Experimental validation shows that this multi-stage CL strategy significantly enhances extraction performance while maintaining training stability (Chapter 3).

1.2.3 RQ3: Can we use speech generative models to generate useful data for training TSE?

Challenge: Limited speaker diversity in standard TSE datasets constrains model generalization to unseen speakers with diverse acoustic characteristics. Existing datasets typically contain fewer than 1,200 unique speakers, which is insufficient for robust speaker representation learning, in contrast to speaker recognition systems trained on databases with thousands of speakers.

Contributions: I develop a comprehensive framework for synthetic speaker generation using advanced voice conversion and speaker anonymization techniques. Specifically, I implement two complementary approaches: (1) a k-nearest neighbor-based voice conversion method that manipulates the latent space of pre-trained speaker representations, and (2) an orthogonal Householder neural network designed for speaker anonymization. These methods generate diverse synthetic interference speakers that expand the variability of training data without requiring additional real recordings. The synthetic speakers are strategically incorporated into the training process through a CL framework, enabling the model to acquire more robust speaker separation capabilities from a richer and more diverse set of training examples (Chapter 4&5).

1.2.4 RQ4: How can we develop training strategies that coordinate multiple complexity factors in challenging multi-speaker scenarios?

Challenge: Most existing TSE frameworks handle individual complexity factors in isolation, failing to address the complex interactions between multiple difficulty dimensions that characterize real-world multi-speaker environments. Current approaches cannot effectively coordinate factors such as SNR, number of interfering speakers, temporal overlap, and speaker characteristics simultaneously.

Contributions: I develop a comprehensive multi-factor CL framework that coordinates multiple complexity dimensions simultaneously, moving beyond the single-factor approaches explored in earlier chapters. This framework introduces TSE-Datamap, which grounds curriculum design in empirical training dynamics rather

than predetermined difficulty assumptions, categorizing training examples based on observed model learning behaviors into easy-to-learn, ambiguous, and hard-to-learn regions. The experimental results demonstrate that coordinated multi-dimensional progression significantly outperforms single-factor optimization, with the framework successfully integrating CL foundations, synthetic speaker generation, and realistic acoustic conditions into a unified training strategy. This contribution addresses the fundamental limitation where complexity factors were handled independently, providing effective strategies for managing the compound challenges of real-world multi-speaker scenarios (Chapter 6).

1.3 Societal Impact of Improved TSE Systems

Advancements in TSE have notable societal implications across multiple domains. For individuals with hearing impairments, improved TSE can enhance hearing aid functionality by selectively amplifying speech from a specific speaker in noisy environments, substantially improving communication capabilities and quality of life. In teleconferencing and remote work settings, robust TSE enables clearer communication in challenging acoustic environments, facilitating more effective collaboration across distributed teams. Moreover, TSE serves as an enabling technology for more accurate automatic speech recognition in multi-speaker environments, which is critical for applications ranging from meeting transcription to voice-controlled systems in smart homes and vehicles. By extracting clean speech signals from complex acoustic environments, TSE contributes to more accessible and inclusive speech technologies that can function reliably in real-world conditions. The development of more generalizable TSE models using synthetic data and improved training scheme also has broad implications for reducing the data collection burden, potentially enabling faster development cycles and more privacy-preserving approaches to speech processing system creation.

1.4 Thesis Organization

The remainder of this thesis is organized as follows:

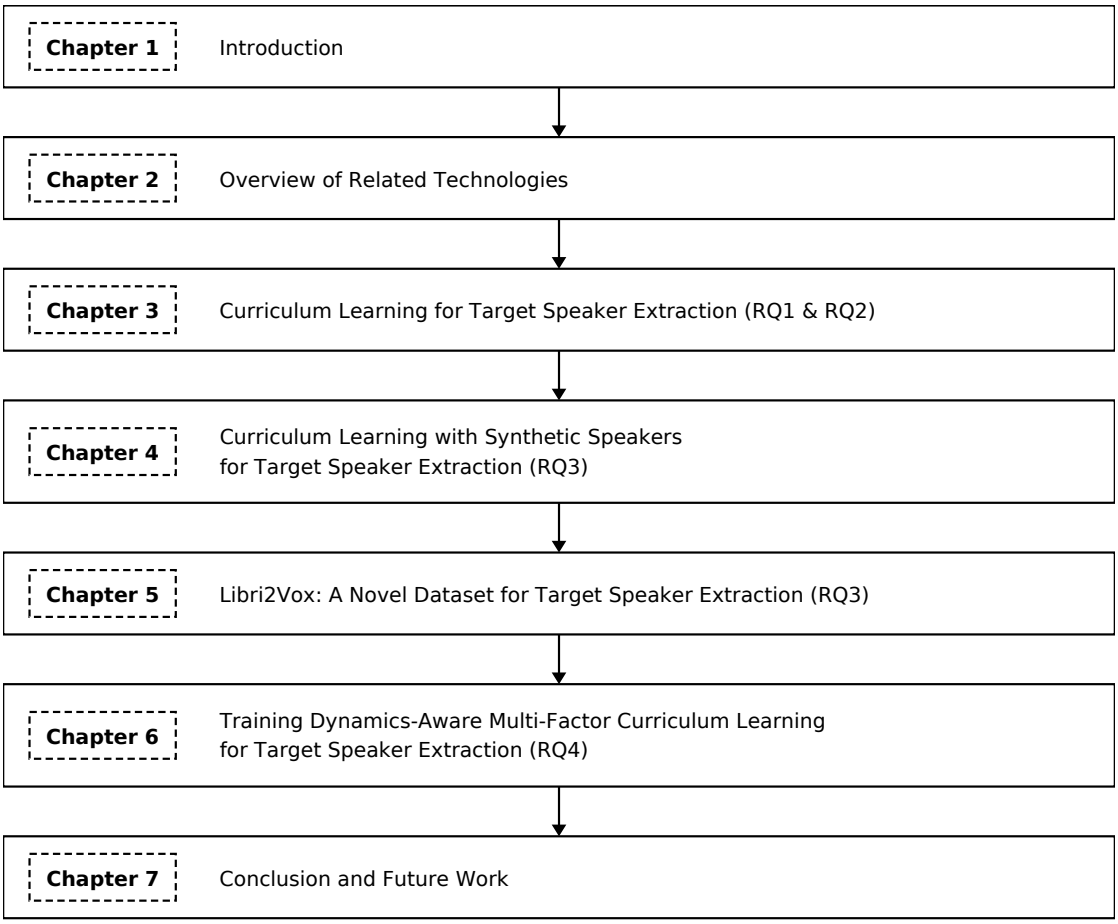


Figure 1.2: Thesis outline.

Chapter 2 provides a comprehensive overview of related technologies, including speech enhancement, speech separation, and target speaker extraction. It reviews fundamental approaches, neural network architectures, evaluation strategy, and existing datasets.

Chapter 3 introduces the proposed CL framework for target speaker extraction, demonstrating how progressive training with increasingly difficult examples improves model performance. We explore various difficulty measures and evaluate their effectiveness across different TSE architectures.

Chapter 4 addresses the challenge of limited speaker diversity by exploring speaker anonymization and voice conversion. It enhances TSE by using synthetic speakers with CL. By generating diverse synthetic voices and implementing a three-stage training approach, the system gradually learns from easy to complex examples, significantly improving performance across various model architectures.

Chapter 5 introduces Libri2Vox, a new dataset combining studio-recorded clean speech with real-world noisy speech to enhance target speaker extraction. This work augmented it with synthetic speakers and implemented CL, significantly improving model performance across multiple architectures.

Chapter 6 presents a training dynamics-aware multi-factor CL framework that coordinates multiple complexity dimensions in challenging multi-speaker scenarios, integrating the CL foundations, synthetic speaker generation, and realistic acoustic conditions from previous chapters into a unified training strategy.

Chapter 7 concludes the thesis by summarizing the key findings, discussing limitations, and outlining promising directions for future research in target speaker extraction.

2

Overview of Related Technologies

This chapter provides a comprehensive overview of TSE technologies, establishing the conceptual framework for our research. We review the principal network architectures employed in modern DNN-based TSE systems, analyze standard training frameworks and their inherent limitations, and discuss evaluation metrics that quantify extraction performance. The chapter establishes the technical foundation for our subsequent contributions in optimizing training strategies and expanding speaker diversity to enhance model robustness across challenging acoustic environments.

2.1 Speech Front-end Foundations

2.1.1 Definitions of Speech Front-end Task

Given a time-domain mixture speech signal $y(n) = s_0(n) + \sum_{i\neq 0} s_i(n) + u(n)$, where $s_0(n)$ represents the target speech, $\sum_{i\neq 0} s_i(n)$ represents interference speech, and $u(n)$ represents environmental noise, we can formally define the following key speech

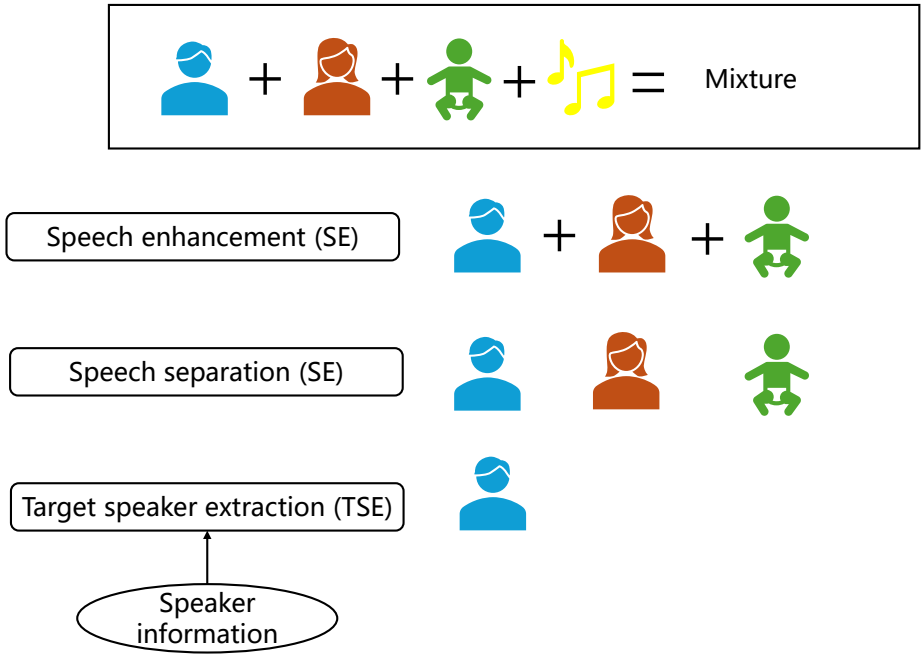


Figure 2.1: Difference between speech enhancement, speech separation, and target speaker extraction, in terms of source and target objectives.

processing technologies[1][2]:

Speech Enhancement (SE) aims to estimate clean speech signal $\hat{y}(n) = \hat{s}_0(n) + \sum_{i \neq 0} \hat{s}_i(n)$ that maximizes both quality and intelligibility:

$$\hat{y}(n) = \mathrm{SE}(y(n)) \tag{2.1}$$

Speech Separation (SS) attempts to recover each individual speech component within the mixture:

$$\hat{s}_i(n) = \mathrm{SS}_i(y(n)) \quad \text{for} \quad i \in \{0, 1, 2, ..., C-1\} \tag{2.2}$$

where C represents the total number of speakers in the mixture.

Target Speaker Extraction (TSE) utilizes a reference utterance $r(n)$ from the

[1]We do not consider echo and reverberation here.
[2]For simplicity, the time index n is omitted in the following thesis.

target speaker to extract only the speech of the specific target speaker:

$$\hat{s}_0(n) = \text{TSE}(y(n), r(n)) \tag{2.3}$$

The distinct objectives and outputs of these technologies are visually represented in Figure 2.1. The top portion depicts the composition of an acoustic mixture containing multiple speakers and background noise. SE aims to preserve all speech components while removing non-speech elements, resulting in a cleaner combined speech signal that still contains all speakers. SS advances this capability by isolating individual speakers, producing separate output channels for each speaker present in the mixture. TSE represents a specialized approach that leverages auxiliary speaker information (shown by the connection to "Speaker information") to selectively extract only the target speaker of interest, effectively filtering out both non-speech noise and other speakers. This visual comparison clearly demonstrates how these technologies address progressively more specific aspects of the speech processing challenge, from broad noise reduction to selective speaker isolation.

2.1.2 Historical Development of Speech Enhancement

SE has evolved through several paradigms, from traditional signal processing approaches to modern deep learning methods. This historical progression provides crucial context for understanding TSE, as TSE model architectures are fundamentally built upon frameworks established in SE while incorporating speaker-specific conditioning mechanisms. TSE systems typically adapt mapping-based methods, masking-based methods, or waveform-based methods from SE by integrating speaker identity information to focus on a target speaker's voice. To establish this foundational connection between SE and TSE, we first provide an overview of the historical evolution of speech enhancement before examining how these approaches have been extended to address the specific challenges of target speaker extraction.

The evolution of speech enhancement spans several decades, progressing from statistical signal processing to modern deep learning approaches. Initial approaches relied on spectral subtraction [2] and Wiener filtering [3], using statistical models to estimate and remove noise components. More sophisticated statistical models, including Gaussian mixture models (GMMs) [4] and hidden Markov models (HMMs), [5],

captured speech and noise characteristics better. Meanwhile, deep learning techniques, particularly simple feed-forward neural networks, began to supplement statistical models, learning mappings [6] or masking [7, 8] between noisy and clean speech. The advent of such techniques led to significant advancements, with architectures such as convolutional neural networks (CNNs) [9, 10] and recurrent neural networks (RNNs) [11, 12] being employed to effectively denoise speech signals. Additionally, generative adversarial networks (GANs) [13, 14, 15] and transformers [16, 17] have advanced the development of speech enhancement. More recently, diffusion-based probabilistic models have been applied to speech enhancement tasks, offering new perspectives and frameworks [18].

2.1.3 Traditional Speech Enhancement Methods

Signal Processing-based Methods

Conventional speech enhancement based on signal processing rely primarily on statistical approaches to estimate clean speech from noisy observations.

Spectral Subtraction

Spectral subtraction [2] represents one of the earliest conventional approaches to speech enhancement, operating on the hypothesis that noise can be removed by subtracting its estimated spectral magnitude from the noisy speech:

$$|\hat{S}(t,f)|^p = |Y(t,f)|^p - |\hat{U}(t,f)|^p \tag{2.4}$$

where p is an empirically chosen exponential value, $|\hat{S}(t,f)|$ is the estimated clean speech magnitude at time frame t and frequency f, $|Y(t,f)|$ is the noisy speech magnitude, and $|\hat{U}(t,f)|$ is the estimated noise magnitude. While conceptually straightforward, spectral subtraction often introduces artifacts such as musical noise, particularly in low SNR conditions [19].

Wiener Filtering

In order to tackle the above problem, Wiener filtering [3] adopts a more statistically optimal approach, estimating the clean speech by minimizing the mean-square error between the estimated and true speech signals:

$$\hat{S}(t,f) = \frac{\xi_{t,f}}{\xi_{t,f}+1} Y(t,f) \tag{2.5}$$

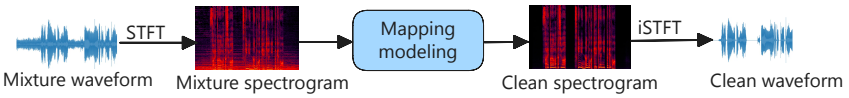
where $\xi_{t,f}$ denotes the priori SNR at time frame t and frequency f. This parameter represents the ratio of clean speech power spectrum to noise power spectrum, and is typically estimated using techniques such as the decision-directed approach [20]:

$$\hat{\xi}_{t,f} = \alpha \frac{|\hat{S}(t-1,f)|^2}{\hat{\sigma}_n^2(t-1,f)} + (1-\alpha)\max[\gamma_{t,f}-1, 0] \tag{2.6}$$

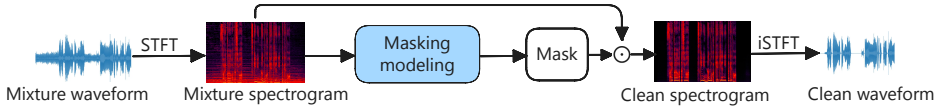
where α is a smoothing factor , $\hat{\sigma}_n^2$ is the estimated noise variance, and $\gamma_{t,f}$ is the posteriori SNR defined as $|Y(t,f)|^2/\hat{\sigma}_n^2(t,f)$. The accuracy of $\xi_{t,f}$ estimation directly impacts the quality of enhanced speech, with more precise estimations leading to better noise reduction while preserving speech components. Wiener filtering generally achieves superior performance to spectral subtraction but remains vulnerable under non-stationary noise conditions.

2.1.4 Neural Network-based Methods

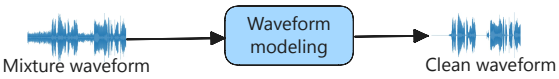
Deep neural networks have substantially advanced speech enhancement and fundamentally changed the solution paradigm. Compared to the conventional models mentioned above, it has a greater capacity to learn optimal processing paradigms from large amounts of data. We review the following three neural network-based TSE network to provide a comprehensive understanding of the current technical landscape and establish a foundation for our proposed approach. An overview of them is shown in Figure 2.2.



(a) Mapping modeling.



(b) Masking modeling.



(c) Waveform modeling.

Figure 2.2: Difference between mapping modeling, masking modeling, and waveform modeling.

Mapping-based Methods

Mapping-based approaches directly estimate the clean speech spectrogram from the noisy input without the explicit use of masks:

$$\hat{S}(t,f) = f_\theta(Y(t,f)) \tag{2.7}$$

where f_θ represents a neural network with parameters θ. Previous work by [6] demonstrated the effectiveness of mapping-based methods using deep neural networks to predict clean log-power spectra from noisy inputs.

Masking-based Methods

Masking-based approaches estimate a time-frequency mask that, when applied to the noisy spectrogram, approximates the clean speech spectrogram:

$$\hat{S}(t,f) = M(t,f) \cdot Y(t,f) \tag{2.8}$$

Various mask formulations have been proposed, including the ideal binary mask (IBM) [21], ideal ratio mask (IRM) [22], and complex ratio mask (CRM) [7]. Masks are typically estimated using neural architectures such as convolutional networks [23],

recurrent networks [24], or their combinations [25].

Waveform-based Methods

Recent advances have enabled direct modeling of waveform samples without relying on time-frequency representations:

$$\hat{s}(n) = f_\theta(y(n)) \tag{2.9}$$

where n is the waveform sample index, and $\hat{s}(n)$ and $y(n)$ represent the estimated clean and noisy speech waveforms, respectively. Representative works at this direction, such as WaveNet-based architectures [26] and Time-domain audio separation network (TasNet) [9], have demonstrated promising performance by avoiding phase-related issues inherent in spectrogram-based approaches.

2.1.5 Limitations of Conventional Speech Enhancement and Speech Separation

Conventional speech separation techniques have demonstrated remarkable capabilities in isolating individual speakers from complex acoustic mixtures. State-of-the-art frameworks including permutation invariant training (PIT) [27], deep clustering [28], and Conv-TasNet [9] have progressively advanced the field by leveraging sophisticated deep learning paradigms. Despite these advances, speech separation frameworks encounter fundamental limitations that constrain their practical deployment. The inherent output permutation ambiguity necessitates additional speaker identification mechanisms [29], creating a two-stage process that increases system complexity. Moreover, the architectural constraint of predefined output channels limits adaptability to dynamic environments with variable speaker counts [30]. The computational burden of processing all speakers simultaneously becomes particularly problematic in resource-constrained applications [31]. Performance degradation with increasing speaker density further compounds these limitations [32], particularly in real-world environments with complex acoustic scenarios.

2.2 Target Speech Extraction

2.2.1 TSE Framework

TSE represents a paradigm shift that directly addresses the limitations of conventional separation approaches. Compared to the early-mentioned problems, rather than attempting to isolate all constituent speakers, TSE reformulates the problem by focusing exclusively on extracting a specified target speaker from the acoustic mixture. This targeted approach inherently resolves the permutation ambiguity by producing a single output corresponding unambiguously to the target speaker. TSE frameworks demonstrate inherent adaptability to scenarios with varying numbers of interference speakers without requiring architectural reconfiguration. The computational efficiency gained through focused processing aligns well with applications where only a specific speaker is of interest, such as personalized hearing assistive devices, selective meeting transcription systems, and voice-controlled interfaces in multi-speaker environments.

2.2.2 Traditional TSE methods

The development of TSE methods can be traced back to the classical "cocktail-party" problem [33], which seeks to emulate the human auditory system's remarkable capability to selectively attend to specific speakers within complex acoustic environments [34]. Early approaches relied predominantly on spatial filtering techniques implemented through fixed beamformers to enhance signals arriving from predetermined directions [35, 36]. These methods, while effective in controlled environments, necessitated multi-microphone arrays and prior knowledge of speaker positioning.

The subsequent emergence of blind source separation (BSS) approaches, particularly those based on Independent component analysis (ICA), marked a significant advancement in the field [37]. These statistical techniques exploited the assumed independence between sources but encountered persistent challenges with permutation ambiguity across frequency bands. This limitation prompted the development of independent vector analysis [38, 39], which addressed the frequency permutation problem through joint modeling of frequency bands [37], establishing methodological foundations that would later influence TSE approaches.

The development of TSE has evolved through independent vector extraction

frameworks that incorporate auxiliary information, such as voice activity patterns obtained from visual cues [40] or speaker characteristics extracted from enrollment audio [41]. Despite their technical advances, these approaches remained constrained by their reliance on multi-channel inputs. A significant transition occurred with the introduction of single-channel approaches based on factorial hidden Markov models [42, 43] and non-negative matrix factorization [44, 45]. These techniques leveraged spectral models learned from clean speech data but were limited in their computational efficiency and generalization capability to unseen speakers.

2.2.3 Early Neural Network Methods for TSE

The integration of DNNs into TSE frameworks marked a transformative milestone in methodological advancement [46]. Building upon the success of neural approaches in general speech separation tasks, researchers adapted and enhanced these architectures specifically for targeted extraction scenarios. The pioneering SpeakerBeam framework [47] introduced a neural architecture that directly incorporated speaker embeddings derived from enrollment audio into the extraction process. This innovation enabled speaker-independent extraction scenarios, where models could effectively generalize to previously unseen speakers, dramatically enhancing system flexibility and applicability. Parallel developments in visual-cue-based neural methods expanded the modality space for TSE systems. Approaches such as "Looking-to-Listen" [48] exploited visual information, particularly lip movements captured through video, to guide the extraction process. These visually-informed systems demonstrated particular robustness in challenging scenarios involving speakers with similar vocal characteristics or adverse acoustic conditions where purely audio-based systems struggled [49]. Recent advancements [50] have further explored neural spatial filtering approaches that leverage multi-channel microphone arrays without encountering the permutation ambiguity issues characteristic of traditional ICA-based methods. These neural spatial filters enable precise speaker targeting based on direction, effectively combining the advantages of traditional beamforming with the flexibility and learning capacity of neural networks [51].

Current state-of-the-art in TSE research is represented by multimodal approaches that integrate complementary information from audio, visual, and spatial domains [52].

These integrated systems, exemplified by multimodal extensions of the SpeakerBeam [53], leverage the distinctive strengths of each modality, resulting in enhanced robustness and extraction accuracy across diverse and challenging real-world acoustic environments [54]. The evolution from conventional signal processing methods to sophisticated neural approaches reflects the field's continuous pursuit of improved performance, expanded flexibility, and enhanced robustness in TSE frameworks [29]. By directly addressing the fundamental limitations of conventional speech separation techniques, TSE offers a more targeted and efficient solution to applications requiring selective attention to specific speakers in complex acoustic environments [55].

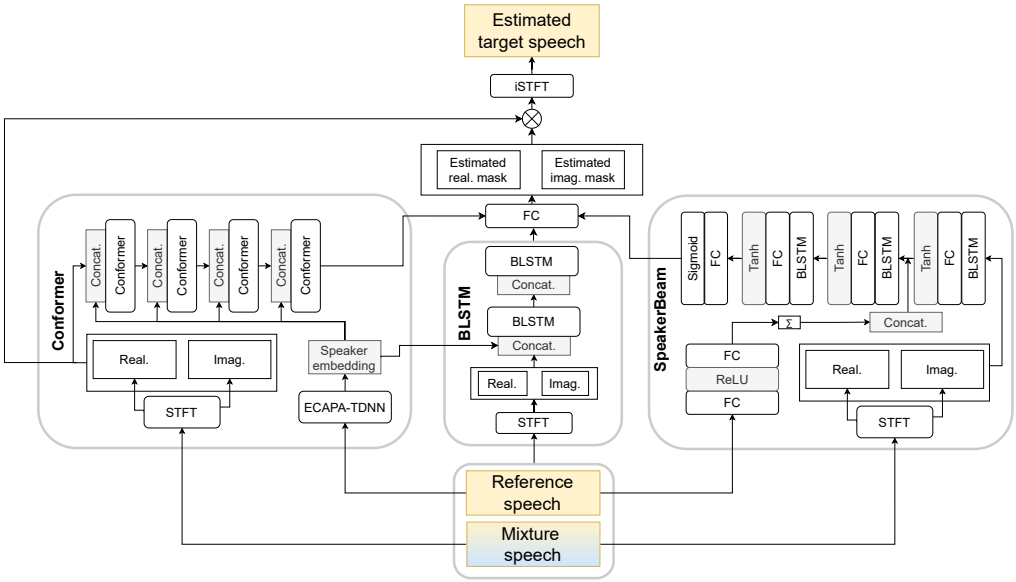


Figure 2.3: An unified illustration of the details of the TSE networks applied for this thesis: Conformer, BLSTM, and SpeakerBeam. Reference speech is processed through ECAPA-TDNN to extract speaker embeddings, which condition both Conformer and BLSTM models. The SpeakerBeam module processes raw waveforms directly. The system generates real and imaginary masks that are applied to the mixture representation before inverse transformation to produce the estimated target speech, which is optimized against ground truth targets during training.

2.3 Neural Network for Target Speaker Extraction

The major DNN-based TSE architectures include those shown in Figure 2.3. Fundamentally requiring two streams of information: the mixture signal containing multiple speakers, and speaker characteristics derived from reference utterances, those TSE systems can be broadly categorized based on their core components and information fusion strategies.

SpeakerBeam [47] implements target speaker extraction using a two-stage approach. In the first stage, a speaker adaptation layer processes the reference speech to extract speaker characteristics and computes a speaker-dependent bias term that modulates the gate activations in an Long Short-Term Memory (LSTM) based extraction network; in the second stage, the extraction network uses these modulated activations to filter the mixture and isolate the target speaker's voice. The network architecture comprises three Bidirectional LSTM (BLSTM) layers, processing concatenated real and imaginary spectra components.

VoiceFilter [56] adopts a CNN-LSTM architecture with explicit speaker embedding conditioning, wherein a separately trained d-vector system extracts fixed-dimensional embeddings from the reference speech. These embeddings are concatenated with intermediate features extracted by convolutional layers and passed to LSTM layers that model temporal dependencies, so that a magnitude mask is ultimately produced and applied to the mixture spectrogram, with phase reconstruction performed using the noisy phase. The architecture contains dilated convolutional layers followed by a bidirectional LSTM and fully connected layers for mask estimation.

Conformer-TSE [1] leverages multi-head self-attention layers to capture long-range dependencies, convolutional modules to extract local patterns and provide relative positional cues, and feed-forward networks to transform features at each position independently; such models typically estimate complex ratio masks for both the real and imaginary components, thereby enhancing extraction quality. The Conformer architecture incorporates four stacked blocks, each containing feed-forward layers, multi-head attention, and convolutional modules with residual connections, effectively combining local and global feature modeling.

Band-Split RNN for TSE (BSRNN-TSE, [57]) is a neural network architecture that serves as the backbone for speech extraction TSE systems. As implemented in

the paper, BSRNN-TSE divides the frequency spectrum into multiple subbands for more effective processing. The frequency band below 1.5 kHz is split using a 100 Hz bandwidth, the band between 1.5 kHz and 3.5 kHz uses a 200 Hz bandwidth, the band between 3.5 kHz and 6 kHz employs a 500 Hz bandwidth, and the frequency band between 6 kHz and 8 kHz is treated as a single subband. This frequency division results in 32 total subbands. The speech extractor is constructed with a BLSTM. This band-splitting approach allows the model to process different frequency regions independently, which is particularly beneficial for separating speakers whose vocal characteristics may be more prominent in specific frequency ranges.

DualStream Contextual Fusion Network (DCF-Net) [58] presents a state-of-the-art approach. Operating in the time-frequency domain, DCF-Net introduces a DualStream Fusion Block that captures contextual information and models the interactions between the contextualized enrollment and mixture representations across both spatial and channel dimensions, fusing rich and consistent representations from two parallel streams to effectively guide the extraction network.

Each of these architectures employs distinct strategies, whether through adaptive biasing, multi-modal conditioning, or advanced fusion of contextual cues, to push the frontier of target speaker extraction in increasingly challenging acoustic environments.

All the TSE systems discussed above rely on speaker embeddings extracted from reference speech to guide the TSE process. The effectiveness of these systems heavily depends on the quality of speaker representations. Therefore, we now describe the speaker encoder used throughout our experiments.

2.3.1 Speaker Encoder Based on ECAPA-TDNN

As intuitively illustrated in Figure 2.1, the speaker encoder is widely recognized as an essential component in modern TSE systems, with DNNs being extensively adopted for recent speaker encoder implementations due to their rapid development and effectiveness. For this thesis, we adopt Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) [59] as the speaker encoder for extracting robust speaker embeddings from enrollment utterances. These embeddings guide the TSE network to accurately isolate the target speaker's voice from mixed speech.

ECAPA-TDNN enhances the standard TDNN architecture by introducing channel attention mechanisms, residual connections, and dense connectivity. Channel attention adaptively emphasizes frequency bands most relevant to speaker characteristics, enhancing embedding discriminability. Residual and dense connections facilitate stable gradient flow and efficient feature propagation, leading to improved speaker representation learning.

Integrating ECAPA-TDNN ensures the TSE system generates discriminative representations, which is beneficial especially under challenging acoustic conditions.

2.4 Evaluation Metrics

2.4.1 Common Evaluation Metrics

Signal-to-Noise Ratio (SNR)

Measures the ratio between the power of the clean signal and the power of the noise:

$$\text{SNR} = 10 \log_{10} \frac{\sum_{n=1}^{N} |s_0(n)|^2}{\sum_{n=1}^{N} |\hat{s}_0(n) - s_0(n)|^2} \tag{2.10}$$

where n is the sample index ranging from 1 to N, and N is the total number of samples.

Signal-to-Distortion Ratio (SDR)

SDR is a fundamental objective metric for evaluating speech separation systems. Let $s_0(n)$ denote the clean target speech signal at sample n, $s_i(n)$ the interference speech signal from other speakers, $u(n)$ the environmental noise signal, and $y(n) = s_0(n) + s_i(n) + u(n)$ the observed mixture signal. The estimated target signal produced by a separation algorithm is denoted as $\hat{s}_0(n)$.

To evaluate the quality of $\hat{s}_0$, it is decomposed into four orthogonal components:

$$\hat{s}_0(n) = s_{\text{proj}}(n) + e_{\text{interf}}(n) + e_{\text{noise}}(n) + e_{\text{artif}}(n) \tag{2.11}$$

Here, s_{proj} is the orthogonal projection of $\hat{s}_0$ onto s_0, defined as:

$$s_{\text{proj}}(n) = \frac{\sum_{m=1}^{N} \hat{s}_0(m) s_0(m)}{\sum_{m=1}^{N} |s_0(m)|^2} s_0(n) \tag{2.12}$$

If the interference signals s_i are available, the interference component is defined as:

$$e_{\text{interf}}(n) = \sum_{i \neq 0} \frac{\sum_{m=1}^{N} \hat{s}_0(m) s_i(m)}{\sum_{m=1}^{N} |s_i(m)|^2} s_i(n) \tag{2.13}$$

Similarly, if the environmental noise u is known, the noise component is computed as:

$$e_{\text{noise}}(n) = \frac{\sum_{m=1}^{N} \hat{s}_0(m) u(m)}{\sum_{m=1}^{N} |u(m)|^2} u(n) \tag{2.14}$$

The artifact component, representing the part of $\hat{s}_0$ that is not explained by the target, interference, or noise components, is given by:

$$e_{\text{artif}}(n) = \hat{s}_0(n) - s_{\text{proj}}(n) - e_{\text{interf}}(n) - e_{\text{noise}}(n) \tag{2.15}$$

Based on this decomposition, SDR is defined as:

$$\text{SDR} = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} |s_{\text{target}}(n)|^2}{\sum_{n=1}^{N} |e_{\text{interf}}(n) + e_{\text{noise}}(n) + e_{\text{artif}}(n)|^2} \right) \tag{2.16}$$

Scale-Invariant Signal-to-Noise Ratio (SI-SNR)

A scale-invariant version that accounts for overall level differences:

$$\text{SI-SNR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \tag{2.17}$$

where $s_{\text{target}} = \frac{\langle \hat{s}, s_0 \rangle}{\|s_0\|^2} s_0$ and $e_{\text{noise}} = \hat{s} - s_{\text{target}}$. The main difference between SNR and SI-SNR is how they handle signal scaling. SNR directly compares the original signal to the estimation error without considering amplitude variations. In contrast, SI-SNR first projects the estimated signal onto the target signal, creating a scale-invariant target component, before measuring the ratio between this aligned target and remaining

noise. This makes SI-SNR more suitable for evaluating speech front-end systems, as it focuses on signal distortion rather than amplitude differences.

Perceptual Evaluation of Speech Quality (PESQ) PESQ is an ITU-T standard (P.862) for objective assessment of speech quality by comparing a reference signal with a degraded signal. PESQ scores range from -0.5 to 4.5, with higher values indicating better perceived quality. This metric correlates highly with subjective listening tests [60].

Short-Time Objective Intelligibility (STOI) STOI measures speech intelligibility by computing the correlation between the temporal envelopes of clean and processed speech signals over short-time segments. STOI scores range from 0 to 1, with higher values indicating better intelligibility. It is particularly robust for time-frequency weighted noisy speech [61].

Deep Noise Suppression Mean Opinion Score (DNSMOS) DNSMOS is a neural network-based predictor trained to estimate human mean opinion scores (MOS) for enhanced speech. DNSMOS is graded on a scale of 1 to 5, with higher scores indicating better perceived quality. It enables real-time, non-intrusive evaluation of speech quality without the need for a reference signal [62].

2.5 Standard Training Approaches

Conventional TSE training typically employs random sampling from the entire dataset, with loss functions based on signal reconstruction quality. This approach randomly selects samples from the training dataset during each iteration, constructing mixture signals containing target speakers, interference speakers, and background noise. For each sample, the model attempts to extract the target speaker's speech from the mixture signal and optimizes model parameters by minimizing reconstruction loss. This random sampling strategy assumes all training samples have equal learning value, disregarding the potentially varying importance of different samples in the learning process.

Algorithm 2.1 Standard TSE Training with Random Sampling

Input: Dataset $\mathcal{D}$ containing target speakers $\mathcal{S}_0$, interference speakers $\mathcal{S}_i$, and background noise $\mathcal{U}$
Input: Batch size B, training iterations T, model parameters θ
Input: Speaker encoder $E(\cdot)$, TSE model $\text{TSE}(\cdot, \cdot; \theta)$, loss function $\mathcal{L}(\cdot, \cdot)$
for $t = 1$ to T **do**
  $\mathcal{B} \leftarrow \emptyset$ ▷ Initialize empty batch
  for $b = 1$ to B **do**
    Sample target speaker $s_0 \sim \mathcal{S}_0$
    Sample utterance $s_0(n) \sim \mathcal{D}(s_0)$
    Sample enrollment utterance $r(n) \sim \mathcal{D}(s_0)$ where $r(n) \neq s_0(n)$
    Sample interference speaker $s_i \sim \mathcal{S}_i$
    Sample interference utterance $s_i(n) \sim \mathcal{D}(s_i)$
    Sample background noise $u(n) \sim \mathcal{U}$
    Create mixture $y(n) = s_0(n) + s_i(n) + u(n)$
    Extract speaker embedding $v_0 = E(r(n))$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{(y(n), v_0, s_0(n))\}$
  end for
  Compute predictions $\hat{s}_{0,i}(n) = \text{TSE}(y_i(n), v_{0,i}; \theta)$ for all $(y_i(n), v_{0,i}, s_{0,i}(n)) \in \mathcal{B}$
  Compute loss $\mathcal{L}_{\text{batch}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}(\hat{s}_{0,i}(n), s_{0,i}(n))$
  Update parameters θ using gradient of $\mathcal{L}_{\text{batch}}$
end for

As shown in Algorithm 2.1, each training iteration randomly samples speech from target and interference speakers, along with background noise. For each target speaker, two distinct speech samples are drawn: one for creating the mixture signal (u_{target}) and another for generating the speaker embedding vector (e_{target}). This method ensures that the model learns to extract target speakers based on their embedding vectors rather than merely memorizing speech characteristics from the training data. The model receives the mixture signal and target speaker embedding as inputs and produces the estimated target speech as output.

This random sampling approach treats all training examples uniformly, regardless of their inherent difficulty. The model parameters θ are updated using gradient descent to minimize the reconstruction loss between the estimated and ground truth target speech. Common loss functions include mean squared error (MSE), SNR defined by Eq. (2.10) , and SI-SDR defined by Eq. (2.17), also known as SI-SNR [63].

During training, the SI-SNR loss function receives particular attention as it is

invariant to signal amplitude scaling, enabling better evaluation of speech extraction quality. This loss function first calculates the projection of the estimated signal $\hat{s_0}$ onto the direction of the true signal s_0 as s_{target}, then computes the error e between the estimated signal and this projection. The final SI-SNR loss is the logarithm of the ratio between the projected signal energy and the error signal energy, with higher values indicating better extraction quality.

While this approach has proven effective for TSE models, it disregards the inherent complexity gradient in the extraction task. By treating simple and challenging examples equally during training, the model may struggle to learn effective strategies for more difficult scenarios, such as those involving similar-sounding target and interference speakers or low SNR conditions. In practice, TSE task difficulty is influenced by multiple factors, including the voice similarity between target and interference speakers, the SNR of the mixture, target speaker speech characteristics (such as volume, speech rate, and pitch), and the type and intensity of background noise. In random sampling training strategies, models tend to prioritize learning capabilities for handling simple samples while performing poorly on more challenging scenarios, such as extracting speech from similar-sounding speakers or under extremely low SNR conditions. These limitations motivate the exploration of more sophisticated training strategies, such as CL, which we examine in subsequent chapters.

This standard training approach treats all training examples uniformly, regardless of their inherent difficulty. The model parameters θ are updated using gradient descent to minimize the reconstruction loss between the estimated and ground truth target speech, typically using SNR defined by Eq. (2.10) or SI-SDR defined by Eq. (2.17) [63].

2.6 Data Augmentation Techniques

2.6.1 Source-preserving Augmentation Approaches

Data augmentation for TSE can be systematically categorized into source-preserving and non-source preserving approaches. Source-preserving augmentations modify only the input mixture while maintaining the integrity of the original target signal, effectively preserving extraction targets while introducing controlled variability. These include Gaussian noise addition with optimal amplitude parameters [64, 65],

gain adjustment within moderate ranges to simulate microphone proximity variations [66], time masking spanning a fraction the signal length to simulate temporal interruptions [66, 67],frequency masking across a percentage of bandwidth to replicate channel effects [66], short noise bursts to simulate environmental events [68], and both time stretching and pitch shifting to account for speaker variability [69, 65]. These transformation techniques collectively enhance model robustness against the diverse acoustic conditions encountered in real-world deployment scenarios.

2.6.2 Non-source Preserving Approaches and Performance Analysis

Non-source preserving augmentations represent a more advanced paradigm that strategically modifies both mixture and target signals, fundamentally altering the extraction task parameters. Dynamic mixing generates entirely new mixtures during training by randomly selecting and combining speakers from the corpus [31, 70], while Mixup [71, 72] performs convex combinations of different mixtures weighted by parameters sampled from Beta distributions. By adapting computer vision techniques to the audio domain, CutMix [73] replaces temporal segments between mixtures to maintain perceptual coherence while expanding training diversity. Comprehensive evaluations across LibriMix [74], TIMIT [75], and VCTK [76] demonstrate that the strategic integration of CutMix [73], Data-only Mixup [77], and Dynamic mixing [78] yields exceptional generalization capabilities. This performance advantage persists across architectural variations, with both Conv-TasNet [9] and dual-path recurrent neural network (DPRNN) [12] based extraction models showing consistent benefits. The efficacy of these techniques can be attributed to complementary mechanisms: adversarial learning effects from controlled distortions [79], prevention of dataset-specific memorization [80], and expansion of the acoustic context space [81]. Particularly noteworthy is the retained performance advantage even when training with substantially reduced datasets (25% of the original data), highlighting the resource efficiency of these augmentation strategies in computationally constrained environments [64, 82].

2.7 Datasets and Benchmark

In this section, we provide detailed descriptions of several datasets and the benchmark commonly used in speech enhancement, speech separation, and target speaker extraction tasks.

WSJ0 The Wall Street Journal (WSJ0) corpus [83] is a widely used speech dataset originally designed for large-vocabulary continuous speech recognition (LVCSR) [83]. It consists of high-quality read speech recordings collected from various speakers reading financial news articles. Due to its clean and well-structured nature, WSJ0 has been extensively used in speech processing research, including speech recognition, enhancement, and separation.

WSJ0-2mix is a dataset derived from the WSJ0 corpus, specifically created for speech separation tasks [28]. It contains artificially mixed two-speaker utterances with different signal-to-noise ratio (SNR), typically ranging from 0 to 5 dB. The dataset is structured with a training set comprising approximately 20,000 mixtures corresponding to around 30 hours of speech involving 101 speakers. The validation set contains about 5,000 mixtures (roughly 8 hours) using the same speaker pool as training, while the test set includes 3,000 mixtures (around 5 hours) created from 18 speakers not present in the training set. This careful partitioning ensures robust evaluation of model generalization to unseen speakers.

LibriSpeech Corpus The LibriVox project is a volunteer-driven initiative that provides free audiobooks of public domain literary works [84]. Established in 2005, it has since accumulated a vast collection of spoken recordings across multiple languages, contributed by volunteers worldwide. These audiobooks vary in speaker identity, gender, and reading styles, making them a valuable resource for speech-related research, including ASR, TTS, speaker identification, and speech separation. Based on LibriVox, the LibriSpeech corpus was created as a large-scale, high-quality speech dataset for ASR [85]. It contains approximately 1,000 hours of 16kHz English speech read aloud, carefully curated to ensure clean recordings with precise transcriptions. Due to its extensive speaker diversity and well-structured nature, LibriSpeech has become a standard benchmark in the speech processing community.

LibriTTS [86] is a significant extension of the LibriSpeech corpus, specifically designed to address the limitations of previous datasets for TTS research. The corpus comprises approximately 585 hours of speech data at a sampling rate of 24kHz, derived from audiobook recordings in the LibriVox project. Unlike LibriSpeech, which primarily focuses on ASR tasks, LibriTTS maintains crucial speaker and textual metadata, provides normalized text transcriptions, and features higher audio quality suitable for modern neural TTS systems. The dataset is carefully balanced across genders and includes a diverse range of 2,456 speakers, facilitating robust speaker adaptation and multi-speaker TTS model development. LibriTTS is partitioned into training subsets of varying sizes and acoustic conditions (train-clean-100, train-clean-360, train-other-500), along with development and test sets that maintain consistent characteristics with the training data. The high-quality audio, coupled with precise alignments between text and speech, has established LibriTTS as an essential resource for advancing neural speech synthesis research, particularly for models requiring diverse speaker characteristics and prosodic variations.

Libri2Mix [74] is derived from LibriSpeech and is specifically designed for speech separation tasks. It consists of artificially mixed two-speaker utterances, following a strict minimum duration protocol, where the longer utterance in each pair is truncated to match the shorter one, ensuring a 100% overlap rate. The dataset includes Train-360 with approximately 50,800 mixtures totaling around 212 hours involving 921 speakers, and Train-100 with about 13,900 mixtures covering roughly 58 hours with 251 speakers. Both validation and test sets contain approximately 3,000 mixtures (about 11 hours) with 40 speakers each. Combining all subsets, the training data spans roughly 270 hours with a total of 1,172 speakers. Each mixture is generated by randomly selecting two utterances from different speakers and mixing them at random signal-to-noise ratios (SNRs), typically within a fixed range. The dataset is widely used for speech separation and target speaker extraction (TSE) research.

Libri2Talker [87] is an extension of Libri2Mix that maximizes data efficiency by swapping the roles of the target and interference speakers in each two-speaker mixture. This process effectively doubles the number of training samples, providing additional variability for TSE models. It retains the 100% overlap rate and the minimum duration protocol, ensuring challenging conditions for speaker extraction and verification tasks. The dataset consists of a training set with 127,056 mixtures created from 1,172 speakers

spanning 460 hours, a validation set with 2,344 mixtures using the same speaker pool as the training set, and a test set comprising 6,000 mixtures from Libri2Mix with an additional 2,260 enrollment utterances from the LibriSpeech test-clean set. Libri2Talker is particularly useful for target speaker extraction (TSE) and speaker verification tasks, as it allows models to learn robust extraction strategies under different interference conditions.

Comparison of Libri2Mix, Libri2Talker, and LibriTTS The LibriSpeech derivatives exhibit distinctive characteristics optimized for their intended applications. Libri2Talker substantially expands upon Libri2Mix by implementing a role-swapping method, effectively doubling the available training examples while preserving the acoustic fidelity inherent to the source LibriSpeech corpus. This data augmentation approach facilitates enhanced model learning of speaker-specific features, resulting in improved generalization capabilities across varied acoustic environments. The computational efficiency of Libri2Talker is particularly noteworthy, as it increases diversity without necessitating additional recording infrastructure. In contrast, LibriTTS diverges from these separation-oriented datasets by prioritizing high-fidelity audio components (24kHz sampling rate versus the 16kHz in Libri2Mix/Libri2Talker) and maintaining critical speaker and textual metadata required for advanced speech synthesis tasks. The three datasets represent a progression in the LibriSpeech ecosystem, with Libri2Mix establishing fundamental separation capabilities, Libri2Talker enhancing extraction robustness through innovative data manipulation techniques, and LibriTTS extending the application domain toward high-quality speech synthesis with precise text-speech alignments. This complementary relationship enables researchers to leverage domain-specific optimizations while benefiting from the consistent speaker characteristics present across all three derivatives.

VoxCeleb The VoxCeleb1 [88] dataset is a large-scale audio-visual collection designed for speaker identification and verification tasks. It comprises over 100,000 utterances from 1,251 celebrities, extracted from interview videos uploaded to YouTube. The dataset is gender-balanced, with 55% male speakers, and includes a wide range of ethnicities, accents, professions, and ages. The data was collected using an automated pipeline that involves obtaining videos from YouTube, performing active speaker

Dataset (Partition)	#Speakers	#Utterances	Duration (Hours)
Libri2Mix (Train-360)	921	50,800	212
Libri2Mix (Train-100)	251	13,900	58
Libri2Mix (Validation & Test)	40	3,000 each	11 each
Libri2Talker (Train)	1,172	127,056	460
Libri2Talker (Validation)	1,172	2,344	8
Libri2Talker (Test)	40	6,000	22
LibriTTS (train-clean-100)	247	33,236	54
LibriTTS (train-clean-360)	904	116,500	191
LibriTTS (train-other-500)	1,160	205,044	310
LibriTTS (dev-clean & test-clean)	78	5,736	9
LibriTTS (dev-other & test-other)	72	12,286	21

Table 2.1: Comparison of Libri2Mix, Libri2Talker, and LibriTTS datasets.

verification using a two-stream synchronization CNN, and confirming the identity of the speaker using CNN-based facial recognition [88].

Building upon VoxCeleb1, the VoxCeleb2 [89] dataset expands the speaker pool and the number of utterances. It contains over 1 million utterances from 6,112 celebrities, also extracted from YouTube videos. The development set includes 5,994 speakers with approximately 1,092,009 utterances, averaging about 7.8 seconds per utterance, totaling around 2,370 hours of audio. The evaluation set consists of 118 speakers with 36,237 utterances. Similar to VoxCeleb1, VoxCeleb2 captures real-world conditions with background noise, varied accents, and diverse recording environments [89].

CN-Celeb The CN-Celeb dataset is a large-scale speaker recognition dataset collected in the 'in the wild' manner, focusing on Chinese celebrities. It comprises two subsets:

- **CN-Celeb1[90]:** Contains more than 130,000 utterances from 1,000 Chinese celebrities, covering 11 different genres, including entertainment, interview, singing, game play, movie, vlog, live broadcast, speech, drama, recitation, and advertisement. This diversity introduces complex inter-session variations, both intrinsic (speaking style, physiological status) and extrinsic (recording device, background noise).

- **CN-Celeb2[91]:** Expands the dataset with over 520,000 utterances from 2,000 Chinese celebrities, maintaining the coverage of 11 genres. The data collection process was organized by the Center for Speech and Language Technologies at

Tsinghua University and reflects real-world scenarios with significant variability in recording conditions.

Comparison of VoxCeleb and CN-Celeb Datasets

Dataset	#. Speakers	#. Utterances	Total Duration (Hours)
VoxCeleb1	1,251	153,516	1,211
VoxCeleb2	6,112	1,128,246	2,794
CN-Celeb1	1,000	130,109	274
CN-Celeb2	2,000	529,485	1,090

Table 2.2: Comparison of VoxCeleb and CN-Celeb datasets.

Both VoxCeleb and CN-Celeb provide extensive resources for speaker recognition research. VoxCeleb datasets are primarily in English and capture a wide range of real-world conditions, making them suitable for developing robust speaker recognition systems. CN-Celeb datasets focus on Chinese speakers and cover multiple genres, introducing additional challenges such as genre mismatch and complex inter-session variations, thereby offering valuable resources for advancing speaker recognition technologies in diverse and unconstrained environments.

Deep Noise Suppression Challenge Dataset The Deep Noise Suppression (DNS) Challenge dataset [92] is curated to evaluate and advance speech enhancement systems, focusing on improving perceptual speech quality in noisy environments. The dataset has evolved over successive challenges, each introducing more comprehensive and diverse data. The clean speech component encompasses a vast collection of recordings across multiple languages and speaking styles. For instance, the DNS Challenge 3 provided approximately 760.53 hours of clean speech, consisting of 562.72 hours of read speech, 8.80 hours of singing voice, 3.6 hours of emotion-rich expressive speech, and 185.41 hours of Mandarin speech. This multi-domain approach ensures that enhancement models can generalize across diverse speaking conditions and linguistic characteristics. Complementing the speech corpus, the challenge incorporates a diverse set of noise clips totaling approximately 181 hours, methodically sourced from various environments including AudioSet [93], Freesound [94], and DEMAND [95]. This careful curation ensures robustness against different acoustic interference

types commonly encountered in real-world scenarios. To simulate realistic spatial acoustics, the dataset includes over 118,000 room impulse responses (RIRs), comprising both real-world measurements and synthetically generated profiles. These RIRs are accompanied by critical acoustic parameters such as Reverberation Time (T60) and Clarity (C50), enabling precise evaluation of enhancement algorithms under controlled acoustic conditions. The comprehensive nature of the DNS Challenge dataset, with its systematic organization of clean speech, environmental noise, and acoustic simulations, has established it as a rigorous benchmark for advancing deep learning-based speech enhancement technologies that can operate effectively across diverse acoustic environments and speaking contexts.

2.7.1 State-of-the-art Performance

Recent state-of-the-art results in target speaker extraction demonstrate impressive performance benchmarks across standardized datasets. Systems evaluated on the WSJ0-2mix benchmark have achieved an SI-SDR improvement of approximately 21.7 dB, as evidenced by the DCF-Net architecture [58]. Similarly, on the Libri2Mix dataset, a BSRNN-based approach [57] has demonstrated an SI-SDR improvement of approximately 15.39 dB. These metrics highlight the substantial progress in TSE technology, though it should be noted that performance figures may vary depending on model architectures, training frameworks, and evaluation protocols employed across different research implementations.

2.8 Limitations of Current Approaches

Despite significant advances, current TSE systems face several critical limitations:

2.8.1 Acoustic and Speaker Diversities

Current TSE systems exhibit constrained generalization capabilities. Most systems train on datasets with fewer than 1,200 unique speakers, creating a fundamental mismatch with real-world diversity and contrasting with speaker recognition systems that leverage thousands of speakers for robust performance.

The acoustic environment mismatch compounds these issues, as training on artificially mixed data fails to capture complex natural acoustic characteristics, resulting in performance degradation when encountering real-world reverberation, dynamic background noise, or device-specific signatures.

Performance deteriorates significantly in similar-speaker scenarios where target and interference speakers share acoustic similarities (same gender, age group, or accent), particularly problematic in family settings or homogeneous environments.

2.8.2 Training Strategy Limitations

Conventional random sampling on the training data treats extraction scenarios of varying difficulty equally, failing to distinguish between straightforward cases (different-gender speakers) and challenging scenarios (similar-sounding speakers), potentially undermining the model's capacity to handle difficult extractions critical for real-world performance.

Current approaches typically utilize a fixed set of speakers and acoustic conditions, restricting exposure to the full spectrum of real-world variability. This constrained distribution hampers generalization, particularly for unseen speakers or novel acoustic environments.

The standard practice of randomized batch selection throughout training may impede optimal learning progression, as models simultaneously encounter simple and challenging examples without a structured curriculum, potentially leading to suboptimal convergence.

These limitations necessitate more sophisticated approaches that address the complexity gradient in TSE tasks. CL, which progressively introduces more difficult examples in the training data, offers a promising direction. Additionally, expanding training data diversity through synthetic speaker generation could effectively address the constrained speaker representation in current datasets. These advanced strategies form the core focus of our research.

3

Curriculum Learning for Target Speaker Extraction

3.1 Curriculum Learning

CL [96] is a training method in machine learning that mimics how humans learn by gradually introducing more complex concepts to models. This approach builds on basic teaching ideas, where well-organized learning steps help models gain and keep knowledge more effectively. Unlike conventional training approaches that randomly sample from the entire dataset, effectively exposing the model to the whole dataset of complexity, CL implements a step-by-step progression from simpler to more challenging examples.

CL is supported by cognitive science, where it has been demonstrated that learning step by step, from easy to hard, improves understanding, generalization, and speed of learning in complex tasks [97]. In computational learning theory, this method can be formalized as a sequence of weighted training distributions that gradually transition

from focusing on simpler examples to incorporating increasingly difficult instances, ultimately converging to the target distribution of interest.

From an optimization perspective, CL functions similarly to continuation methods [98, 99], where models initially optimize a smoothed, simplified version of the objective function before tackling progressively more complex formulations. Such progressive approaches helps navigate the complex loss landscapes characteristic of deep learning, potentially avoiding problematic local minima and saddle points that might otherwise trap optimization procedures. [100] provides theoretical evidence that CL can improve convergence speed by guiding the model through steeper areas of the gradient, effectively accelerating the learning process with positive guidance. From a data distribution perspective, CL helps models initially focus on high-confidence, cleaner data regions before gradually incorporating more challenging examples.

The formal introduction of CL as a comprehensive training paradigm occurred with [101], who systematically demonstrated its effectiveness across multiple supervised learning domains, including visual recognition and natural language processing. This seminal work established CL as a principled approach rather than merely an ad-hoc training trick. Subsequently, the methodology has experienced rapid expansion and adaptation across diverse machine learning domains. In computer vision, CL has proven effective for object recognition and image classification tasks [96, 102]. CL strategies have also been applied for machine translation and language modeling [103, 104]. Speech processing systems have benefited from curriculum approaches in automatic speech recognition tasks [105], while reinforcement learning environments have integrated CL for complex control problems [106, 107], demonstrating the broad applicability and effectiveness of this training paradigm across the machine learning landscape.

Empirical evidence across multiple domains suggests that these structured learning progressions can lead to improved convergence rates, enhanced model generalization, and increased robustness to noise and outliers in training data. In practice, CL is typically instantiated by two design choices: a difficulty measure and a training scheduler [96]. The difficulty measure defines a scalar (or ordinal) score $C(d)$ for each training example d, indicating its relative easiness/hardness, while the training scheduler specifies when and how much harder data should be exposed to the model during optimization. This abstraction provides a task-agnostic view of CL: data are first

ranked or grouped according to $C(\cdot)$, and then the scheduler constructs a sequence of training subsets (or sampling distributions) that gradually approaches uniform sampling over the full dataset.

Difficulty measurers can be predefined or automatic. Predefined measures encode prior knowledge about what makes examples hard (e.g., structural complexity, diversity, or noise), whereas automatic measures estimate difficulty from data-driven signals, such as model loss or a learned scoring model. In practice, predefined measures are attractive for their interpretability and low overhead, while automatic measures are often more flexible when difficulty factors are numerous or difficult to quantify explicitly.

Training schedulers are commonly categorized as discrete or continuous. Discrete schedulers update the training subset after a fixed number of epochs (or after convergence on the current subset). A representative example is the baby-step strategy, which partitions the ranked data into K buckets from easy to hard, starts training with the easiest bucket, and then progressively merges harder buckets into the training set. Continuous schedulers, on the other hand, control the curriculum by a pacing (or competence) function $\lambda(t) \in (0, 1]$ that maps epoch t to the proportion of easiest samples available for training, with $\lambda(t)$ being monotone non-decreasing and reaching 1 at the end of curriculum. Both scheduler families share the same goal: to bias early optimization toward easier regions of the data distribution and gradually expand coverage to the full task distribution.

These two components naturally connect CL to TSE. In TSE, each training example is a mixture/reference pair, and its difficulty can be characterized by acoustic and speaker-related factors (e.g., target–interference similarity, SNR/SDR, overlap ratio, number of interferers, or real vs. synthetic interferers). Once a difficulty score (or bins) is defined, a scheduler can either (i) predefine the training order by progressively enlarging the subset from easy to hard, or (ii) adaptively select/weight examples using model feedback (self-paced learning), which aligns with the implementation described later in this chapter.

Algorithm 3.2 shows the widely used baby-step strategy. The core idea is to (i) rank training examples by a chosen difficulty measure $C(\cdot)$, (ii) partition them into K buckets from easy to hard, and (iii) progressively merge harder buckets into the active training set. This generic template is included here as a reference point; later in this

Algorithm 3.2 Baby-step curriculum training scheduler

Require: Training set $\mathcal{D}$, difficulty measure $C(\cdot)$, number of buckets K, update criterion (fixed epochs p or convergence)
Ensure: Trained model parameters θ

```
Initialize model parameters θ
Sort D in ascending order by difficulty score C(d)
Split sorted data into K buckets {D_1, ..., D_K} from easy to hard
D_train ← ∅
for s = 1 to K do
    D_train ← D_train ∪ D_s
    while update criterion is not met on D_train do
        Sample mini-batches from D_train and update θ by gradient descent
    end while
end for
```

chapter we instantiate the same principle for TSE, where the buckets correspond to mixture/reference pairs stratified by speaker/acoustic difficulty factors.

3.1.1 Motivations of CL for TSE

In the context of TSE, which aims to isolate a specific speaker's voice from a mixture of overlapping speakers, CL offers particular advantages. TSE tasks exhibit a natural hierarchy of difficulty that makes them especially well-suited for curriculum-based training. For instance, separating speakers of different genders with distinct vocal characteristics is inherently easier than distinguishing between speakers of the same gender with similar pitch and timbre. Similarly, extracting speech at high SNR presents fewer challenges than processing heavily corrupted mixtures (low SNR). This clear difficulty progression provides an ideal foundation for structured curriculum design. The complexity of speech separation tasks makes them ideal candidates for curriculum-based approaches, where learning from simpler separation scenarios can build robust foundations before tackling more challenging cases.

In TSE, randomly sampling the entire dataset forces models to handle easy and difficult examples simultaneously, leading to unstable training and suboptimal performance. Our proposed CL strategy addresses this by starting with easier examples and gradually introducing more challenging ones. This progressive approach allows

models to first establish strong foundational skills on simpler cases before tackling complex scenarios, such as separating speakers with similar vocal characteristics. The structured learning progression improves convergence stability and enhances the model's ability to handle challenging extraction tasks.

3.2 Error analysis of the neural network-based TSE

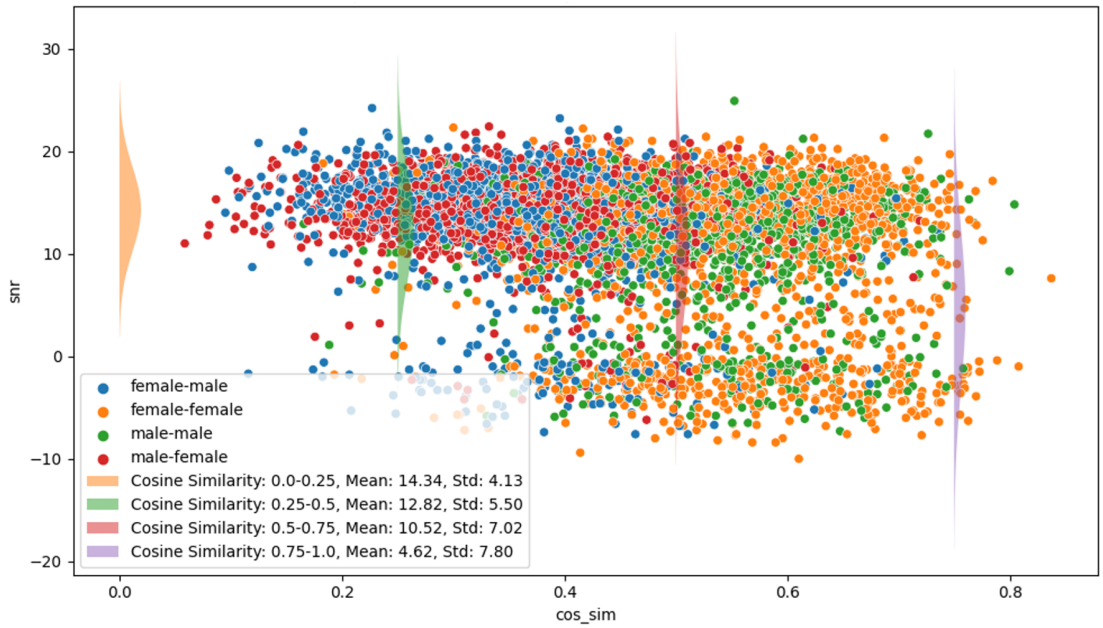


Figure 3.1: Scatter plot of cosine similarity vs SNR with normal distributions. If the TSE system successfully extracts the target speaker, the SNR should be positive.

Before introducing the CL approach for TSE, we first conducted a systematic error analysis of the neural network-based TSE systems introduced in Chapter 2. This analysis serves as a crucial foundation for developing effective difficulty measures and training schedulers. By evaluating model performance on the commonly-acquired TSE data, we mapped the relationship between extraction quality and various speaker characteristics, particularly focusing on the impact of speaker similarity on separation difficulty.

Speaker similarity was measured using cosine similarity between speaker embeddings extracted by the ECAPA-TDNN encoder described in Section 2.3.1, computed

as sim = $\cos(\mathbf{e}_{\text{target}}, \mathbf{e}_{\text{interference}})$ where $\mathbf{e}_{\text{target}}$ and $\mathbf{e}_{\text{interference}}$ represent the normalized embedding vectors for target and interference speakers respectively. Performance statistics including mean and standard deviation of SNR improvement were calculated across different similarity bins to quantify the relationship between speaker similarity and extraction difficulty. Gender-pair configurations were categorized into four types: male-male (MM), female-female (FF), male-female (MF), and female-male (FM) combinations, enabling analysis of how gender similarity affects separation performance across different acoustic characteristics.

For this analysis, we employed the Conformer-based TSE model detailed in Section 2.3 and the Libri2talker dataset introduced in Section 2.7. We used the standard training approach described in Section 2.1 that treats all training examples equally regardless of their inherent complexity.

Figure 3.1 visualizes the relationship between speaker similarity and model performance by plotting SNR improvement against cosine similarity for mixed speech containing target and interference speakers. The horizontal axis represents cosine similarity between speaker embeddings (obtained through our ECAPA-TDNN encoder), with higher values indicating greater acoustic similarity between speakers. The vertical axis shows SNR improvement, calculated as the difference between two key SNR measurements as equation 3.1. This visualization method was developed by first training a standard TSE model using conventional procedures, then evaluating this model on our training dataset to measure the SNR improvement as defined above.

$$\text{iSNR(n)} = 10\log_{10}\left(\frac{\|s_0(n)\|^2}{\|\hat{s}(n) - s_0(n)\|^2}\right) - 10\log_{10}\left(\frac{\|s_0(n)\|^2}{\|y(n) - s_0(n)\|^2}\right) \tag{3.1}$$

This SNR improvement directly measures how much the model improves the signal quality compared to the original mixture. A positive value indicates that the TSE successfully extracted the target speaker. Here, $s_0(n)$ is the target speech, $\hat{s}(n)$ is the model's estimated speech, and $y(n)$ is the input mixture.

Three key observations emerge from this analysis: First, as cosine similarity increases, the mean SNR improvement decreases systematically, as evidenced by the distribution means (14.34 dB for cosine similarity 0.0-0.25, decreasing to 4.62 dB for cosine similarity 0.75-1.0). This quantitative decline confirms that acoustically similar speakers present substantially greater separation challenges for the model.

Second, the standard deviation values increase with similarity (from 4.13 dB to 7.80 dB), indicating greater performance variability for difficult cases. Third, examining the distribution pattern in Figure 3.1 reveals two distinct clusters: one with positive SNRs representing successful extractions, and another near zero SNR where the TSE failed to extract the target speaker. Within the successful extraction cluster, speaker pairs with different genders predominantly appear on the left side (lower similarity region), while same-gender pairs tend to appear on the right side (higher similarity region). Notably, in the failed extraction cluster where SNR is near zero, most pairs share the same gender and exhibit high similarity between target and interference speakers, as highlighted by the color-coded gender pairings (orange for female-female, green for male-male) that cluster predominantly in the high cosine similarity region and frequently display negative SNR improvements.

Notably, the figure not only validates our speaker-dependent difficulty measures but also naturally introduces our mixture-dependent metrics. The regions with negative SNR values (below the horizontal axis) represent particularly challenging cases where the model actually degrades the signal quality rather than improving it. These cases automatically identify hard samples that would benefit from later-stage curriculum inclusion. Conversely, samples with high SNR improvement (upper portions of the graph) represent easier cases suitable for early curriculum stages. This analysis creates a natural bridge between our speaker-dependent and the mixture-dependent factors discussed in the subsequent sections.

3.2.1 Speaker-Dependent Factors

Gender

Pilot experiments showed in the Figure 3.2 revealed that the initial TSE model exhibited higher error rates when target and interfering speakers shared the same gender. Consequently, gender information of both target and interfering speakers can serve as an effective difficulty measure for curriculum design. A potential CL strategy involves initially training with speaker pairs where target and interfering speakers have different genders, then progressively incorporating more challenging same-gender speaker pairs in later training stages. This approach allows the model to first establish robust separation capabilities on easier cross-gender scenarios before tackling the

more complex same-gender interference cases, where acoustic similarities pose greater challenges for TSE.

Speaker Similarity

We compute the cosine similarity between the speaker embeddings of the target and interference speakers (the same embeddings as described in Section 3.3.1)

$$\text{sim} = \cos\Big(E(s_0(n)),\, E(s_i(n))\Big).$$

Lower similarity (i.e., more dissimilar speakers) corresponds to easier samples, while higher similarity indicates a higher difficulty. This continuous measure provides significant advantages over binary distinctions like gender by establishing a precise difficulty level that enables smoother training progression. While same-gender speaker pairs generally exhibit higher similarity scores, this embedding-based approach reveals that certain same-gender combinations may actually be easier to separate than some cross-gender pairs, demonstrating the limitations of simple binary categorization. The speaker embedding approach captures comprehensive acoustic characteristics that directly influence separation complexity, including fundamental frequency patterns, speaking rate variations, accent characteristics, and spectral distribution features. Implementation relies on pre-trained speaker encoders such as ECAPA-TDNN aforementioned, which transform speech segments into fixed-dimensional embedding vectors optimized for speaker discrimination. By quantifying similarity within this specialized embedding space, the approach establishes an accurate difficulty gradient that correlates strongly with actual extraction challenges, enabling more effective curriculum design than crude categorical distinctions.

3.2.2 Mixture-Dependent Factors

SDR

For artificially mixed speech data, SDR (as a difficulty measure, not to be confused with the SDR used as an evaluation metric) is computed as:

$$\text{SDR} = \frac{1}{N}\sum_{n=1}^{N} 10\log_{10}\left(\frac{\|s_0(n)\|^2}{\|s_i(n)\|^2}\right) \tag{3.2}$$

where $s_0(n)$ represents the target speaker's signal and $s_i(n)$ represents the interference signal.

Lower SDR values imply that the target speech is more heavily masked by interference, indicating higher difficulty. This evaluation assumes the use of artificially created training data with access to clean individual signals. We further assume that the artificial mixtures are free of noise and that the TSE system introduces no artifacts, effectively functioning as an ideal TSE. SDR provides a direct measure of how well the target speaker stands out in the mixture, making it particularly valuable for synthetic data where we can control the mixing process to create examples at desired difficulty levels.

SNR

Similarly, we acquire the early-defined SNR as the evaluation metric. Lower SNR values indicate samples that are more challenging for the extraction model. Unlike SDR, which measures intrinsic properties of the mixture, SNR requires a pre-trained model (described in Section 3.2) to produce the estimated target speech $\hat{s}_0$. This makes SNR a model-driven measure that adapts to the specific characteristics of the current extraction approach. The use of SNR as a difficulty measure rests on another strong assumption: that we have already trained a TSE model once, allowing us to use this pre-trained model to directly measure extraction performance on training samples.

This model-dependent approach offers advantages by accounting for the particular strengths and weaknesses of the current model architecture. SNR-based difficulty assessment captures model-specific challenges, such as feature representation limitations and optimization landscape complexities. Samples that yield lower SNR values with

the current model represent gaps in the model's capabilities that need to be addressed through targeted training. This enables a more tailored curriculum that specifically targets the model's current limitations, rather than generic acoustic properties.

3.2.3 Dynamic Difficulty Measure

Self-Paced

In addition to the fixed measures above, we propose a dynamic difficulty measure based on the current model training loss. In the self-paced approach, only samples whose loss exceeds a threshold τ_{SP} are used for back-propagation in each mini-batch. This threshold is gradually decreased so that the model progressively incorporates more difficult samples.

Self-paced learning represents the most adaptive difficulty measure, dynamically adjusting to the model's current capabilities. This approach automatically assesses which examples are appropriate for learning at any given stage, adapts as the model's capabilities evolve during training, and implicitly accounts for all factors contributing to extraction difficulty. The approach is particularly valuable when the factors making extraction difficult are numerous or complex to quantify explicitly. Using the loss as a proxy for difficulty allows self-paced learning to capture nuanced interactions between various acoustic factors that might be missed by more explicit measures.

3.3 Experimental Setup

3.3.1 Model Configurations

In our experiments, we implement and evaluate four distinct TSE architectures described in Section 2.3, namely Conformer, BLSTM, SpeakerBeam, and VoiceFilter. All models were trained to generate complex ratio masks for TSE.

Input Representation

For all models, we used complex-valued spectrograms obtained through short-time Fourier transform (STFT) with a 512-point FFT, 25ms window, and 10ms hop size. The input features consist of concatenated real and imaginary parts of the STFT, resulting

in 512-dimensional representation (256 real + 256 imaginary components, with DC component removed).

Speaker Embedding

We used a 192-dimensional x-vector extracted using ECAPA-TDNN [59] as the speaker representation for all models. This speaker encoder was trained on CN-Celeb [91], which outperformed VoxCeleb-trained embeddings in our preliminary experiments.

3.3.2 Model Configuration

Conformer: The Conformer-based model consists of four Conformer blocks with the following hyperparameters: a feed-forward dimension of 1024 with dropout rate of 0.2, multi-head attention with 4 heads and dropout rate of 0.2, convolutional kernel size of 3, and Swish activation function.

BLSTM: The BLSTM model maintains the same input and output processing as the Conformer but replaces the Conformer blocks with three BLSTM layers with 512 hidden units per direction and a final fully connected layer with 512 output dimensions.

SpeakerBeam: Our SpeakerBeam implementation follows [47] with modifications for complex spectra processing. It consists of three BLSTM layers with 512 hidden units per direction and applies speaker adaptation through concatenation along time dimension.

VoiceFilter: The implemented VoiceFilter model follows [56] with adaptations for complex spectra. It comprises eight convolutional layers with kernel sizes from 1 to 7, dilation factors increasing from 1 to 16, one BLSTM layer with 400 hidden units per direction, and two fully connected layers of sizes 600 and 512.

3.3.3 Curriculum Strategy Implementation

Predefined Approach. For most of our CL strategies, the training scheduler follows a straightforward approach. We divide the training data into subsets based on a predefined difficulty measure and sort these subsets from easy to hard. We begin training the TSE model parameters using only the easiest subset. After a fixed number of training epochs or once the training loss converges, we merge the next subset into

Table 3.1: Dataset and its partitions for this study. No overlap exists between the test partition and the other two partitions at either the utterance or speaker level.

Dataset	Partition	#utterances	#speakers
Libri2talker	train	127,056	1,172
	dev	2,344	1,172
	test	6,000	40

the training portion. The model parameters are further updated using this expanded training set. This process repeats until all subsets are incorporated and the model parameters are well-trained.

Self-Paced Approach. The self-paced strategy employs a different training scheduler. We form training mini-batches using standard random sampling. For each sample within a mini-batch, we compute the objective value using the current model. The curriculum is implemented by adjusting the threshold that determines which samples contribute to the objective function before gradient calculation. For the SNR-based objective function, we gradually decrease this threshold over time. This ensures easy samples dominate gradient computation initially, with difficult samples incorporated later in training. This approach helps manage challenging samples that would destabilize early learning, ultimately improving training stability and the network's generalization ability.

3.4 Experiments and Results

3.4.1 Dataset

We evaluate our CL strategies using the Libri2talker dataset [87], a two-talker mixture dataset derived from LibriSpeech [85]. Table 3.1 lists the dataset partitions, ensuring no overlap between the test set and the training/validation sets at either the utterance or speaker level.

3.4.2 Experimental Conditions

All audio is resampled to 16 kHz. Reference speeches are padded or segmented to 15 seconds, while mixture signals are clipped to 6 seconds. Speaker embeddings are

Table 3.2: iSDR(dB) results on the test set using the CL-based TSE. “Used data" shows the percentage of the training data utilized in the 1st phase. The 1st phase results were obtained from systems trained using only “easy" samples. The 2nd phase results were from systems additionally trained on all samples based on the model weights from the 1st phase.

	Optimal τ	**Used data**	**1st phase**	**2nd phase**
random	-	100%	12.50	12.57
τ_{gender}	Different	50.0%	7.64	11.40
τ_{spk}	0.6	83.8%	13.03	**13.44**
τ_{SDR}	1	44.3%	8.75	13.40
τ_{SNR}	10	82.4%	12.79	12.99

extracted using a pre-trained ECAPA-TDNN [59] trained on the CN-Celeb dataset []. The STFT is computed with a 32-ms window, 8-ms hop, and 512-point FFT.

Training is performed using the Adam optimizer [108] with an initial learning rate of 1×10^{-3}. The learning rate increases gradually over the first 5000 batches, then decays with an inverse square-root schedule to a minimum of 1×10^{-5}. Each mini-batch contains 48 pairs of target and interference speakers.

For evaluation, improvement is measured using the improvement in SDR (iSDR), defined as the relative increase in SDR compared to the mixture. All CL strategies (except self-paced) follow a two-stage training protocol: Stage 1 trains for 100 epochs using easy samples selected based on the difficulty measures, and Stage 2 continues training for an additional 5 epochs using the **full training set**.

The specific data selection strategies are as follows:

Random: Standard training approach where samples are randomly selected from the entire dataset without any difficulty-based filtering. This serves as our baseline and represents the conventional TSE training procedure.

Gender: In stage 1, only samples with different genders between the target and interference speakers are selected.

Speaker Similarity: In stage 1, samples with cosine similarity below a threshold τ_{spk} are selected. We evaluate $\tau_{spk} \in \{0.5, 0.6, 0.7\}$.

SDR: In stage 1, samples are selected based on an SDR threshold τ_{SDR}, varied from -5 dB to 5 dB in 1 dB increments.

SNR: In stage 1, samples are selected based on an SNR threshold τ_{SNR}, with evaluation thresholds of 0 dB, 5 dB, and 10 dB.

Self-paced: The self-paced method divides training into 5 stages. Initially, the model is trained for 1 epoch using the full dataset; then, three stages are applied with configurations (start epoch, end epoch, τ_{SP}) of (1, 30, 10 dB), (31, 60, 5 dB), and (61, 80, 0 dB); finally, the model is trained for an additional 20 epochs on the full dataset.

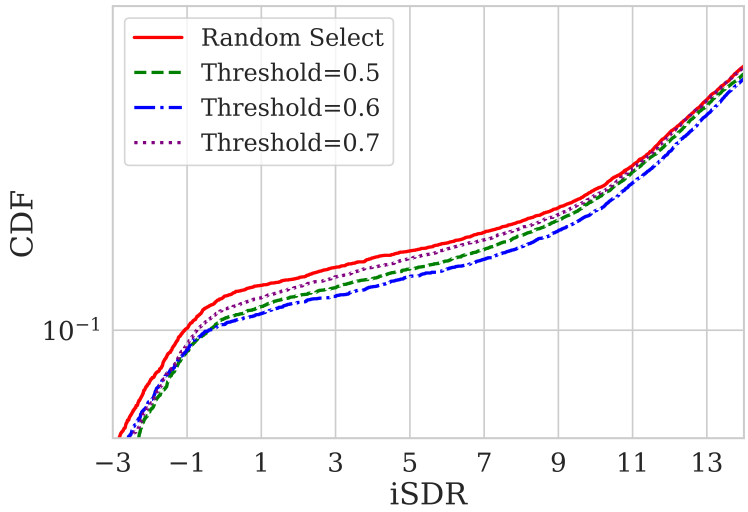


Figure 3.2: Cumulative density function of prediction results from the TSE model based on speaker similarity. Higher CDF values indicate a greater proportion of samples achieving the corresponding iSDR (dB). "Threshold" corresponds to τ_{spk}, and "Random Select" is the baseline.

3.4.3 Results

We conducted extensive experiments on the test set using various CL strategies including random selection (baseline), Gender-based, Speaker similarity-based, SDR-based, SNR-based, and Self-paced CL. The results are presented in Table 3.2. The optimal thresholds τ for each difficulty measure were carefully tuned on the development set. Below, we discuss our findings in detail.

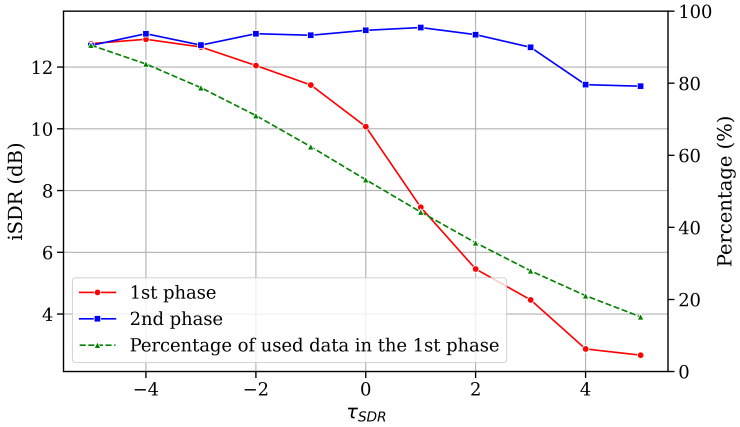


Figure 3.3: iSDR changes on the dev set with varying τ_{SDR} in two training stages, along with the data usage percentage in Stage 1.

Random Selection: Our baseline system, trained for 100 epochs with random sample selection, achieved a 12.50 dB SDR. An additional 5 epochs of training yielded only a marginal improvement of 0.07 dB, suggesting that the model had largely converged after the initial training phase.

Gender-based CL: For convenience, in Table 3.2 we use $\tau_{gender} \in$ [Same, Different] to indicate whether the gender between the target and interference speakers is the same. Surprisingly, the gender-based CL method resulted in performance worse than random selection, with the resulting iSDR in the 2nd phase being 1.17 dB lower than the baseline (11.40 dB vs. 12.57 dB). This suggests that gender alone may be too coarse a difficulty measure for effective CL in the TSE task.

Speaker Similarity-based CL: The results strongly validate the effectiveness of cosine distance-based speaker similarity as a reliable difficulty measure within our proposed CL framework. Setting τ_{spk} to 0.6 resulted in a notable improvement, achieving 13.44 dB iSDR in the 2nd phase, which represents a significant 0.87 dB improvement over the baseline. This approach utilized 83.8% of the training data in the 1st phase, suggesting an optimal balance between data selection and model exposure to challenging examples.

To gain deeper insights into how the speaker similarity-based CL method improved TSE performance, we computed iSDR for all samples in the test set and compared the model trained using our CL approach with the baseline. We performed kernel density estimation on the obtained SDRs and derived cumulative density functions (CDFs),

as illustrated in Figure 3.2. The analysis reveals that the 'Random Select' method demonstrates a higher CDF probability than our CL methods when the iSDR is low, indicating that it contains more samples with minor or negative improvement. As iSDR increases, the gap between the methods becomes smaller, which suggests that our proposed CL approach includes more samples with large improvements, particularly for challenging speaker pairs.

SDR-based CL: When the SDR threshold τ_{SDR} was set to 1 dB, we observed a significant increase in performance from 8.75 dB in the 1st phase to 13.40 dB in the 2nd phase. This approach utilized only 44.3% of the training data in the 1st phase, indicating that focusing initially on easier samples can establish a strong foundation for the model. To further validate this trend, we plotted the iSDR changes on the development set across different thresholds for both training phases in Figure 3.3, alongside the percentage of data used in the 1st phase. The results consistently show an increase in iSDR across all thresholds, confirming that gradually exposing the model to more challenging samples is highly effective for achieving better performance. Additionally, the plot highlights that the selection of the optimal threshold to balance between 'easier' and more complex data is both sensitive and crucial for maximizing performance gains.

SNR-based CL: While the speaker similarity-based CL achieved promising results, the SNR-based CL showed relatively modest improvements despite using a similar data proportion scheme (82.4% of training data). The optimal τ_{SNR} of 10 dB resulted in an iSDR of 12.99 dB, which is only 0.42 dB higher than the baseline. This comparatively limited improvement underscores the importance of designing appropriate difficulty measures for effective CL in TSE tasks.

Self-paced CL: The results using our proposed self-paced CL strategy are presented in Table 3.3. This approach employed a more sophisticated five-phase training schedule, dynamically adjusting the difficulty threshold based on the model's current performance. After completing all five phases, the model achieved an impressive 13.54 dB iSDR, representing a 0.97 dB improvement over the baseline—the largest gain across all methods we evaluated.

Particularly noteworthy is the significant improvement observed in the 4th phase, where the model only back-propagated gradients when the SNR loss was non-negative. At this stage, even without incorporating the most challenging examples,

Table 3.3: iSDR(dB) results on the test set using the self-paced CL approach with 5 training phases, as described in Section 4.2.

1st phase	2nd phase	3rd phase	4th phase	5th phase
3.30	7.79	10.93	12.66	**13.54**

Table 3.4: iSDR(dB) of CL methods with different network architectures.

Networks	random	similarity(0.6)	self-paced
Naive-BLSTM	8.09	8.30	**9.60**
SpeakerBeam[109]	9.47	9.78	**9.88**
VoiceFilter [56]	6.90	7.17	**7.54**

the model already outperformed the baseline. This suggests that the negative SNR loss may correspond to cases where the wrong speaker is extracted, and deferring such challenging cases to later phases is beneficial for overall model performance.

To evaluate the generalizability of our CL strategies across different model architectures, we conducted additional experiments using alternative TSE networks. The results, presented in Table 3.4, demonstrate that our proposed CL methods consistently outperform random selection across all architectures. The Naive-BLSTM model, which simply replaces the 4 Conformer blocks with a 2-layer 512-dimensional BLSTM while maintaining the same input and output structure, showed the most dramatic improvement with the self-paced method, gaining 1.51 dB over random selection. The SpeakerBeam and VoiceFilter architectures also benefited from our CL approaches, with consistent improvements observed across both selected CL methods.

The above experimental results demonstrate that gradually introducing more challenging samples via CL significantly improves the iSDR performance of TSE models across all tested architectures. The systematic progression from easier to more difficult examples enables models to build robust foundational capabilities before tackling complex separation scenarios, leading to more stable convergence and enhanced generalization performance.

Although the self-paced method achieves the highest performance with a remarkable 0.97 dB improvement over the baseline, its increased design complexity and computational overhead might be a trade-off for certain practical applications. The self-paced approach requires dynamic threshold adjustment and continuous performance monitoring throughout the training process, which may not be feasible in

resource-constrained environments or time-sensitive deployment scenarios. When simplicity is desired, the speaker similarity-based strategy provides a practical and effective alternative that balances performance gains with implementation efficiency. This approach offers several advantages: it requires minimal hyperparameter tuning, can be easily integrated into existing training pipelines, and provides consistent improvements across different model architectures without significant computational overhead. The speaker similarity threshold can be predetermined based on dataset characteristics, making it particularly suitable for automated training workflows.

3.4.4 Discussion: knowledge-driven vs. data-driven curriculum design.

Curriculum design can be broadly viewed as a choice between knowledge-driven and data-driven strategies. Knowledge-driven curricula use predefined difficulty measures derived from domain insights (e.g., mixture-related or speaker/acoustic factors), which are easy to interpret and inexpensive to compute, but may be sensitive to the assumed difficulty ordering and may not reflect the actual weaknesses of a specific model. In contrast, data-driven curricula define difficulty using model feedback or auxiliary estimators. In our setting, SNR-based scoring already serves as a model-driven difficulty measure that adapts to the current extraction approach, while the self-paced strategy further makes difficulty dynamic by selecting samples according to the current training loss. This aligns with the distinction between predefined and automatic curriculum learning summarized in [96], where automatic methods reduce reliance on expert-designed rules by incorporating model-dependent signals.

Beyond loss-based signals, an alternative data-driven option is to use automatic quality assessment as a alternative difficulty or filtering signal. Non-intrusive predictors such as DNSMOS or UTMOS can estimate perceptual quality without access to a clean reference, and could be used to down-weight or exclude unreliable synthetic samples, or to form an easy-to-hard curriculum based on predicted quality. Such quality-aware curricula may be particularly useful when training data are collected at scale and include diverse noise, artifacts, or low-quality synthetic speech. Exploring how quality predictors correlate with downstream TSE training dynamics, and how to integrate them into difficulty measures or schedulers, is an interesting direction for future work.

3.5 Conclusion and Future Work

This chapter presented a CL framework for target speaker extraction. By integrating different difficulty measures with progressive training and self-paced mechanisms, our approach helps the model handle challenging acoustic conditions effectively. Evaluations on the Libri2talker dataset show that our CL strategies significantly outperform traditional random sampling across various TSE architectures. The self-paced CL method achieved the best performance with a 1.04 dB improvement in iSDR over the baseline, while speaker similarity-based CL offered a good balance between performance gains and implementation complexity.

We address two fundamental research questions concerning TSE training optimization. First, we investigate how to select training data for more effective TSE training (RQ1), demonstrating that careful consideration of attributes such as gender, speaker similarity, SNR, and SDR significantly impacts model performance. Second, we explore how to strategically utilize the selected data at different training stages (RQ2), finding that CL approaches, beginning with 'easy' examples to build foundational skills before progressively introducing more challenging data, substantially enhance training outcomes and generalization capabilities.

For future work, we plan to develop adaptive difficulty measurements that evolve during training, potentially combining multiple metrics to better reflect the model's capabilities. We also aim to explore hybrid curriculum strategies that integrate self-paced learning with structured speaker similarity-based progression. Additionally, we will investigate computational efficiency improvements and conduct more detailed analysis of specific TSE performance aspects.

With the advances addressed in this chapter, the advanced TSE systems meanwhile still struggle when processing speakers with similar vocal characteristics, and the limited speaker diversity in standard datasets restricts model generalization. To address those issues, the next chapter explores combining CL strategies with synthetic interference speakers generated through voice conversion. By incorporating synthetically created voices into the training curriculum, we aim to further diversify the training data and improve the capability of the model to distinguish between similar speakers.

4

Curriculum Learning with Synthetic Speakers for Target Speaker Extraction

4.1 Introduction

Furthering the CL framework established in the last chapter, in this chapter, we introduce a novel approach that combines CL with synthetic speaker generation to enhance TSE performance [110]. The proposed approach incorporates an additional training stage that leverages synthetic interference speakers created through voice conversion techniques. This approach effectively expands the diversity of training data without requiring additional recordings, allowing TSE models to develop more robust extraction capabilities across a wider range of **speaker conditions**.

Our proposed method addresses two critical limitations in current TSE systems. First, it overcomes the constraint of limited speaker diversity in standard datasets by generating synthetic interference speakers with controlled characteristics. Second, it integrates these synthetic speakers within a structured curriculum that ensures

improved learning progression, first establishing foundational extraction capabilities before introducing increasingly challenging scenarios.

Through comprehensive experiments across multiple TSE architectures, we demonstrate that this combined approach significantly improves extraction performance compared to conventional training methods and standard CL. The results highlight the potential of synthetic speaker generation as a valuable tool for enhancing TSE systems.

4.2 Background and Motivation

4.2.1 Limitations of Existing Datasets

Standard datasets used for TSE training, such as LibriSpeech [85] and WSJ0 [111], contain a relatively limited number of unique speakers, which fails to represent the diversity of speaker characteristics encountered in real-world applications. For comparison, state-of-the-art speaker recognition systems typically train on datasets containing thousands of speakers [112] (e.g. VoxCeleb2 has more than 6,000 speakers in total), whereas TSE models are often limited to fewer than 1200 speakers, e.g., models trained on Libri2Mix or VoxCeleb subsets. From our initial experiments, this limitation particularly affects the TSE model's ability to handle challenging cases where target and interference speakers share similar vocal characteristics.

4.2.2 Synthetic Speaker Generation for Speech Processing

Synthetic data generation offers a promising approach to address the limited diversity of real speech datasets without requiring additional recordings [113, 114]. By artificially generating speech samples with controlled variations, synthetic data can expand training data distribution and improve model robustness to unseen speaker conditions. Several key technologies enable the generation of synthetic speech with specific characteristics relevant to TSE training:

Voice Conversion

Voice conversion (VC) transforms speech from a source speaker to match the characteristics of a target speaker while preserving linguistic content [115, 116]. This capability

can be particularly valuable for generating diverse interference speakers in TSE training data. Formally, speaker conditional VC can be represented as $\hat{s}_{\text{target}} = \text{VC}(s_{\text{source}}, e_{\text{target}})$, where e_{target} represents the target speaker's characteristics.

Parallel VC [117, 118] require aligned speech from source and target speakers, typically using sequence-to-sequence models with attention mechanisms to learn the mapping between speaker-specific acoustic features. Non-parallel approaches like CycleGAN-VC [119] operate without aligned data, using adversarial training to learn transformations between source and target domains while preserving content. Many-to-many VC systems extend this capability to arbitrary speaker pairs through a unified model that disentangles speaker identity from content [120]. Recent advances in zero-shot VC allow conversion to unseen target speakers using only a short reference sample [121], significantly expanding the potential speaker diversity available for TSE training.

Speaker Anonymization

Speaker anonymization (SA) provides another approach for generating synthetic speakers by transforming speech to conceal the speaker's identity while maintaining content intelligibility [122, 123]. X-vector-based anonymization [124] replaces the original speaker's x-vector [125] with a synthetic embedding derived from multiple speakers, effectively creating a new speaker identity that does not exist in the original dataset.

Neural voice scrambling apply transformations to prosodic and spectral features to obscure identity while preserving speech naturalness [126]. Orthogonal Householder neural networks perform speaker anonymization through specialized architectures that apply orthogonal transformations to speaker embeddings, preserving distance relationships while creating new speaker identities [127]. Latent space transformation approaches like SALT manipulate the latent representations of self-supervised speech models to generate anonymized speech with controlled speaker characteristics [122], enabling fine-grained control over speaker similarity.

Neural network based speech synthesis

Neural speech synthesis models enable the generation of high-quality synthetic speech with controlled characteristics [128, 129]. Neural vocoders such as WaveNet [130], HiFi-GAN [131], and WaveRNN [132] convert acoustic features to high-quality waveforms, enabling realistic speech generation from intermediate representations. End-to-end TTS (TTS) systems like Tacotron 2 [133], FastSpeech 2 [134], and VALL-E [135] generate natural speech directly from text with controllable speaker characteristics.

Zero-shot synthesis models such as XTTS [121] and NaturalSpeech 3 [128] can generate speech for unseen speakers from limited reference samples, allowing expansion of speaker diversity beyond the training set. By combining these models with controlled variation in acoustic parameters, researchers can generate diverse synthetic speech data with precisely defined characteristics.

Synthetic Speaker Applications in Speech Recognition and Verification

Synthetic speaker generation has demonstrated effectiveness across various speech processing tasks, though its application to TSE remains relatively unexplored. In ASR, synthetic speaker augmentation has been shown to improve robustness to accent variations and speaker characteristics [136, 137]. Speaker verification systems have particularly benefited from synthetic speakers that expand training data diversity, improving generalization to unseen speakers and recording conditions [138, 139].

These successful applications of synthetic speakers suggest significant potential for TSE training. By generating synthetic interference speakers with controlled similarity to target speakers, TSE training could be made more comprehensive and systematic, potentially improving performance on challenging extraction scenarios involving similar-sounding speakers or complex acoustic environments.

Why VC instead of TTS?

In principle, TTS can generate speech with controllable speaker characteristics and may further expand speaker diversity beyond the original datasets. However, for our purpose of generating realistic interferers for TSE training, we adopt VC rather than TTS for several practical reasons. First, VC keeps the linguistic content and timing of real recordings while mainly changing speaker identity, which makes the synthetic

interferers closer to real conversational speech in terms of speaking style and temporal structure. Second, VC does not require transcripts, allowing us to directly convert large amounts of existing speech data at scale. Third, when the source recordings already contain background noise and channel effects, VC tends to preserve these characteristics, whereas TTS typically produces cleaner and more homogeneous signals unless additional modeling is introduced, potentially increasing the domain mismatch to real-world mixtures. Finally, VC makes it easier to control how similar a synthetic speaker is to a given real speaker by adjusting speaker embeddings or latent interpolation. Therefore, although TTS-based augmentation is a promising direction, in this chapter we focus on VC as a more scalable and realistic way to create diverse interference speakers for TSE.

4.2.3 Motivations

Building upon our previous work on CL for TSE in Chapter 3, we recognize the potential to further enhance model performance by incorporating synthetic speakers within the curriculum. This motivation stems from two complementary observations that drive our approach.

First, regarding synthetic interference speakers, current TSE datasets suffer from limited speaker diversity that constrains model generalization. While the original two-stage curriculum approach demonstrated significant improvements by progressively introducing more challenging examples based on speaker similarity, the constraint of available speakers in real datasets limits the extent of diversity we can achieve. Synthetic speaker generation offers a solution by creating interference speakers with controlled characteristics that expand beyond the limitations of recorded datasets. In our approach, we specifically chose speech-to-speech methods, including VC and SA, over TTS approaches due to their superior performance in preserving linguistic content while transforming speaker characteristics.

Second, enhanced CL represents a natural extension of our established framework. The addition of an extra training stage with synthetic speakers creates a learning pathway where models first establish foundational extraction capabilities on real speakers before encountering the increased diversity and complexity introduced by synthetic voices. This structured approach ensures that synthetic speakers serve as an

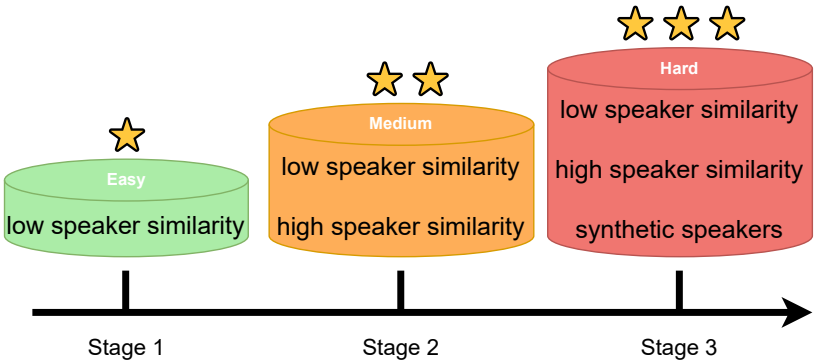


Figure 4.1: Illustration of the proposed three-stage CL approach for TSE covered by Chapter 4. Stage 1 focuses on examples with low speaker similarity, Stage 2 introduces high similarity pairs, and Stage 3 incorporates synthetic interference speakers.

advanced training challenge rather than a source of confusion during early learning stages. The systematic integration of synthetic speakers into our CL framework represents a promising direction for improving TSE model robustness that we explore in this chapter.

4.3 Synthetic Speaker Augmentation for CL Based TSE Training

4.3.1 Main Concept

Our proposed approach involves extending the previously established CL framework proposed in Chapter 3 with an additional training stage that incorporates synthetic interference speakers. As illustrated in Figure 4.1, the three-stage CL follows a systematic progression:

- Stage 1 focuses on "easier" examples where target and interference speakers have low similarity measured by the cosine distance between their respective speaker embeddings, establishing foundational extraction capabilities.
- Stage 2 introduces more challenging examples with higher speaker similarity,

allowing the model to develop more refined extraction strategies.

- Stage 3 exposes the model to both real and synthetic interference speakers, further enhancing its ability to handle diverse and complex acoustic scenarios.

This progressive training approach ensures that the model builds robust fundamentals before tackling increasingly challenging extraction tasks, with synthetic speakers representing the most complex scenario due to their increased diversity and potential for creating more challenging mixtures.

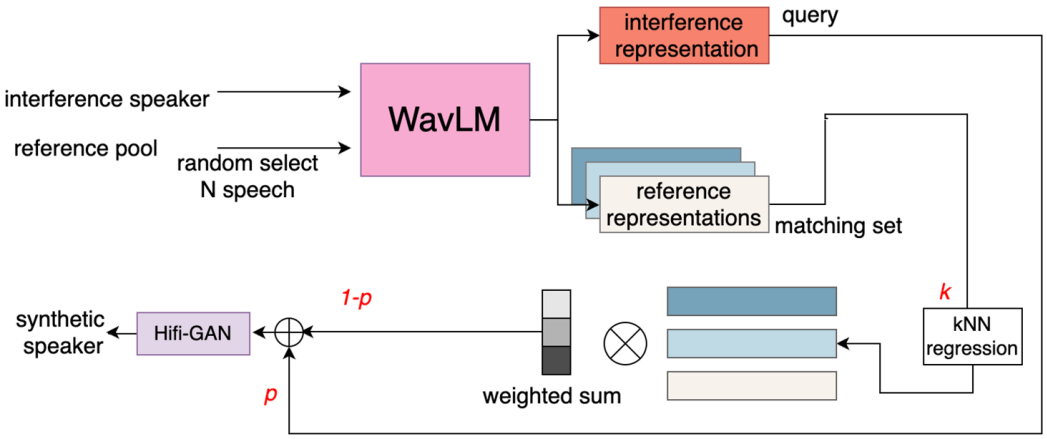


Figure 4.2: Synthetic speaker generation using the k-NN VC / SALT system. The framework consists of an WavLM encoder that extracts latent representations, a k-NN regression module that performs neighbor selection and weighted averaging, and a HiFi-GAN vocoder that generates the synthetic speech.

4.3.2 Synthetic Speaker Generation Based on Voice Conversion

For generating synthetic interference speakers, we employ a k-nearest neighbor (k-NN) based voice conversion approach, specifically the Speaker Anonymization via Latent space Transformation (SALT) system [122][1]. This framework was selected after preliminary experiments indicated its superior performance in generating diverse yet realistic speaker characteristics with proper preservation of the voice and speech-related content from the original audio.

[1]https://github.com/BakerBunker/SALT

SALT consists of three primary components working in coordinated succession to achieve effective VC. The first component is a self-supervised latent encoder based on a pre-trained WavLM encoder [140] that extracts latent representations correlated with speaker characteristics. This is followed by a k-NN regression module that performs neighbor selection and weighted averaging of reference speaker representations to create the desired speaker identity transformation. Finally, a neural vocoder based on HiFi-GAN technology [131] converts the modified latent representations back to high-quality speech waveforms with remarkable acoustic fidelity.

The WavLM encoder specifically extracts features from the third transformer layer, which have been shown to capture speaker characteristics effectively [140]. This produces a 1024-dimensional representation vector for every 20 ms of audio. The HiFi-GAN vocoder is correspondingly trained to reconstruct high-quality waveforms using these specific layer features as input, ensuring consistency between the encoding and decoding processes.

The VC process follows a systematic workflow where WavLM representations are first extracted from the original interference speech. These representations undergo careful modification through a specialized combination of k-NN selection and weighted averaging with representations from reference speakers. The resulting modified representations are then processed by the vocoder to generate synthetic speech that maintains natural speech quality while exhibiting distinct speaker characteristics. This approach ensures that the generated speakers achieve a critical balance between authenticity and differentiation, which is essential for robust evaluation scenarios involving mixtures that contain interference speakers.

4.3.3 Generation progress of synthetic interference speakers

We employ SALT to generate synthetic interference speakers through a systematic four-step process. Given an interference speaker's audio s_t in TSE, we first extract its WavLM representation WavLM(s_t). Next, we use this representation as a query to perform k nearest neighbors(k-NN) selection, identifying k-NN from N reference speakers based on cosine distance similarity.

In the third step, we assign random weights w_j to each selected neighbor's representation $\mathbf{R}_j$. These weights are sampled from a normal distribution and normalized to sum

to unity. Finally, we compute a weighted combination of the selected representations and blend it with the original speaker's representation using linear interpolation:

$$(\mathbf{R}_1, \cdots, \mathbf{R}_k) = \text{kNN}(\text{WavLM}(s_t)) \tag{4.1}$$

$$\mathbf{O} = (1 - p) \sum_{j=1}^{k} w_j \mathbf{R}_j + p\text{WavLM}(s_t) \tag{4.2}$$

where p controls the interpolation balance. The resulting representation $\mathbf{O}$ is then fed into the HiFi-GAN vocoder to produce the synthetic waveform.

This approach offers significant advantages for TSE training. The parameters k and p provide fine-grained control over speaker diversity and similarity to the original speaker. By varying these parameters and generating different random weight combinations, we can create an extensive collection of diverse synthetic interference speakers that effectively expand the training data without requiring additional recordings.

4.4 Experimental Setup

4.4.1 Dataset

We conducted experiments using the Libri2talker dataset [87], a two-talker simulated mixture dataset derived from LibriSpeech. The dataset contains 127,056 data triplets for training, each consisting of a mixture, target speech, and reference speech. For each data triplet, we generated a synthetic interference speaker using the SALT system, effectively doubling the potential training data volume (254,112 data triplets).

4.4.2 Feature Extraction and Model Configuration

All audio was processed at a 16 kHz sampling rate, with reference speech normalized to 15 seconds through padding or segmentation and mixture signals clipped to 6 seconds. The STFT was computed using a 32 ms window, 8 ms hop, and 512-point FFT. Speaker embeddings were extracted using a pre-trained ECAPA-TDNN [59].

For the main TSE system, we employed the Conformer-based architecture with four Conformer blocks from Chapter 3. This model estimates complex ratio masks [7] for the real and imaginary parts of the complex-valued spectrum. We also evaluated our approach with alternative architectures, including a naive BLSTM, SpeakerBeam [109], and VoiceFilter [56].

4.4.3 Voice Conversion Setup

For generating synthetic speakers, we used the WavLM-Base model [2] trained on the LibriSpeech dataset, extracting latent representations from the third transformer layer, which has been shown to capture speaker characteristics effectively [140]. The reference pool consisted of 20 randomly selected speakers from the LibriSpeech train-clean-100 dataset, with 50 randomly selected utterances per speaker used for feature extraction.

For each interference speaker in the training set, we generated a corresponding synthetic version using the SALT system. We explored various configurations of the hyperparameters k (number of neighbors) and p (interpolation factor), with $k \in \{4, 8, 20\}$ and $p \in \{0.2, 0.5, 0.8\}$.

4.4.4 Training Protocol of TSE

Training followed a three-stage CL approach and showed excellent results in Chapter 3.

1. **Stage 1**: 100 epochs using only examples with speaker similarity below a threshold of 0.6.
2. **Stage 2**: 5 epochs using the full training dataset.
3. **Stage 3**: Additional training with both real and synthetic speakers.

For Stage 3, we explored different ratios of real to synthetic speakers within each mini-batch, with the optimal configuration using an equal proportion of each. Training used the Adam optimizer [108] with an initial learning rate of 1×10^{-3}, gradually increasing over the first 5,000 batches and then decaying with an inverse square-root schedule to a minimum of 1×10^{-5}.

[2] https://huggingface.co/microsoft/wavlm-base

Table 4.1: Experimental results before and after adding Stage 3, which uses synthetic speakers, to the TSE model based on Conformer in [1]. Hyperparameters are k=4 and p=0.5.

TSE method	Stage2	Stage3	iSDR [dB]
Conformer w/o CL (Chapter 3)	×	×	12.57
Conformer (Chapter 3)	✓	×	13.44
Conformer (real only)	✓	✓	13.98
Conformer (real+syn)	✓	✓	**14.43**

4.5 Results and Analysis

4.5.1 Improvement from Synthetic Speakers

Table 4.1 presents the comparative results of different training approaches using the Conformer-based TSE model. The baseline without CL achieved 12.57 dB iSDR, while the two-stage CL improved this to 13.44 dB. Simply adding Stage 3 with additional training on real examples (Conformer real only) further improved performance to 13.98 dB. However, the most significant improvement came from incorporating both real and synthetic speakers in Stage 3 (Conformer real+syn), which achieved 14.43 dB iSDR, representing a substantial 0.99 dB improvement over the two-stage curriculum and 1.86 dB over the baseline without CL. These results clearly demonstrate the benefits of both CL and the integration of synthetic speakers, with their combination providing the best overall performance.

4.5.2 Impact of Real-Synthetic Speaker Ratio

To understand the optimal balance between real and synthetic speakers, we conducted experiments varying the proportion of synthetic speakers within each mini-batch. Figure 4.3 illustrates the iSDR performance as a function of this proportion, ranging from 0.0 (only real speakers) to 1.0 (only synthetic speakers).

The results indicate that a ratio of 0.5 (equal numbers of real and synthetic speakers) yields the best performance (14.43 dB). However, performance sharply declines to 4.37 dB when only synthetic speakers are used (ratio of 1.0). This dramatic drop can be attributed to a domain mismatch [141, 142], as the model initially trained on real data struggles when fine-tuned exclusively with synthetic data that differs significantly

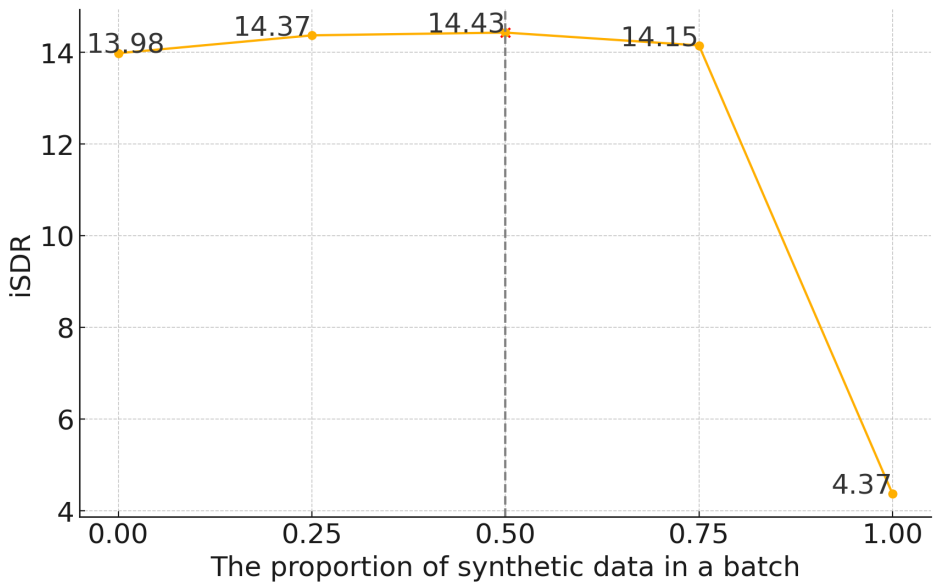


Figure 4.3: iSDR performance as a function of the proportion of synthetic speakers within a mini-batch. An x-axis value of 1.0 indicates the exclusive use of synthetic interference speakers.

Table 4.2: iSDR (dB) results of methods with varying parameters k and p.

	p=0.2	p=0.5	p=0.8
k=4	14.16	**14.43**	14.15
k=8	14.17	14.32	14.25
k=20	14.39	14.42	14.36

from the real test set distribution.

This finding highlights the importance of maintaining a balance between real and synthetic data during training, leveraging the complementary benefits of both data types while avoiding over-reliance on synthetic examples.

4.5.3 Effect of Voice Conversion Hyperparameters

The quality and diversity of generated synthetic speakers depend significantly on the k-NN and interpolation hyperparameters. Table 4.2 presents the iSDR results across various configurations of k (number of neighbors) and p (interpolation factor).

A general trend of improvement is observed as the value of k increases, with k=20

Table 4.3: Results with different numbers of synthetic speakers, in terms of its resulting size of the data and iSDR.

TSE dataset	Data size #utts	iSDR (dB)
1:1	127,000	14.43
Different k & p	127,000 × 5	14.35
Same k & p	127,000 × 5	14.49

consistently yielding better results than k=4 or k=8 across different p values. This suggests that incorporating a larger number of reference speakers in the VC process produces more diverse and effective synthetic speakers for TSE training.

The interpolation factor p=0.5 provides the most stable performance across different k values, with consistently high iSDR scores. This balanced interpolation factor maintains sufficient characteristics from the original speaker while introducing beneficial diversity through the reference speakers.

4.5.4 Impact of Synthetic Data Quantity

We further investigated whether increasing the volume of synthetic data beyond a 1:1 ratio with real data would yield additional improvements. Two approaches were tested: (1) generating data with multiple combinations of k and p parameters, and (2) keeping k and p fixed but generating multiple versions with different random weights.

Table 4.3 shows that simply increasing the amount of synthetic data to five times the original volume did not significantly improve performance beyond what was achieved with the 1:1 ratio. The "Different k & p" approach achieved 14.35 dB iSDR, while "Same k & p" with multiple weight variations reached 14.49 dB, compared to 14.43 dB with the original 1:1 ratio.

These results suggest that the current TSE model architecture may already be approaching its capacity to benefit from additional synthetic data, and that the diversity provided by the 1:1 ratio is sufficient for the current test set characteristics.

4.5.5 Generalization Across TSE Architectures

To evaluate whether our approach generalizes beyond the Conformer, we conducted experiments using three additional TSE models: Naive-BLSTM, SpeakerBeam, and

Table 4.4: iSDR results using different TSE architectures. Hyperparameters for speaker generation are k=4 and p=0.5.

TSE architecture	base	real only	real+syn
Naive-BLSTM	8.30	9.79	**10.32**
SpeakerBeam [109]	9.78	10.10	**10.51**
VoiceFilter [56]	7.17	8.66	**9.32**

VoiceFilter. Table 4.4 presents the results for each architecture under three training configurations: baseline with two-stage CL (base), Stage 3 with only real data (real only), and Stage 3 with both real and synthetic data (real+syn).

Across all architectures, the addition of Stage 3 with real and synthetic data consistently yielded the highest performance improvements. The Naive-BLSTM model improved from 8.30 dB to 10.32 dB, SpeakerBeam from 9.78 dB to 10.51 dB, and VoiceFilter from 7.17 dB to a remarkable 9.32 dB. These substantial gains across diverse architectures demonstrate the broad applicability of our approach, suggesting that the benefits of synthetic speaker integration are not tied to specific architectural design choices.

4.5.6 Visualization for Speaker Embedding Space

To investigate the speaker diversity between real and synthetic interference speakers, we employed a dimensionality reduction approach to visualize high-dimensional speaker embeddings. Figure 4.4 illustrates the interference speaker embedding space extracted from real data (Libri2Talker) and synthetic data (SALT-generated). We extracted speaker embeddings using the pre-trained ECAPA-TDNN model described in Section 4.4.2 and applied t-SNE for two-dimensional visualization.

The visualization reveals that synthetic data (red) forms a concentrated distribution in the central region of the embedding space, while real data (purple) exhibits a more dispersed pattern covering the broader embedding space. Importantly, the synthetic speakers effectively fill the core regions of the speaker embedding space and complement areas where real data coverage is limited. This complementary distribution pattern demonstrates that our SALT-generated synthetic speakers provide valuable supplementation to the original speaker diversity, enhancing the overall

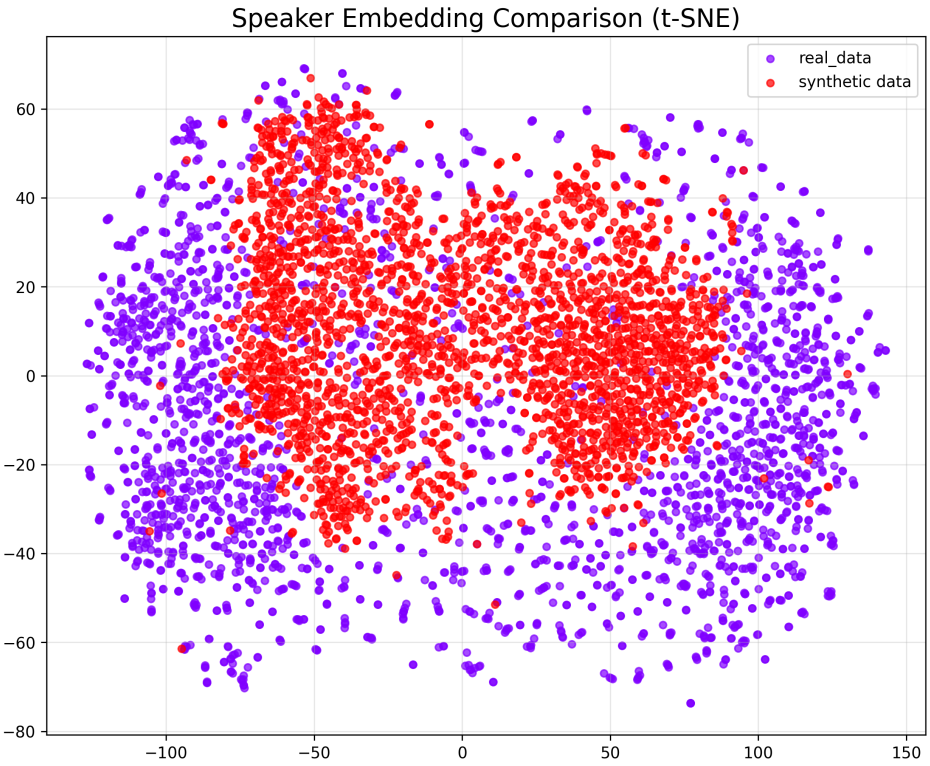


Figure 4.4: Speaker Embedding Space Visualization.

coverage of the speaker embedding space for more robust TSE training.

4.6 Discussion

The experimental results clearly demonstrate the effectiveness of incorporating synthetic interference speakers within a CL framework for TSE. Below presents several key insights derived:

First, the progressive nature of the three-stage CL enables the model to build a strong foundation on simpler examples before tackling more complex scenarios, with synthetic speakers representing an additional level of complexity that further enhances model robustness.

Second, the balance between real and synthetic data is crucial, with optimal performance achieved using a 50%-50% ratio. This suggests that while synthetic data provides valuable diversity, maintaining a connection to the real data distribution

remains essential for effective generalization.

Third, the VC hyperparameters can have a notable impact the effectiveness of synthetic speakers, with moderate interpolation factors and larger numbers of reference neighbors yielding the best results. This highlights the importance of carefully controlling the related characteristics of generated speakers to maximize their benefit for TSE training.

Finally, the consistent improvements across multiple TSE architectures indicate that our approach addresses fundamental limitations in current TSE training data rather than exploiting architecture-specific characteristics. Such applicability suggests that synthetic speaker generation could become a standard component of TSE training, complementing other data augmentation and CL strategies.

However, it is important to note that simply increasing the volume of synthetic data did not yield proportional performance improvement, suggesting that current test sets may not fully capture the potential benefits of enhanced speaker diversity. This points to the need for more challenging evaluation benchmarks that better reflect real-world diversities and complexities.

4.7 Conclusion

In this chapter, we addressed the critical research question of whether speech generative models can generate useful data for TSE training (RQ3). We presented an enhanced CL approach that incorporates synthetically generated interference speakers, extending our previous two-stage curriculum with an additional stage that integrates both real and synthetic speakers. The k-NN-based VC proved effective for generating diverse interference speakers without requiring additional recordings, addressing a fundamental limitation in current TSE datasets regarding speaker diversity. Our experimental results demonstrated that synthetic speakers can indeed provide substantial benefits when properly integrated, with the optimal approach involving an equal balance of real and synthetic speakers and carefully tuned VC parameters.

The consistent performance gains across different architectures, ranging from 1.0 to 2.2 dB iSDR improvements, highlight the broad applicability of our approach. These results establish synthetic speaker generation as a valuable technique for enhancing TSE training, complementing CL strategies to further improve extraction performance.

Future work will focus on increasing the diversity of synthetic speakers by exploring alternative generation approaches and investigating more challenging evaluation scenarios, including multi-speaker interference and speakers with greater similarity to the target. Additionally, the development of more comprehensive test sets that better reflect real-world speaker diversity will be essential for fully evaluating the potential benefits of synthetically enhanced training.

While the CL framework and synthetic speaker augmentation techniques demonstrate substantial TSE performance improvements, a crucial consideration is raised: can model robustness advance further by confronting fundamental training data limitations? Our strategic training methodologies achieve notable gains, yet remain bounded by simple combination upon existing datasets. This realization motivates exploring whether combining our training innovations with more realistic acoustic data could unlock additional performance improvements, leading us to develop a new dataset that bridges controlled experimental conditions with **real-world acoustic complexity**.

5

Libri2Vox : A Novel Dataset for Target Speaker Extraction

5.1 Introduction

Current TSE systems face a critical bottleneck in real-world deployment. Despite achieving impressive performance, these systems exhibit substantial degradation when confronted with complex acoustic environments [143]. This performance gap stems from fundamental limitations in existing training approaches and datasets that affect two critical dimensions: speaker diversity and acoustic realism. As has been established in Chapter 4, limited speaker diversity significantly constrains TSE performance, which we addressed through synthetic speaker generation to expand the range of acoustic characteristics available during training. However, speaker diversity represents only one dimension of the challenge. This chapter extends our investigation beyond speaker diversity to address the equally critical issue of acoustic background complexity. While Chapter 4 focused on enhancing the variety of speakers encountered during training

through synthetic voice generation, this chapter tackles the fundamental mismatch between controlled training environments and real-world acoustic conditions.

Existing datasets such as WSJ0-2mix-extr [144, 145] and Libri2talker [87] consist entirely of artificially mixed clean speech recordings. These controlled mixtures fail to capture the complex interplay of reverberation, channel effects, background chatter, and dynamic noise patterns that characterize real-world acoustic scenes [146]. The speaker diversity limitation further compounds this challenge. With WSJ0-2mix-extr containing only 101 speakers and Libri2talker limited to 1,172 speakers, current datasets provide insufficient coverage of the vast acoustic and linguistic variations encountered in practical applications. This constraint becomes particularly evident when comparing TSE training resources to speaker recognition systems, which routinely leverage datasets with over 6,000 speakers to achieve robust performance.

Our CL framework (Chapter 3) and synthetic speaker generation methods (Chapter 4) have demonstrated that strategic data utilization and augmentation can significantly improve TSE performance. However, these approaches still operate within the constraints of artificially constructed training data. This raises a crucial question: what if we could combine the benefits of our training innovations with more realistic acoustic data? This insight motivates the development of Libri2Vox, a novel dataset that bridges the gap between controlled experimental conditions and real-world complexity. By strategically combining clean target speech from LibriTTS [86] with naturally noisy interference speech from VoxCeleb2 [112], we create a training environment that maintains the supervision necessary for TSE while introducing authentic acoustic variations. VoxCeleb2's recordings capture "in the wild" across diverse environments from YouTube recordings, providing over 6,000 speakers with genuine background noise, reverberation, and channel characteristics.

Building upon our established CL framework proposed in Chapter 3 and synthetic speaker generation techniques in Chapter 4, this chapter demonstrates how Libri2Vox serves as the missing piece in our comprehensive approach to robust TSE. The integration of realistic acoustic conditions with our progressive training strategies creates a synergistic effect, enabling models to learn from both methodological innovations and enhanced data quality.

5.2 Background and Motivation

5.2.1 Limitations of Existing Datasets

Standard datasets used for TSE training, such as ones described in Section 2.7 (namely WSJ0-2mix-extr [144, 145], Libri2mix [74], and Libri2talker [87]), exhibit fundamental limitations that constrain model performance in real-world applications, which manifest in three dimensions: speaker diversity, acoustic realism, and environmental variability.

Regarding speaker diversity, current TSE datasets severely under-represent the variety of voices encountered in practical deployments. WSJ0-2mix-extr contains only 101 speakers, while Libri2mix and Libri2talker, despite being larger, are limited to 1,172 speakers from LibriSpeech [85]. This constraint becomes particularly apparent when compared to state-of-the-art speaker recognition systems, which routinely train on datasets such as VoxCeleb2 containing over 6,000 speakers [112]. The limited speaker pool restricts the model's exposure to diverse vocal characteristics, including variations in pitch, timbre, accent, speaking rate, and prosodic patterns. This limitation particularly affects the model's ability to handle challenging extraction scenarios where target and interference speakers exhibit similar acoustic properties—a common occurrence in real-world applications such as family conversations or regional gatherings.

The lack of real-world acoustic realism is also a big challenge. All existing TSE datasets consist of artificially mixed speech recorded under controlled studio conditions, lacking the natural acoustic representation in real environments such as reverberation, dynamic background noise, and channel effects from diverse recording devices. This artificial construction creates a substantial mismatch with real-world deployment scenarios where TSE systems must function across varied acoustic environments—from quiet offices to noisy restaurants. Consequently, models trained on such clean data perform well on benchmarks with similar data conditions, but the performance can deteriorate dramatically when confronted with complex acoustic conditions, as they learn dataset-specific patterns rather than robust extraction capabilities. Addressing these limitations requires novel data that incorporates real-world acoustic variability while maintaining clean target speech as ground truth for supervised TSE training.

5.3 Libri2Vox Dataset

Current TSE systems confront two primary challenges: constrained speaker diversity and synthetic acoustic environments. Existing TSE datasets, as illustrated in Table 5.1, typically encompass only hundreds to several thousand speakers, substantially constraining model generalization capabilities. This constraint becomes especially evident when systems encounter speakers or acoustic conditions outside the training distribution.

Speaker diversity serves a fundamental role in TSE performance. A comprehensive speaker set allows models to learn robust representations across various speech characteristics, including accent variations, speaking styles, and vocal qualities. Our analysis indicates that expanding the speaker sets directly correlates with enhanced model generalization and performance metrics.

Beyond speaker diversity, acoustic condition represents another compelling factor. Most existing datasets for standard TSE training and evaluation are constructed by artificially combining target and interference speech with noise, which fails to fully capture how noise and speech mixtures occur in real environments. The complexity of how background noise interacts with speech, including varying distances between the speakers and noise sources, or overlapped voices during conversations, makes it challenging to simulate the true nature of how speech mixtures are created in real-time cases. This is where current datasets often fall short. Meanwhile, datasets that have been widely utilized for other tasks (such as VoxCeleb2 used in this study) often contain real-world recordings, some of which include background noise from various environments, making them a closer reflection of the conditions under which TSE models are expected to operate.

To address these limitations, we utilize VoxCeleb2 recordings as interference sources. These recordings inherently contain diverse, real-world acoustic conditions, providing more realistic training scenarios for TSE models compared to traditional artificially mixed clean speech. This approach marks a significant step toward more realistic scenarios, though it still focuses on 2-speaker mixtures rather than the full complexity of unconstrained real-world environments.

Table 5.1: Comparison of WSJ0-2mix-extr, Libri2mix, and Libri2talker Datasets.

Dataset	# Speakers	# Utterances	Duration(h)
WSJ0-2mix-extr	101 (train) 101 (val) 18 (test)	20,000 (train) 5,000 (val) 3,000 (test)	30 8 5
Libri2mix	921 (train-360) 251 (train-100) 40 (val) 40 (test)	50,800 (train-360) 13,900 (train-100) 3,000 val 3,000 test	212 58 11 11
Libri2talker	1,172 (train/val) 1,172 (val) 40 (test)	127,056 (train) 2,344 (val) 6,000 (test)	460 8 22

5.3.1 Data construction

In developing the *Libri2Vox* database, as illustrated in Fig. 5.1, we employed several key steps.

Pre-processing and mixing

Each audio segment is randomly divided into 6-second segments. Segments shorter than six seconds are zero-padded at the end. The dataset is initially processed through the following steps:

- All audio data is pre-processed via amplitude normalization using **sv56**[1] with a scaling factor of -26 dB.
- In *LibriTTS*, target speech shorter than 2s is eliminated, except for the train set. Speakers with fewer than 3 utterances are excluded, resulting in the removal of 31 speakers out of 1,151, retaining 1,120 speakers in total.
- For reference speech processing, we implement the following strategy: If the original utterance is shorter than 10s, we randomly select additional utterances from the same speaker and concatenate them sequentially until the combined duration exceeds 10s (while ensuring the total duration remains below 15s). If the original utterance is already longer than 10s but shorter than 15s, we utilize

[1]https://github.com/foss-for-synopsys-dwc-arc-processors/G722/tree/master/sv56

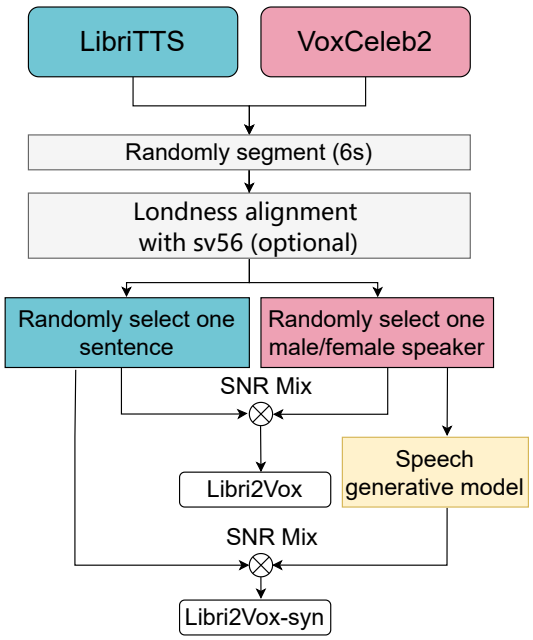


Figure 5.1: Data generation framework of Libri2Vox and its synthetic version.

it as is. If the original utterance exceeds 15s, we truncate it to 15s. This approach ensures sufficient reference material for speaker embedding extraction while maintaining reasonable processing lengths.

For the dataset composition, we employ VoxCeleb2 as the interference speaker dataset and LibriTTS as the target speaker dataset. VoxCeleb2 is a large-scale speaker recognition dataset that was collected "in the wild", indicating that the speech segments are naturally corrupted by real-world noise such as laughter, cross-talk, channel effects, music, and other background sounds. This introduces an element of realism, making the dataset particularly valuable for training models that need to handle noisy environments. VoxCeleb2 is also multilingual, featuring speech from speakers of 145 different nationalities and covering a broad range of accents, ages, and languages.

Concerning speaker diversity characteristics, Libri2Vox inherits rich demographic variety from both source datasets. From VoxCeleb2, we preserve the diversity in nationalities (145 different countries), with a relatively balanced gender distribution (61% male, 39% female). The dataset captures diverse age groups spanning from young adults to seniors across various professions and backgrounds. Speaking style diversity

is particularly robust in the VoxCeleb2 portion, featuring speech from speakers of different ethnicities captured 'in the wild', with background chatter, overlapping speech, public speeches, and entertainment segments, each with varied speech rates and natural prosodic patterns. The natural recording environments in VoxCeleb2 contribute to diversity in speech rates, emotional states, and speaking styles—ranging from formal interviews to casual conversations. Speech segments exhibit different emotional qualities including excitement during red carpet events, neutral tones in studio recordings, and various natural speech patterns. The LibriTTS portion contributes clean, read speech with 24kHz sampling rate from 2,456 speakers, providing high fidelity target speech. While LibriTTS primarily contains read speech with less emotional variation, it includes some accent diversity, particularly with British, Scottish, Welsh, and Irish accents identified in specialized subsets. These diverse acoustic characteristics create a challenging yet realistic environment for TSE systems to operate under real-world conditions. The variety of recording settings (red carpets, outdoor stadiums, indoor studios, etc.) further ensures acoustic diversity beyond controlled laboratory conditions. The combination of these two datasets in Libri2Vox creates a training environment with both controlled, high-quality target speech and diverse, real-world interference conditions.

One potential issue with VoxCeleb2 for this study is that the speech segments are predominantly sourced under uncontrolled conditions. However, conversely, the inherent noise and variations in VoxCeleb2 makes it an ideal source for interference speakers. By employing VoxCeleb2 as the interference speaker source, we can exploit the real-world noise captured during the original recordings, rather than relying on artificially added noise in post-processing. This enables the dataset to more accurately simulate real acoustic conditions in TSE.

VoxCeleb2 includes both a development (dev) set and evaluation set. The dev set contains 5,994 speakers, with a total of 1,092,009 utterances, while the evaluation set includes 118 speakers and 36,237 utterances. The entire dataset contains over 1.1 million utterances, making it a rich resource for training models on diverse speaker characteristics. The average duration of the utterances in VoxCeleb2 is approximately 7.8s. We randomly select 94 speakers from the dev set of VoxCeleb2 to form the validation set and retain 5,900 speakers for training.

The other data source acquired for this study is LibriTTS [86]. It is a refined version

Table 5.2: Libri2Vox statistics. "Total" column shows the total number of utterances combined with corresponding LibriTTS and VoxCeleb2 sets.

Set	#. Speakers		#. Utterances	Duration (h)
	LibriTTS	VoxCeleb2	Total	
Training	1,151	5,900	149,681	250
Validation	40	94	4,664	7.77
Testing	39	118	3,931	6.55

of LibriSpeech [85], where a small portion of the noisy utterances from the latter was eliminated to ensure cleaner speech for tasks such as TTS.

For each target speech from LibriTTS, we randomly select one male and one female interference speaker from VoxCeleb2 in an alternating manner. This indicates that for every target speech, we first choose a male interference speaker then a female interference speaker, ensuring a balanced gender distribution across the dataset. The mixture is then generated by adding the target speech to the interference speech with an additional scaling factor α, which is sampled uniformly from the SNR range of [-5, 5] dB. The same mixing procedure applies for the validation and test sets, with the average SNR across the train, validation, and test sets being approximately -0.11 dB.

5.3.2 Statistics of Libri2Vox

Libri2Vox was pre-partitioned following collection and processing, as demonstrated in Table 5.2. For the training set, *LibriTTS* contributes 1,151 speakers. Each *LibriTTS* utterance produces one corresponding triplets (mixture, reference, target). VoxCeleb2 provides 5,900 interference speakers for this partition, making the training set contain 149,681 triplets, equivalent to 250 hours of mixture data. The validation set comprises 40 *LibriTTS* speakers (approximately 7.77 hours) and 94 VoxCeleb2 speakers, totaling 118 speakers and 4664 triplets. The test set encompasses 39 *LibriTTS* speakers (approximately 6.55 hours) and 118 VoxCeleb2 speakers, with a total of 157 speakers and 3931 triplets.

5.4 Synthetic Libri2Vox Dataset

5.4.1 Why synthetic speakers?

While Libri2Vox can provide substantial speaker diversity through VoxCeleb2, with over 6,000 speakers — significantly exceeding the 1,172 speakers in LibriSpeech-based datasets — our experimental analysis revealed specific limitations that synthetic speaker augmentation can address. Despite VoxCeleb2's large speaker pool, the distribution of speaker characteristics is not uniform across acoustic conditions. Many speakers appear in similar recording environments or share comparable vocal characteristics within specific demographic groups, creating coverage gaps in the speaker-acoustic space [112, 147].

Our analysis of the VoxCeleb2 interference speakers showed that certain challenging speaker combinations remain underrepresented. Specifically, speakers with highly similar vocal characteristics to our LibriTTS target speakers are relatively rare, particularly within the same gender and age groups. This scarcity limits the model's exposure to the most challenging extraction scenarios during training. Furthermore, the fixed nature of real recordings prevents us from creating specific speaker-similarity conditions that would optimally challenge the extraction model at different stages of CL [101, 100].

Synthetic speaker generation addresses these limitations by enabling controlled creation of interference speakers with precise similarity levels to target speakers [116, 148]. Unlike simply adding more real speakers, synthetic generation allows us to strategically fill gaps in the speaker-acoustic space, ensuring comprehensive coverage of challenging extraction scenarios [149, 138]. This targeted augmentation is particularly valuable when combined with our CL framework, as we can generate speakers with gradually increasing similarity to systematically challenge the model throughout training progression as shown in the Chapter 4.

5.4.2 Synthetic interference speaker generation

Building upon Chapter 4 where we introduced SALT for synthetic speaker generation, Chapter 5 extends our approach by incorporating an additional generation method to further enhance speaker diversity. While Chapter 4 focused on using SALT within a

CL framework to improve TSE performance, Chapter 5 addresses the need for even greater acoustic variability by combining multiple synthetic generation techniques with realistic acoustic environments.

Specifically, we present two different generation methods employed to produce synthetic interference speakers that are distinct from the real speakers selected from VoxCeleb dataset: SALT [122] and SynVox2 [150]. In Chapter 4, we used SALT to generate synthetic speakers as part of our three-stage CL approach. Now in Chapter 5, we continue to use SALT but additionally introduce SynVox2 to produce even more diverse synthetic speakers. This dual-method approach allows us to create a richer variety of synthetic interference speakers, addressing the limitations of using a single generation method. Since the description of SALT already appears in Section 4.3.2, below we provide a description of the additional synthetic speaker generation method SynVox2.

SynVox2

SynVox2 was developed for speaker anonymization using an orthogonal Householder neural network (OHNN) [151, 150]. The framework operates through the following three essential components:

- **Disentanglement:** Speech characteristics are extracted using specialized encoders. The Yet Another Algorithm for Pitch Tracking (YAAPT) [152] extracts the fundamental frequency (F0), while the ECAPA-TDNN [59] speaker encoder, trained on VoxCeleb2, produces 192-dimensional speaker identity embeddings. Additionally, a Hidden Unit BERT (HuBERT)-based soft content encoder, fine-tuned on LibriTTS-train-clean-100 from a pre-trained HuBERT model [153], captures linguistic content information.

- **Anonymization:** The system utilizes an OHNN-based anonymizer [151] that transforms original speaker embeddings into anonymized representations through multiple orthogonal Householder transformation layers. The network employs randomly initialized weights and is optimized using classification and distance-based loss functions to ensure the anonymized speaker identities are distinct from both the original and other anonymized speakers.

- **Generation:** The synthesis stage incorporates the extracted content features, F0 information, and anonymized speaker embeddings into a HiFi-GAN model trained on LibriTTS-train-clean-100. This integration generates high-quality anonymized speech waveforms that preserve natural speech characteristics while ensuring distinct speaker identities.

5.4.3 Synthetic data and statistics

Since the synthetic data produced for Libri2Vox-syn only marginally alters the duration of the original VoxCeleb2 recordings, and we randomly select 6-second audio segments for all speakers, the statistics for the Libri2Vox-syn dataset remain identical to those of the original Libri2Vox dataset. This indicates that the number of speakers, total number of utterances, and the total duration of the dataset are consistent across both the real and synthetic versions.

5.5 Experiments

The aim of our experiments is to address two key questions related to TSE using both real and synthetic data:

- **Q1:** How effective is *Libri2Vox* compared to *Libri2Talker* in enhancing TSE performance?
- **Q2:** Does the employment of synthetic speakers in *Libri2Vox-syn* dataset further improve TSE performance, particularly in noisy environments, compared with using models trained only on real speakers?

To answer these questions, we performed a series of experiments using the aforementioned TSE models mentioned in Section 2.3, trained on different combinations of real and synthetic data, with alternated noisy data augmentation. We assess these models on both the *Libri2Talker* and *Libri2Vox* test sets to comprehensively evaluate their performance across different conditions. In addition to the main experiments, we performed several ablation studies to investigate the impact of specific components and configurations on the performance of the TSE models. These ablation studies were designed to provide further insights into the contributions of individual factors such as

the inclusion of synthetic speakers, ratio of real to synthetic data, and effects of noise augmentation.

5.5.1 Experimental setup and model configurations

All TSE models were trained using a custom learning rate scheduler, designed to adjust the learning rate dynamically based on the number of steps. Each step corresponds to one mini-batch, where the mini-batch size was set to 48. The initial learning rate was set to 1×10^{-3}, with a minimum threshold of 1×10^{-5}. The learning rate was warmed up linearly for the first 5000 steps (covering approximately 104,000 samples), after which it followed an inverse square root decay based on the step count. Specifically, after the warm-up phase, the learning rate decayed proportionally to $(\text{warmup_steps}/\text{global_step})^{0.5}$. This dynamic learning rate schedule allowed for smooth transitions during training while avoiding rapid drops in learning rate that could destabilize the optimization process. The Adam optimizer was utilized with the default settings of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. No additional data augmentation, feature normalization, or input trimming was applied during training. All experiments were performed on an Nvidia Tesla A100 GPU, with each model trained for three independent runs using different random seeds. The final results are reported as the average across these runs to ensure robustness and minimize the effects of random initialization. The sampling frequency of the speech waveform was set to 16 kHz. The STFT parameters for the mixture signal included a window length of 32 ms and hop size of 8 ms with a 512-point FFT.

5.5.2 Training strategy

Noisy data augmentation

Our approach incorporated noise samples from the DNS Challenge dataset [154]. Throughout the training process, we applied a 50% probability of augmenting the target speaker's audio with a randomly chosen 6-second noise clip. The signal-to-noise ratio (SNR) for this augmentation was randomly drawn from a uniform distribution spanning $[-5, 10]$ dB.

This noise augmentation strategy aimed to investigate potential improvements

in model robustness, particularly given that VoxCeleb2 inherently contains various environmental noises. Rather than embedding noise directly into the Libri2Vox dataset, we opted for dynamic augmentation during training. This design choice ensures dataset consistency and adaptability, enabling controlled experiments on the impact of supplementary noise while retaining the authentic acoustic characteristics inherent in the VoxCeleb2 recordings.

CL strategy

To improve the model's ability to distinguish between target and interference speakers with varying degrees of similarity, we employed a CL strategy. Following our previous research in the Chapter 4, the training process is partitioned into three stages, as illustrated in Fig. 4.1:

- **Stage 1:** At this stage, the training data comprises primarily of target and interference speaker pairs with low similarity. The objective at this stage is to enable the model to focus on simpler tasks, thus establishing a solid foundation for learning speaker characteristics.
- **Stage 2:** At this stage, the model is exposed to speaker pairs with higher similarity, which gradually increases the complexity of the task.
- **Stage 3:** Finally, at this stage, in addition to low and high similarity speaker pairs, synthetic interference speakers are introduced to further diversify the data and improve the model's generalization capability.

5.6 Results and Discussion

5.6.1 Impact of Libri2Vox

Utilizing Libri2Vox and its synthetic version, we performed extensive experiments to assess the impact of synthetic data integration with CL on TSE performance, with main results presented in Table 5.3. We employed a CL training strategy with three distinct stages and evaluated eight configurations for each model, including baseline without CL, single-stage CL with real data, two-stage CL with real data, three-stage

Table 5.3: Table 5.3: iSDR(0.07 dB), iPESQ(1.514), iSTOI(0.734), iDNSMOS(2.356), results of three stage methods for Libri2Vox test set. Values are reported as mean ± 95% confidence interval half-width. values in '()' represent baseline performance metrics of original mixture signals compared **WITH** the clean target. "Real only" means use of only real data with cosine similarity less than 0.5 in 1st stage (about 71% of all data).

Model	Method	Stage 1	Stage 2	Stage 3	iSDR	iPESQ	iSTOI	iDNSMOS
Conformer	w/o CL (Real only)	✓			15.42 ± 0.076	1.325 ± 0.010	0.198 ± 0.002	0.650 ± 0.012
	w/o CL (Real+SynVox2)	✓			15.46 ± 0.073	1.295 ± 0.009	0.196 ± 0.002	0.640 ± 0.012
	w/o CL (Real+SALT)	✓			15.43 ± 0.069	1.287 ± 0.008	0.194 ± 0.002	0.635 ± 0.011
	w/ 1-stage CL (Real only)	✓			15.39 ± 0.075	1.401 ± 0.010	0.203 ± 0.002	0.664 ± 0.012
	w/ 2-stage CL (Real only)	✓	✓		15.89 ± 0.074	1.424 ± 0.010	0.205 ± 0.002	0.664 ± 0.011
	w/ 3-stage CL (Real only)	✓	✓	✓	16.01 ± 0.072	1.420 ± 0.010	0.205 ± 0.002	0.664 ± 0.011
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	16.20 ± 0.072	1.470 ± 0.010	0.206 ± 0.002	0.692 ± 0.012
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	16.20 ± 0.072	1.466 ± 0.010	0.207 ± 0.002	0.683 ± 0.012
BLSTM	w/o CL (Real only)	✓			12.34 ± 0.085	0.852 ± 0.008	0.157 ± 0.002	0.495 ± 0.012
	w/o CL (Real+SynVox2)	✓			12.53 ± 0.082	0.903 ± 0.009	0.164 ± 0.002	0.501 ± 0.012
	w/o CL (Real+SALT)	✓			12.54 ± 0.080	0.885 ± 0.009	0.161 ± 0.002	0.491 ± 0.011
	w/ 1-stage CL (Real only)	✓			11.73 ± 0.089	0.794 ± 0.008	0.149 ± 0.002	0.473 ± 0.012
	w/ 2-stage CL (Real only)	✓	✓		12.50 ± 0.084	0.922 ± 0.009	0.162 ± 0.002	0.520 ± 0.012
	w/ 3-stage CL (Real only)	✓	✓	✓	12.65 ± 0.083	0.932 ± 0.009	0.165 ± 0.002	0.531 ± 0.012
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	13.07 ± 0.082	0.997 ± 0.010	0.171 ± 0.002	0.554 ± 0.012
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	13.06 ± 0.082	0.986 ± 0.010	0.170 ± 0.002	0.544 ± 0.012
SpeakerBeam	w/o CL (Real only)	✓			11.62 ± 0.082	0.804 ± 0.008	0.150 ± 0.002	0.489 ± 0.012
	w/o CL (Real+SynVox2)	✓			12.01 ± 0.078	0.915 ± 0.009	0.156 ± 0.002	0.503 ± 0.012
	w/o CL (Real+SALT)	✓			12.07 ± 0.078	0.887 ± 0.009	0.153 ± 0.002	0.501 ± 0.012
	w/ 1-stage CL (Real only)	✓			11.17 ± 0.087	0.753 ± 0.008	0.144 ± 0.002	0.473 ± 0.012
	w/ 2-stage CL (Real only)	✓	✓		11.76 ± 0.081	0.869 ± 0.009	0.159 ± 0.002	0.512 ± 0.012
	w/ 3-stage CL (Real only)	✓	✓	✓	11.88 ± 0.079	0.899 ± 0.009	0.163 ± 0.002	0.519 ± 0.012
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	12.35 ± 0.077	0.974 ± 0.010	0.171 ± 0.002	0.547 ± 0.012
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	12.34 ± 0.077	0.959 ± 0.010	0.168 ± 0.002	0.533 ± 0.012
Voicefilter	w/o CL (Real only)	✓			11.92 ± 0.074	0.818 ± 0.008	0.158 ± 0.002	0.453 ± 0.012
	w/o CL (Real+SynVox2)	✓			11.41 ± 0.071	0.747 ± 0.007	0.149 ± 0.002	0.402 ± 0.012
	w/o CL (Real+SALT)	✓			11.46 ± 0.067	0.728 ± 0.007	0.147 ± 0.002	0.398 ± 0.012
	w/ 1-stage CL (Real only)	✓			11.34 ± 0.072	0.709 ± 0.007	0.149 ± 0.002	0.414 ± 0.012
	w/ 2-stage CL (Real only)	✓	✓		12.10 ± 0.075	0.847 ± 0.008	0.161 ± 0.002	0.463 ± 0.012
	w/ 3-stage CL (Real only)	✓	✓	✓	12.15 ± 0.076	0.856 ± 0.008	0.164 ± 0.002	0.468 ± 0.012
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	12.39 ± 0.076	0.927 ± 0.009	0.169 ± 0.002	0.493 ± 0.012
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	12.26 ± 0.076	0.908 ± 0.009	0.166 ± 0.002	0.476 ± 0.012

CL with real data, and two variants of three-stage CL incorporating synthetic data (SynVox2 or SALT).

The results demonstrate consistent performance improvements across all models when employing CL strategies. Conformer exhibited particularly notable gains, with iSDR improving from a baseline of 15.42 to 15.89 dB using two-stage CL with real data. The introduction of synthetic data in the third stage further enhanced performance, achieving 16.20 dB with both SynVox2 and SALT variants. Importantly, this improvement exceeded the control condition using only real data (16.01 dB), confirming that the benefits derive from the synthetic data rather than extended training epochs. To isolate the impact of synthetic data from potential benefits of extended

training, we included a controlled condition under which the third stage continued with real data only (w/ 3-stage CL (Real only)). The results clearly demonstrate that while additional training epochs with real data provided slight improvements, they were inferior to the gains achieved through synthetic data integration, validating the effectiveness of introducing synthetic data into the training pipeline.

To illustrate the benefits of CL in using synthetic data, we compared different integration strategies. Starting with 50% real and 50% synthetic data at stage 1 only (e.g., w/o CL (Real+SynVox2)) does not exhibit substantial improvements compared with using synthetic and real data in Stage 3 (e.g., w/ 3-stage CL (Real + SynVox2)), while CL further enhances the performance of the latter setup.

While BLSTM, SpeakerBeam, and Voicefilter exhibited lower absolute performance compared with Conformer, they displayed proportionally larger relative gains by incorporating CL. BLSTM achieved a 0.73 dB improvement (from 12.34 to 13.07 dB), while SpeakerBeam exhibited gains of 0.73 dB (from 11.62 to 12.35 dB), and Voicefilter demonstrated an improvement of 0.47 dB (from 11.92 to 12.39 dB). Conformer exhibited particularly notable gains, with an improvement of 0.78 dB (from 15.42 dB without CL to 16.20 dB with 3-stage CL incorporating synthetic speakers) over conventional random sampling approaches. Notably, both synthetic data approaches (SynVox2 and SALT) produced comparable improvements, likely due to their shared use of the HiFi-GAN vocoder architecture. These results collectively demonstrate the effectiveness of combining synthetic data augmentation with CL for enhancing TSE system performance.

The experimental results also demonstrate improvements across perceptual metrics. While intelligibility metrics exhibited modest improvements, with iSTOI increasing only marginally from 0.198 to 0.207 for Conformer with SALT data, quality-related metrics demonstrated substantially more significant gains. The iPESQ improvements were particularly notable, with Conformer displaying an increase from 1.325 to 1.470 (11% improvement) when using SynVox2 synthetic data, and similar patterns were observed across all architectures.

Likewise, iDNSMOS scores improved considerably from 0.650 to 0.692 (6.5% improvement) for Conformer with SynVox2. This finding demonstrates that our approach predominantly enhances perceptual quality aspects of extracted speech rather than intelligibility. This suggests that the synthetic data integration through CL

Table 5.4: Downstream ASR performance on the Libri2Vox test set using faster-whisper large-v3 (beam size 5), reported as mean WER (%). Lower is better. We compare the TSE-baseline trained w/o CL (Real only) and the best 3-stage curriculum model from Table 5.3, w/ 3-stage CL (Real + SynVox2).

Architecture	w/o CL (Real only)	w/ 3-stage CL (Real + SynVox2)
Conformer	28.8	28.3
BLSTM	31.6	30.5
SpeakerBeam	41.4	37.1
VoiceFilter	30.8	30.1
Mixture (no TSE)	72.5	

particularly addresses distortions and artifacts that affect perceived speech quality, which is often more critical in real-world applications.

To further verify that the gains in Table 5.3 are not due to random fluctuations, we additionally conduct a significance test by reporting a two-sided 95% confidence interval (CI) for the mean improvement using Student's t distribution, $\bar{x} \pm t_{0.975,n-1} S/\sqrt{n}$, where the CI half-width equals $t_{\text{half_width}} = t_{0.975,n-1} S/\sqrt{n}$. When comparing two systems, if the lower bound of one CI exceeds the upper bound of the other (non-overlapping CIs), we treat the difference as statistically significant. For example, Conformer with w/ 3-stage CL (Real + SynVox2) achieves 16.20 ± 0.072 iSDR, while w/ 3-stage CL (Real only) obtains 16.01 ± 0.072; their CIs are $[16.128, 16.272]$ and $[15.938, 16.082]$, respectively, indicating a significant gain even though the absolute difference is only $\sim$0.2 dB. Compared with the random-sampling baseline w/o CL (Real only) (15.42 ± 0.076), our approach yields an even clearer improvement. Similar CI separations can also be observed on other metrics, demonstrating consistent and statistically significant gains of the proposed strategy.

5.6.2 Downstream ASR evaluation.

We evaluate downstream ASR performance on the Libri2Vox test set using faster-whisper large-v3 [155] with beam size 5, and report mean word error rate (WER) in Table 5.4, where lower is better. As a baseline, we directly transcribe the raw mixtures and obtain a mean WER of 72.5%. We then apply TSE to each mixture together with its enrollment reference and transcribe the enhanced output using the same ASR

configuration. Across all four architectures, both TSE systems substantially improve ASR accuracy compared to transcribing mixtures directly. Specifically, the TSE-baseline trained w/o CL (Real only) achieves mean WERs of 28.8% (Conformer), 31.6% (BLSTM), 41.4% (SpeakerBeam), and 30.8% (VoiceFilter). Using the best curriculum system from Table 5.3, w/ 3-stage CL (Real + SynVox2), further reduces WER to 28.3%, 30.5%, 37.1%, and 30.1%, respectively. Overall, applying TSE before ASR reduces WER by about 31-44 absolute percentage points relative to the mixture baseline, and the curriculum-based model consistently improves over the non-curriculum baseline, with the largest gain observed for SpeakerBeam.

5.6.3 Evaluation on real-world recordings

To further verify the advantages of TSE models trained on Libri2Vox in realistic environments, we performed additional experiments using the ICASSP 2021 Deep Noise Suppression Challenge test set [154]. This test set comprises real-world recordings collected in various acoustic environments, such as restaurants, cafeterias, and public transportation, captured using mobile devices and other recording equipment.

Since the real recordings portion of the test set comprises original reverberant speech samples, and our research does not focus on far-field speech and reverberation issues, we first eliminated far-field reverberant speech samples using the official meta information of each sample. We also excluded speech samples marked with musical and emotional elements, resulting in 87 real recordings with a total duration of approximately 22.7 minutes of speech audio uttered by 13 speakers.

Since these real-world recordings do not include clean oracle reference signals, traditional reference-based metrics such as SDR, cannot be calculated. Instead, we employed DNSMOS since it does not require reference signals. In addition to the TSE models trained on the Libri2Vox dataset, we also employed TSE models trained using the Libri2Talker dataset for comparisons.

Table 5.5 presents the DNSMOS results for different model configurations when tested on real-world recordings. The results clearly demonstrate that models trained solely on Libri2Talker perform poorly on real recordings, achieving a negative DNSMOS gain of -0.666 for BLSTM. This poor performance persists even with noise augmentation, yielding only -0.421. In contrast, models trained with Libri2Vox achieve

Table 5.5: DNSMOS results on real-world recordings from ICASSP 2021 DNS Challenge test set. Higher values indicate better performance. (Mixture DNSMOS:2.270)

	Training Data	iDNSMOS
Conformer	Libri2Talker	-0.722
	Libri2Talker+Noise	-0.534
	Libri2Vox	-0.047
	Libri2Vox+Noise	0.017
	Libri2Vox & 3-stage (Real + SynVox2)	**0.186**
BLSTM	Libri2Talker	-0.666
	Libri2Talker+Noise	-0.421
	Libri2Vox	-0.081
	Libri2Vox+Noise	0.139
	Libri2Vox & 3-stage (Real + SynVox2)	**0.190**
SpeakerBeam	Libri2Talker	-0.701
	Libri2Talker+Noise	-0.542
	Libri2Vox	-0.289
	Libri2Vox+Noise	-0.009
	Libri2Vox & 3-stage (Real + SynVox2)	**0.031**
VoiceFilter	Libri2Talker	-0.739
	Libri2Talker+Noise	-0.369
	Libri2Vox	-0.309
	Libri2Vox+Noise	-0.140
	Libri2Vox & 3-stage (Real + SynVox2)	**0.064**

substantial improvement, attaining a DNSMOS gain of -0.081. This enhancement stems from Libri2Vox's incorporation of real noisy speech from VoxCeleb2. Adding noise augmentation to Libri2Vox further improves performance to 0.139, while our three-stage CL approach with SynVox2-generated synthetic speakers achieves the best results with a DNSMOS gain of 0.190. Similar trends emerge across all architectures (Conformer, SpeakerBeam, and VoiceFilter), with the combination of Libri2Vox and three-stage CL with synthetic speakers consistently achieving optimal performance.

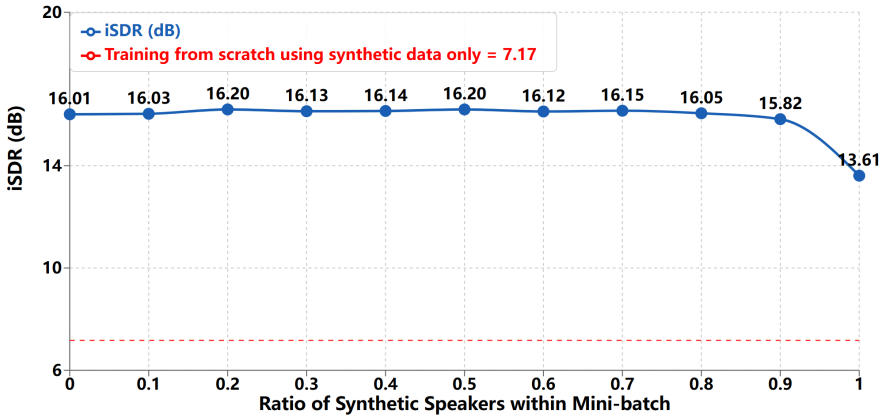


Figure 5.2: Impact of synthetic speaker ratio within one mini-batch at Stage 3 of configuration “w/ 3-stage CL (Real + SALT)" on Conformer. Red dashed line corresponds to performance (7.17 dB) where training started from scratch using synthetic data only.

5.7 Ablation study regarding synthetic data

To further investigate the specific factors contributing to the performance improvements observed in our main experiments, we conducted two ablation studies regarding synthetic speakers. These studies isolate and quantify the individual impact of key components in our approach, providing deeper insights into how different elements contribute to overall system performance.

Comparison of different synthetic data ratios within mini-batch

This experiment determines the optimal ratio of synthetic speakers within each mini-batch at Stage 3 of CL, identifying the balance that maximizes performance while avoiding degradation from excessive synthetic data.

The experiments employ Stage 3 with SALT-generated synthetic speakers, following the configuration “w/ 3-stage CL (Real + SALT)" of Conformer shown in Table 5.3. As shown in Fig. 5.2, optimal performance occurs with synthetic speaker ratios of 0.2 and 0.5, both yielding an iSDR of 16.20 dB. These configurations outperform the baseline configuration using only real data (ratio = 0.0), which achieves an iSDR of 16.01 dB. However, increasing the synthetic speaker ratio beyond these optimal values causes performance degradation. When synthetic speakers comprise 90% of the training data, the iSDR decreases to 15.82 dB. This degradation intensifies with the ratio reaching 1.0,

resulting in an iSDR of 13.61 dB.

It is worth mentioning that as shown with the red dashed line in the figure, if we start training the model from scratch using only synthetic data instead of using the aforementioned configuration, the resulting iSDR is 7.17 dB. This poor result stems from the significant mismatch between the synthetic training data and real test data. Therefore, training the model with real data first then incorporating synthetic data step by step proves necessary.

Comparison of different amount of synthetic data

In the previous experiments, the number of synthetic speakers used for TSE model training remained fixed. Given that SynVox2 and SALT generate distinct speaker sets, their combination effectively increases the total number of synthetic speakers. The same SALT method can also generate different speaker sets by changing the parameter configuration.[2] Since synthetic data can be generated infinitely, we evaluate whether increasing the number of synthetic datasets impacts performance. The amount of synthetic data per epoch remains consistent (randomly sampled from multiple synthetic datasets to match the amount of real data), but the total training process exposes the model to several times more synthetic data than before. This study determines whether increasing the number of synthetic speakers improves BLSTM performance.

Two scenarios emerge: one increases the number of speakers in Stage 3, while another introduces Stage 4 after Stage 3 to increase the total number of synthetic speakers across Stages 3 and 4. Both scenarios maintain consistent configurations for Stages 1 and 2 as specified in Table 5.3 under "w/ 2-stage CL (Real only)". Specifically, Stage 1 uses data with cosine similarity below 0.5, and Stage 2 incorporates the complete dataset.

Table 5.6 displays experimental outcomes for various synthetic speaker configurations. Experiment No. 8, for instance, utilizes SALT during Stage 3 followed by SynVox2 in Stage 4.

Simultaneous deployment of SALT-based and SynVox2-based synthetic datasets during Stage 3 (No. 4) yields performance enhancement to 13.16 dB, demonstrating advantages from expanded speaker diversity. Progressive improvements appear

[2]The SALT' version was generated using parameters k=4, p=0.5 without adding background noise, while the SALT" version was generated using parameters k=4, p=1.

Table 5.6: Comparison of different numbers of synthetic speakers. "Stage 3" and "Stage 4" columns indicate which synthetic data types were used in each stage, while "-" indicates no additional data were used in that stage. Each dataset maintains same size to ensure consistent experimental conditions.

No.	Stage3	Stage4	iSDR
1	SALT	-	13.08
2	SALT'	-	13.11
3	SALT"	-	13.24
4	SALT, SynVox2	-	13.16
5	SALT, SALT'	-	13.18
6	SALT, SALT', SALT"	-	13.31
7	SALT, SALT', SALT", SynVox2	-	**13.36**
8	SALT	SynVox2	**13.44**
9	SynVox2	SALT	13.43
10	SynVox2	SALT, SynVox2	13.30

through incorporating SALT' and SALT" into Stage 3, achieving 13.18 and 13.31 dB respectively, culminating in 13.36 dB when combining all four datasets.

Nevertheless, CL training methodologies produce the most substantial performance improvements. Experiments No. 8 and 9, which alternate synthetic datasets across Stages 3 and 4, attain superior iSDR values of 13.44 and 13.43 dB respectively. This sequential approach surpasses concurrent dataset deployment (No. 7, 13.36 dB), highlighting the effectiveness of structured synthetic data introduction.

5.8 Ablation study regarding datasets

Two ablation studies examine dataset characteristics exclusively, excluding synthetic speakers from this analysis.

5.8.1 Cross-dataset evaluation

We perform cross-dataset evaluation utilizing existing Libri2Talker and WSJ0-2mix-extr alongside our proposed Libri2Vox. TSE models undergo training on individual datasets both with and without noise augmentation, followed by evaluation on test sets from the remaining datasets. While Libri2Talker and Libri2Vox both source target speakers from

Table 5.7: Experimental results of cross-database comparisons in iSDR(dB).

Model	Libri2Talker	Libri2Vox	WSJ0-2mix	Noise aug.	Libri2Talker	Libri2Vox	WSJ0-2mix-extr
Conformer	✓				12.01	12.45	6.76
	✓			✓	12.21	12.76	7.54
		✓			-0.16	15.42	0.18
		✓		✓	-0.85	15.33	0.09
			✓		-4.37	1.67	10.71
			✓	✓	-3.38	-1.41	11.88
BLSTM	✓				9.02	8.57	6.39
	✓			✓	9.32	9.85	7.25
		✓			-0.5	12.32	0.05
		✓		✓	0.21	13.16	0.35
			✓		-1.72	1.49	7.53
			✓	✓	-4.71	0.23	8.94
SpeakerBeam	✓				9.47	8.95	7.78
	✓			✓	9.32	9.70	8.46
		✓			-1.57	11.62	-0.08
		✓		✓	-1.60	12.40	2.85
			✓		0.60	2.14	8.96
			✓	✓	-6.28	-5.55	9.28
Voicefilter	✓				8.88	8.51	5.21
	✓			✓	8.76	9.11	5.81
		✓			-3.30	11.92	0.81
		✓		✓	-2.86	11.63	1.13
			✓		1.73	2.99	8.24
			✓	✓	-0.60	1.56	8.72

LibriSpeech/LibriTTS, their interference speakers originate from distinct databases. These dataset-specific characteristics prevent effective generalization across test sets, creating domain mismatch challenges.

Table 5.7 reports the experimental findings. Predictably, models achieve peak performance when evaluated on test sets matching their training data, exhibiting substantial degradation during cross-dataset testing. Libri2Vox training yields optimal results on its corresponding test set, reaching 15.42 dB iSDR using Conformer architecture. Yet this configuration suffers dramatic performance drops when applied to Libri2Talker or WSJ0-2mix-extr test sets, producing iSDR values of -0.16 dB and 0.18 dB respectively.

Comparable patterns emerge for Libri2Talker-trained models, which demonstrate reduced effectiveness on Libri2Vox or WSJ0-2mix-extr evaluations. WSJ0-2mix-extr training produces particularly poor cross-dataset results, with BLSTM attaining merely -1.72 dB on Libri2Talker testing and 1.49 dB on Libri2Vox. These consistent

patterns reveal fundamental generalization constraints inherent to single-dataset training approaches. The findings suggest potential complementarity among datasets, motivating further investigation in subsequent analysis.

Noise augmentation demonstrates effectiveness across various training configurations. Models utilizing Libri2Talker benefit most from augmentation, whereas Libri2Vox and WSJ0-2mix exhibit inconsistent improvements. This observation indicates the inherent diversity offered by Libri2Vox through VoxCeleb2 interference speakers, potentially rendering supplementary noise augmentation redundant or detrimental.

Table 5.8: Performance comparison (iSDR in dB) in two-stage CL with different test set using multiple training datasets. TSE model was BLSTM.

Dataset used in Stage 1	Datasets used in Stage 2	Libri2Talker	Libri2Vox	WSJ0-2mix-extr
Libri2Talker		9.02	8.57	6.39
Libri2Talker	Libri2Talker, Libri2Vox, WSJ02mix	10.23	11.98	**11.16**
Libri2Vox		-0.50	12.32	0.05
Libri2Vox	Libri2Talker, Libri2Vox, WSJ02mix	**10.36**	**13.02**	10.55
WSJ0-2mix		-1.72	1.49	7.53
WSJ0-2mix	Libri2Talker, Libri2Vox, WSJ02mix	8.83	11.04	9.81

5.8.2 Two-stage CL with multiple datasets

We demonstrate that employing multiple datasets within CL effectively mitigates domain mismatch. Diverging from previous approaches, our second CL stage strategically integrates multiple datasets concurrently during training.

Table 5.8 illustrates performance metrics for various two-stage CL configurations utilizing BLSTM architecture. Initializing Stage 1 with Libri2Vox, then progressing to multi-dataset training in Stage 2, produces superior cross-dataset generalization: 10.36 dB on Libri2Talker, 13.02 dB on Libri2Vox, and 10.55 dB on WSJ0-2mix-extr. This configuration substantially surpasses alternatives where Libri2Talker or WSJ0-2mix-extr serve as Stage 1 foundations.

Remarkably, Libri2Vox establishes a more robust CL foundation than Libri2Talker despite containing fewer data samples. This effectiveness stems from Libri2Vox's authentic acoustic diversity, which creates stronger initial representations amenable to subsequent refinement through supplementary data sources.

Additional performance gains likely await through incorporating synthetic variants of Libri2Talker and WSJ0-2mix-extr datasets within Stage 3 CL, paralleling Table 5.3

results. Such extensions, while promising, exceed our current investigation scope and represent future research directions.

5.9 Conclusion

This chapter advances our solution to RQ3: Can speech generative models produce valuable training data for TSE? Our findings conclusively affirm that speech generative models successfully create synthetic data that substantially enhances TSE performance through strategic training integration. We introduce Libri2Vox, an innovative dataset addressing TSE challenges within authentic acoustic environments. This dataset merges clean LibriTTS speech with naturally noisy VoxCeleb2 interference, establishing a diverse training corpus exceeding 7,000 speakers. Our approach connects idealized synthetic mixtures with uncontrolled recordings through VoxCeleb2's authentic acoustic variations, channel characteristics, and environmental conditions.

Demonstrating synthetic data generation's effectiveness, we enhance Libri2Vox using two complementary speech generative architectures: SynVox2 implementing orthogonal Householder neural networks and SALT utilizing k-NN-based VC. Experimental results show these synthetic speakers, integrated via three-stage CL, produce consistent performance gains across TSE architectures. Models achieve 0.78 dB iSDR improvements over baselines, with optimal configurations employing 0.5 synthetic-to-real speaker ratios during training. Crucially, real-world recording evaluations validate that Libri2Vox-trained models incorporating synthetic augmentation attain positive quality improvements, whereas conventional artificial mixture training yields significant under-performance. This confirms that combining authentic acoustic variations with synthetic speaker diversity directly enhances real-world application performance.

These results are especially encouraging since synthetic data generation enables unlimited speaker diversity expansion surpassing real-world recording constraints. Unlike physical recordings limited by speaker availability and recording conditions, synthetic speakers permit virtually boundless generation, facilitating continuous training diversity expansion without practical collection limitations. Nevertheless, synthetic speaker data presents inherent constraints within speech generative models. The variations of the generated data often remain limited, and identifying optimal

training strategies poses ongoing challenges. Frequently, minimal generated data portions prove genuinely valuable for necessary distribution coverage. Extended generation may yield diminishing results through redundant data providing negligible model insights [156]. This indicates that the “new knowledge" introduced by speech generative models can rapidly become repetitive, leading to redundant training that provides limited benefit to the model.

Future research will investigate optimal synthetic data characteristics for TSE through data selection methodologies [157, 158]. Network gradient tracking [159, 160] across training steps could identify valuable training data. Moreover, effective dataset distillation techniques generating representative synthetic data may illuminate essential TSE task representations.

6

Training Dynamics-Aware Multi-Factor Curriculum Learning for Target Speaker Extraction

6.1 Introduction

The preceding chapters have systematically addressed multiple fundamental challenges in TSE through novel data and training methodologies. Chapter 3 introduced CL that progressively trains models from simple to complex examples, achieving substantial performance improvements. Chapter 4 demonstrated how incorporating synthetic speakers generated through voice conversion further enhanced model capabilities. Chapter 5 presented Libri2Vox, bridging controlled experiments with real-world acoustic complexity. However, these approaches primarily addressed individual complexity factors in isolation from others. Meanwhile, real-world TSE scenarios present multiple interacting challenges that cannot be effectively captured through

single-factor optimization.

This chapter resolves these constraints through a comprehensive multi-factor CL that simultaneously coordinates multiple complexity dimensions. While Chapter 3 demonstrated the effectiveness of single-factor CL using individual complexity factors such as speaker similarity, these approaches treated complexity dimensions in isolation, failing to address compound challenges where multiple factors interact simultaneously —such as low SNR conditions combined with high speaker similarity and multiple interfering speakers.

Initial attempts at multi-factor CL relied on predefined difficulty measures that combined multiple factors through handcrafted rules. However, these predetermined approaches still suffered from a fundamental limitation: they assumed theoretical difficulty relationships rather than reflecting how models actually learn from different examples. What appears "difficulty" according to predefined metrics may not align with actual model learning behaviors during training.

The core innovation of this chapter centers on curriculum design grounded in empirical training dynamics rather than a priori difficulty specifications. We present TSE-Datamap, which categorizes training examples according to observed model learning behaviors. Through confidence and variability metric tracking throughout training epochs, we construct a 2-dimensional representation that delineates three characteristic data categories (easy-to-learn, ambiguous, and hard-to-learn), each providing distinct contributions to model learning and generalization capacity.

This training dynamics-aware approach ensures curriculum design follows actual model learning behaviors rather than theoretical assumptions about factor difficulty. Building upon our established CL framework from Chapter 3 and synthetic speaker generation methodologies from Chapter 4, this approach completes our comprehensive robust TSE framework through synergistic integration of progressive training strategies with realistic acoustic conditions.

This chapter provides two primary contributions: 1) We develop a multi-factor CL strategy for TSE that coordinates multiple complexity factors simultaneously, facilitating progressive learning from simple to complex scenarios; 2) We present TSE-Datamap, establishing curriculum design foundations through observed training dynamics rather than predetermined difficulty metrics. Experimental results demonstrate our approach consistently surpasses both single-factor curricula and predefined multi-

factor approaches, achieving significant performance improvements in challenging multi-speaker scenarios.

6.2 Multi-Factor Complexity in TSE

This section introduces the four primary factors that can simultaneously influence the TSE task difficulty. The investigation covered in this chapter explicitly examines those factors and analyzes their effects on TSE performance.

SNR. The SNR becomes more complex in multi-speaker mixtures case:

$$\mathrm{SNR} = 10 \log_{10} \frac{\|s_0(n)\|^2}{\left\| \sum_{i=1}^{K} s_i(n) + u(n) \right\|^2} \tag{6.1}$$

When the SNR approaches or falls below 0 dB, the combined interference power surpasses the target speech energy. The adverse effect of low SNR is further amplified when accompanied by multiple interfering speakers or high speaker similarity, interactions that cannot be captured by single-factor curricula.

Number of Interfering Speakers. Traditional TSE frameworks typically handle single interference scenarios [161, 144], requiring models to differentiate between two speakers within the acoustic mixture. When confronted with multiple interfering speakers, models must extract targets from intricate spectral interactions generated by overlapping speech signals. Acoustic complexity increases non-linearly with the quantity of simultaneous speakers K in Eq. 6.1, as each additional speaker contributes supplementary spectral interference and masking phenomena across diverse frequency ranges [162, 163].

Temporal Overlap Ratio. This metric quantifies the temporal proportion during which target and interfering speakers generate concurrent speech, producing variable effects on fundamental acoustic and speaker characteristics. Minimal overlap conditions furnish substantial clean target segments that function as distinct acoustic references. Conversely, extensive overlap ratios may convert the extraction task into simultaneous source separation without frequent speaker alternations, a scenario that contemporary TSE architectures can manage with reasonable stability. Intermediate overlap ratios constitute the most difficult conditions, as frequent transitions between target and

interfering speakers require persistent speaker monitoring and effective management of continuous acoustic variations [164, 163]. This phenomenon has been documented in modern TSE implementations, as demonstrated in previous research [165, 166]. Within this study, the ratio spans from 0 to 1, where 0 denotes complete overlap and 1 indicates absence of speaker overlap in the mixture.

Synthetic and Real Interfering Speakers. The incorporation of synthetic interference speakers introduces additional variations that may not be manifested in real recordings. Synthetic samples can inhabit atypical regions within the acoustic feature space, generating extraction challenges distinct from those presented by authentic speech interference. While natural interfering speakers contribute genuine acoustic characteristics, synthetic data enables models to address boundary cases and mitigates overfitting to particular speaker pattern [146, 167].

6.2.1 TSE-Datamap: A Training Dynamics Framework

Previous CL approaches in Chapters 3, 4, and 5 relied on predefined difficulty measures such as speaker similarity and SNR. However, these metrics may not accurately reflect actual model learning behavior. Inspired by dataset cartography [168], we propose TSE-Datamap, which grounds curriculum design in observed training dynamics. Unlike traditional CL methodologies that utilize predefined heuristics and disregard model feedback during training progression, our approach positions training examples based on empirical learning patterns observed over time. Through systematic model behavior monitoring, TSE-Datamap identifies three distinct regions (easy, ambiguous, hard) that provide direct guidance for our sampling methodology.

Training Dynamics Metrics

Let us denote the training loss as our evaluation metric, designated as M_i, for each training example i. In order to to examine comprehensive training dynamics, we subsequently compute two essential statistics across E training epochs, which is indexed via e, and the metric value for example i during epoch e is then $M_i^{(e)}$:

- The **confidence** metric quantifies the average performance across epochs:

$$\mu_i = \frac{1}{E} \sum_{e=1}^{E} M_i^{(e)} \tag{6.2}$$

- The **variability** metric captures prediction consistency through standard deviation:

$$\sigma_i = \sqrt{\frac{1}{E} \sum_{e=1}^{E} (M_i^{(e)} - \mu_i)^2} \tag{6.3}$$

We implement an SNR-based loss function that directly optimizes signal quality. For each epoch, the metric is computed as:

$$M_i(s_{0,i}, \hat{s}_{0,i}) = -10 \log_{10} \frac{\|s_{0,i}\|^2}{\|s_{0,i} - \hat{s}_{0,i}\|^2} \tag{6.4}$$

where $s_{0,i}$ and $\hat{s}_{0,i}$ denote the ground truth target and predicted speech signals respectively for training example i.

Datamap Construction and Region Identification

The confidence and variability metrics establish a 2-dimensional **datamap** where each training example occupies a distinct position determined by its training dynamics, revealing three characteristic learning patterns:

- **Easy-to-learn**: These exhibit high confidence and low variability, typically encompassing clear, high-SNR speech samples with minimal interference.
- **Ambiguous**: Distinguished by high variability, these demonstrate unstable predictions as models fluctuate between alternative hypotheses. These samples contain rich discriminative content, compelling models to develop robust decision boundaries and enhancing generalization capabilities.
- **Hard-to-learn**: These present low confidence and low variability, indicating consistently imprecise predictions. These frequently arise from extremely challenging conditions such as very low SNR environments where discriminative information is severely limited.

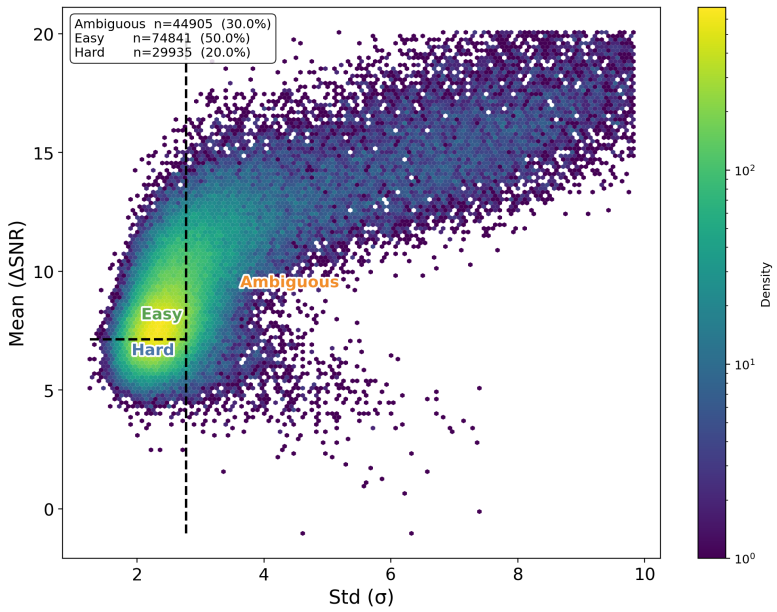


Figure 6.1: TSE-Datamap computed. Each point denotes a sample, with x-axis the standard deviation of ΔSNR and y-axis the mean ΔSNR. Number of training epochs in this study is 50.

To construct the TSE-Datamap visualization shown in Figure 6.1, we track the training dynamics of each example across $E = 50$ epochs. Based on the resulting distribution, we categorize examples into three regions: we first select 30% of samples with highest variability as Ambiguous, then from the remaining 70%, we classify the top 50% with highest confidence as Easy-to-learn and the bottom 20% as Hard-to-learn.

6.3 Experimental Evaluation

6.3.1 Dataset and Training Configuration

Dataset Composition

Following the experimental framework established in Chapter 5, we utilize Libri2Vox as our primary dataset [169]. As detailed in Chapter 5, this dataset addresses the fundamental limitations of existing TSE datasets by incorporating realistic acoustic conditions while maintaining sufficient speaker diversity for robust model training.

The dataset partitioning follows the specifications outlined in Chapter 5: the

training partition encompasses 1,151 LibriTTS speakers and 5900 VoxCeleb2 speakers across 149681 utterances, totaling 250 hours of mixture data. The validation partition contains 40 LibriTTS speakers and 94 VoxCeleb2 speakers with 4664 utterances spanning 7.77 hours, while the test partition includes 39 LibriTTS speakers and 118 VoxCeleb2 speakers across 3931 utterances corresponding to 6.55 hours. The original mixing protocol samples SNR uniformly from the range $[-5, 5]$ dB, establishing baseline acoustic conditions that reflect realistic deployment scenarios.

Building upon the synthetic data augmentation techniques introduced in Chapter 4, we additionally leverage the synthetic variant *Libri2Vox-syn*, generated using advanced speech synthesis models including SALT [122] and SynVox2 [150]. This synthetic augmentation enables exploration of multi-factor CL with controlled speaker characteristics while expanding the acoustic diversity beyond naturally occurring recordings.

For comprehensive evaluation across varying complexity levels, we construct specialized test sets containing 1, 2, and 3 interfering speakers through systematic mixing of LibriTTS and VoxCeleb2 test partitions. Single interfering speaker mixtures maintain SNR sampling from $[-5, 5]$ dB, preserving consistency with the original dataset specifications [169]. However, for multi-speaker scenarios involving 2 or 3 interfering speakers, we adjust the SNR range to $[0, 10]$ dB to prevent energy overflow and clipping artifacts that commonly arise from multiple concurrent speakers. All evaluation mixtures employ complete temporal overlap (ratio = 0.0) and exclusively utilize real audio without synthetic interference, ensuring consistent evaluation conditions across complexity levels.

TSE Model Architecture

Our experimental framework employs a model that consists of 2 BLSTM layers with 256-dimensional hidden states, succeeded by one 512 fully connected layers for feature integration and mask generation. The acoustic features are 80-dimensional log-mel filterbanks extracted from input mixtures. The speaker-specific information is provided through speaker embeddings from pre-trained, frozen ECAPA-TDNN network [170]. The whole TSE model generates time-frequency masks that selectively isolate target speech components from complex acoustic mixtures. Comprehensive architectural

details and implementation specifics are documented in [169].

As demonstrated in Chapter 3, CL exhibit strong transferability across diverse TSE architectures [171]. Such independence suggests that our multi-factor CL approach can be effectively applied to alternative separation networks, without requiring substantial modifications to the underlying training methodology.

Training Configuration

We utilize an Adam [108] optimizer, initialized with a learning rate of 2e-4, implemented within a warm-restart scheduling framework that balances training stability with convergence efficiency. The learning rate schedule incorporates a warmup phase spanning 5,000 optimization steps, during which the learning rate increases exponentially to 1e-3, followed by Noam-style decay [172] proportional to the inverse square root of step count, ultimately decreasing to a minimum threshold of 1e-5. This scheduling approach ensures stable optimization during early training phases while maintaining adaptive learning capacity throughout extended training periods.

The training also includes early stopping, where we monitor validation loss progression with a patience threshold of 6 epochs per stage, effectively preventing overfitting while ensuring comprehensive convergence within each training phase. This staged early stopping approach accommodates the multi-phase nature of CL, allowing each curriculum stage to achieve optimal performance before transitioning to increased complexity levels.

Performance evaluation employs SDR as the primary metric, supplemented by improvement iSDR measurements relative to input mixture quality. These metrics provide comprehensive assessment of extraction quality and have been widely adopted in previous TSE research [161, 144].

6.4 Results and Analysis

6.4.1 Analysis of Single-Factor vs. Multi-Factor

To examine the individual contributions of different complexity dimensions and demonstrate the efficacy of our multi-factor methodology, we perform comprehensive comparisons between single-factor curricula and our proposed multi-dimensional

Table 6.1: Impact of single-factor and multi-factor CL. Each row represents a different curriculum configuration with varying factors. The curriculum stages column shows the progression of training difficulty across three stages, with parameters in the format [*num_speakers*, *snr_low*,*snr_high*, *overlap_ratio*, *inter_source*]. *num_speakers*: number of interfering speakers, [*snr_low*, *snr_high*]: SNR range in dB, *overlap_ratio*: fraction of temporal overlap, ranged from 0.0 to 0.4, *inter_source*: whether interference comes from real or synthetic (syn) speakers.

Exp.	**Curriculum Stages**	**iSDR 2spk**	**iSDR 3spk**	**iSDR 4spk**
baseline	random sampling	12.38	8.56	7.16
baseline-CL	[[1,0,10,0,real], [2,0,10,0,real], [3,0,10,0,real]]	12.58	8.78	7.40
single factor: #snr	[[1,5,10,0,real], [2,0,10,0,real], [3,0,5,0,real]]	13.04	9.71	8.62
single factor: #overlap ratio	[[1,0,10,0,real], [2,0,10,0.2,real], [3,0,10,0.4,real]]	12.62	9.87	8.76
single factor: #inter_source_syn	[[1,0,10,0,syn], [2,0,10,0,syn], [3,0,10,0,syn]]	11.22	8.43	9.19
single factor: #inter_source_real/syn	[[1,0,10,0,real/syn], [2,0,10,0,real/syn], [3,0,10,0,real/syn]]	12.73	9.51	8.96
single factor: #speaker count	[[1,0,10,0,real], [1,0,10,0,real], [1,0,10,0,real]]	12.80	7.79	6.73
multiple factors	[[1,5,10,0,real/syn], [2,0,10,0.2,real/syn], [3,0,5,0.4,real/syn]]	**13.22**	**10.08**	**9.21**

strategy. All curriculum designs adopt three stages. In the single-factor setting, only one factor is varied across stages while all remaining stage parameters are kept identical to the baseline-CL configuration. In contrast, the multi-factor curriculum jointly varies multiple factors across all three stages, i.e., the stage definitions change simultaneously along several difficulty axes. Our baseline configuration implements uniform random sampling without CL, where training samples are generated with SNR uniformly distributed across $[0, 10]$ dB, interfering speaker count selected from $\{1, 2, 3\}$, overlap ratio chosen from $\{0, 0.2, 0.4\}$, and interference type sampled from $\{\text{real}, \text{syn}, \text{real/syn}\}$. This baseline approach produces iSDR values of 12.38 dB, 8.56 dB, and 7.16 dB for 2-speaker, 3-speaker, and 4-speaker configurations respectively. The baseline-CL variant implements a fundamental three-stage CL with static parameters, systematically escalating speaker count from 1 to 3 across training stages while preserving other factors, yielding modest enhancements to 12.58 dB, 8.78 dB, and 7.40 dB respectively. Utilizing baseline-CL as our comparative benchmark, we systematically assess single-factor curricula through ablation of individual complexity factors.

The results are presented in Table 6.1. Among single-factor curriculum approaches, the SNR-focused curriculum demonstrates significant gains with 13.04 dB, 9.71 dB, and 8.62 dB through systematic SNR threshold adjustment across training stages. The single-factor speaker count experiment constrains all stages to utilize solely single interfering speakers, contrasting with standard CL that systematically advances from 1 to 3 speakers per stage, producing performance levels of 12.80 dB, 7.79 dB, and 6.73 dB. Exclusive training with single interference scenarios yields superior performance on 2-speaker mixtures relative to baseline-CL, yet exhibits substantial deterioration on multi-speaker scenarios, demonstrating that incorporating multi-speaker training introduces performance trade-offs that marginally compromise 2-speaker performance while enhancing complex scenario handling.

Our multi-factor strategy attains optimal overall performance with 13.22 dB, 10.08 dB, and 9.21 dB respectively. This represents considerable improvements of up to approximately 24.5% over baseline in 4-speaker scenarios, with performance gains escalating as interfering speaker quantity increases, confirming that coordinated multi-dimensional progression facilitates more effective learning than single-factor optimization approaches.

Table 6.2: Performance comparison of different stage orderings in CL. Each stage takes one third of total epochs. Baseline uses random sampling without curriculum.

Order	iSDR 2spk	iSDR 3spk	iSDR 4spk
baseline	12.38	8.56	7.16
multiple factors	**13.22**	**10.08**	9.21
E / A / H	13.15	9.85	**9.32**
E / H / A	12.93	9.63	9.18
A / E / H	12.82	9.61	9.22
A / H / E	12.96	9.72	9.17
H / E / A	12.77	9.54	**9.32**
H / A / E	12.90	9.63	9.10
E / A / H (forgetting)	8.83	5.43	5.52

6.4.2 TSE-Datamap Analysis

To establish the optimal sequencing of difficulty regions for CL, we conduct comprehensive experiments exploring all possible permutations of easy (E), ambiguous (A), and hard (H) examples. Each permutation constitutes a distinct strategy where training advances through three stages with equivalent duration. The categorization methodology for each region follows the procedures outlined in Section 2.4.

Experimental outcomes from various TSE-Datamap configurations are presented in Table 6.2, incorporating both baseline and multi-factor systems that achieved optimal performance in Table 6.1. The E / A / H scheduling demonstrates superior performance among all tested sequences, producing absolute improvements of 0.77, 1.29, and 2.16 dB over baseline, while surpassing the handcrafted multi-factor approach by 0.11 dB. Performance enhancements increase with additional interferers, demonstrating the efficacy of the proposed method against increasing complexity.

Within easy-first methodologies, E / A / H surpasses E / H / A, indicating that ambiguous data preceding hard examples provides superior decision boundary calibration. Conversely, sequences positioning hard examples earlier (A / E / H, A / H / E) or initiating with them (H / E / A, H / A / E) typically under-perform due to unstable early optimization dynamics. Overall, E / A / H delivers the most effective progression, particularly in challenging multi-speaker configurations.

We additionally implement an E / A / H (forgetting) experiment, where stage-specific data matches E / A / H, but each stage exclusively utilizes its designated region data

Table 6.3: iSDR (dB) under fixed data quantity drawn from Datamap regions. All settings use 70% of the training set for equal-quantity comparison.

Setting	2spk	3spk	4spk
ambi 70%	**11.67**	**8.67**	**8.61**
baseline 70%	11.37	8.19	7.17
hard 70%	10.10	6.75	6.99
easy 70%	9.71	6.50	6.40

without preserving previous stage knowledge. This configuration exhibits substantial performance deterioration, validating that catastrophic forgetting emerges when previously acquired knowledge is not preserved throughout curriculum progression.

6.4.3 Fixed-Quantity Ablation of Datamap Regions

To distinguish the learning contributions of distinct TSE-Datamap regions from data volume effects, we implement controlled ablation studies under fixed data quantity constraints. All methodologies are restricted to utilize precisely 70% of the complete training dataset. The baseline70% configuration samples uniformly across the entire training distribution, while easy70%, ambi70%, and hard70% designate 70% of their data toward their corresponding target regions. For all configurations, the remaining 30% is distributed equally among the alternative two regions.

Table 6.3 demonstrates that under equivalent data quantity restrictions, ambi70% consistently surpasses the baseline70% across all mixture complexity levels. The easy70% configuration exhibits the poorest performance among the four methodologies, potentially attributable to these examples being rapidly processed, yielding diminishing informative gradients throughout training progression. Conversely, ambiguous examples sustain consistent informativeness across training due to their intrinsic learning complexity, forcing the model toward more robust decision boundaries that improve generalization under equivalent data quantities.

6.5 Conclusion

This chapter addressed RQ4: How can we develop training strategies that handle the complex interactions between multiple difficulty factors in challenging multi-speaker scenarios? We have presented a training dynamics-aware multi-factor CL framework that moves beyond the single-factor approaches explored in earlier chapters to coordinate multiple complexity dimensions simultaneously.

The core of the proposed methodology is TSE-Datamap, which guides curriculum design in empirical training dynamics rather than predetermined assumptions. This approach categorizes training examples based on observed model learning behaviors into easy-to-learn, ambiguous, and hard-to-learn regions, ensuring curriculum sequencing aligns with actual model capabilities. The multi-factor framework simultaneously handles the interactions between signal-to-noise ratio, speaker count, temporal overlap, and synthetic/real speaker proportions, achieving substantial improvements particularly in challenging multi-speaker scenarios.

The success of TSE-Datamap suggests broader applicability to other speech processing tasks facing similar challenges of heterogeneous difficulty and limited training resources. Future work will explore theoretical foundations of training dynamics and develop automated curriculum design methods that further reduce human intervention in training pipeline optimization.

7 Conclusion and Future Work

7.1 Summary of Research Contributions

This thesis has investigated into improving modern DNN-based TSE systems through improved training scheme, strategic training data selection and synthetic speaker generation. It has addressed fundamental limitations in current TSE approaches by developing novel training strategies that enhance model robustness and generalization capabilities. Our research contributions span four key research questions that systematically build upon each other to create a unified framework for robust TSE.

7.1.1 Addressing the Research Questions

RQ1: How can we identify and select useful training data for more effective TSE training?

Through extensive empirical analysis in Chapter 3, we have demonstrated that not all training examples contribute equally to model learning. We have identified

and evaluated multiple difficulty measures for TSE training data, including speaker-dependent factors (gender differences and speaker similarity based on cosine distance) and mixture-dependent factors (SDR and SNR). Our findings reveal that speaker similarity, measured through cosine distance between speaker embeddings, provides the most reliable indicator of extraction difficulty for TSE models. This comprehensive analysis enables the development of data selection and learning strategies that prioritize training examples based on their learning value rather than treating all samples uniformly.

RQ2: How do we use the selected data at different training stages to improve TSE performance?

Building upon the difficulty measures identified in RQ1, we have developed and validated a multi-stage CL framework that strategically schedules training progression from easier to more challenging examples. Our experimental results in Chapter 3 show that this structured approach significantly outperforms conventional random sampling, achieving improvements of up to 0.97 dB in iSDR using self-paced CL. The key insight is that models benefit from first establishing robust extraction capabilities on simpler cases (e.g., different-gender speakers with low similarity) before tackling more challenging scenarios (e.g., same-gender speakers with high similarity). This progressive CL training strategy is effective across multiple modern TSE models.

RQ3: Can we use speech generative models to generate useful data for training TSE?

Chapters 4 and 5 have demonstrated that modern speech generative models can create valuable synthetic training data for TSE. We have explored two complementary approaches: VC by SALT and SA via SynVox2. These techniques enable the generation of diverse synthetic interference speakers that expanded training data variety without requiring additional recordings. When integrated within our CL framework as a third training stage, synthetic speakers yield performance improvement consistently across different TSE models. The development of Libri2Vox further validates this finding by combining clean target speech with noisy interference, bridging between controlled experimental conditions and real-world complexities.

RQ4: How can we develop training strategies that handle the complex interactions between multiple difficulty factors in challenging multi-speaker scenarios?

Chapter 6 has addressed the complex interactions between multiple difficulty factors by developing a comprehensive multi-factor CL framework. Building upon the CL foundations from Chapter 3 and synthetic speaker generation from Chapter 4, we have coordinated multiple complexity dimensions including SNR, number of interfering speakers, temporal overlap ratios, and synthetic versus real speaker characteristics. The key innovation is TSE-Datamap, which grounds curriculum design in empirical training dynamics rather than predetermined assumptions, categorizing training examples based on observed model learning behaviors. This comprehensive framework has demonstrated that coordinated multi-dimensional progression facilitates more effective learning than single-factor approaches, providing a unified training strategy for real-world multi-speaker environments.

7.1.2 Chapter-wise Contributions

Chapter 2 has established the technical foundation by providing a comprehensive overview of TSE technologies, analyzing the evolution from traditional speech enhancement to modern neural network-based approaches. This chapter has identified key limitations in current TSE systems, particularly the constraints of limited speaker diversity and the gap between artificial training conditions and real-world deployment scenarios.

Chapter 3 has introduced our CL framework, demonstrating how CL with carefully selected data improves TSE performance. The systematic evaluation of different difficulty measures and training schedulers provides practical insights for implementing CL in TSE systems. The self-paced learning approach, which dynamically adjusts to model capabilities during training, achieved the best results while maintaining implementation flexibility.

Chapter 4 has extended the CL framework by incorporating synthetic speakers generated through VC. This chapter validates the enhancing impact of synthetic data on model generalization. The optimal balance between real and synthetic data, as well as the importance of VC hyperparameters, are also important aspects that guide practical implementation.

Chapter 5 has presented Libri2Vox, a novel dataset that addresses the acoustic realism gap in TSE training. By combining clean LibriTTS recordings with naturally

noisy VoxCeleb2 speech, we have created a training resource that better represents real-world conditions while maintaining the supervision necessary for TSE. The integration of synthetic speakers with this realistic dataset demonstrates complementary benefits, particularly for deployment in authentic acoustic environments where real-world data variability and noise conditions can be directly leveraged to improve model robustness and generalization performance.

Chapter 6 has presented a training dynamics-aware multi-factor CL framework that coordinates multiple complexity dimensions simultaneously in multi-speaker scenarios. We have introduced TSE-Datamap, which categorizes training examples based on empirical training dynamics rather than predetermined assumptions. This unified framework integrates the CL, synthetic speaker generation, and realistic acoustic conditions from previous chapters into a comprehensive training strategy that handles the complex interactions between multiple difficulty factors in real-world environments.

7.2 Limitations and Future Work

While our research has made contributions to advancing TSE systems, several limitations and opportunities for future investigation remain.

7.2.1 Enhanced Difficulty Measures

Our current difficulty measures rely mainly on speaker similarity and signal quality metrics. However, these measures may not fully capture what makes a TSE task truly challenging. Future work could develop more comprehensive difficulty measures in several ways. First, difficulty measures that consider perceptual factors may be concerned, such as how similar two voices sound to human listeners, rather than just similarity. Second, linguistic factors like speaking rate, accent strength, or vocabulary complexity can influence extraction difficulty but are not currently considered. Third, instead of manually selecting and combining difficulty measures, one can train neural networks to automatically learn which combinations of factors best predict extraction difficulty. This learning could happen dynamically during training, allowing the curriculum to adapt as the model improves. Finally, current measures treat each

utterance as a single difficulty score, but difficulty may vary throughout an utterance — for example, when speakers talk simultaneously versus taking turns. Developing time-varying difficulty measures could enable more fine-grained curriculum design.

7.2.2 Advanced Synthetic Speaker Generation

Our synthetic speaker generation methods using VC showed promising results, but they have limitations. The current approach can only create variations based on existing speakers in our reference pool, which restricts the diversity of generated voices. Future work could explore several improvements. First, newer generation methods like diffusion-based ones [173, 174] can create more diverse synthetic speakers while keeping them natural-sounding. These models have shown success in other domains [175, 176, 177] and could offer better control over specific voice characteristics like pitch, speaking style, or accent. Second, we could design synthetic speakers specifically to challenge known weaknesses in TSE models. For example, if a model struggles with high-pitched voices, we could generate more training examples in that range. Third, adversarial training could help create the hardest possible interference speakers that still sound realistic. This would push models to learn more robust features. Finally, in many real applications, both audio and video data are available [49, 48, 53]. Future synthetic speaker generation can thus include visual information like lip movements or spatial audio cues, creating more realistic training scenarios for multi-modal TSE systems.

7.2.3 Real-world Deployment Considerations

While our models with proposed methodologies have performed well on real-world interference and noise, real-world deployment presents additional challenges that need further investigation. First, we need more diverse evaluation data. Current benchmarks mainly use English speech in controlled conditions, but real applications must handle multiple languages, heavy accents, and challenging environments like busy restaurants or outdoor settings. Testing on such diverse data would reveal where current models fail. Second, real-world acoustic conditions change constantly. For example, a model trained on office recordings may perform poorly in a moving car. Therefore, future systems may need rapid adaptation, being able to quickly adjust to

new environments using just a few minutes of local audio data, without suffering from catastrophic forgetting. Third, current TSE models are still computationally large and slow for devices such as hearing aids or smartphones. Therefore, more compact and efficient models that maintain good performance are needed. Approaching this need could involve designing new architectures or compressing existing models through techniques like quantization or pruning.

7.2.4 Theoretical Understanding

Our experiments show that CL and synthetic data clearly improve TSE performance, but we have not gained a full understanding on why these methods work so well. This theoretical gap limits our ability to design even better training strategies. Future research could address several key questions. First, why does training on easy examples first benefits the learning process? One hypothesis is that CL creates a smoother learning path, helping models avoid getting stuck in poor solutions early in training. Mathematical analysis could test this idea and potentially reveal better ways to order training examples. Second, how much synthetic data is actually useful? Our experiments suggest that adding more synthetic speakers eventually stops helping, but we don't know why this limit exists. Understanding this relationship could help us generate only the most valuable synthetic data, saving computational resources. Third, what makes some synthetic speakers more helpful than others for training? If we understood which properties of synthetic data lead to better generalization, we could design generation methods that specifically create these beneficial properties. Developing this theoretical foundation would move TSE training from trial-and-error approaches toward more principled, predictable methods.

7.2.5 Extension to Complex Scenarios

Current TSE systems assume relatively simple scenarios, but real-world applications often involve more complex situations. Several important extensions deserve future attention. First, real conversations are dynamic: people join and leave discussions, move around, and interrupt each other. This dynamicity poses a fundamental challenge for current models, which process fixed-length segments and cannot adapt when the number of speakers changes mid-conversation. Developing TSE systems that

can handle these scenarios would make them much more practical. Second, TSE is often just the first step in a larger system. For example, in meeting transcription, we need to first extract the voice from each speaker, then convert it to text. Current approaches treat these as separate problems, but end-to-end systems that jointly optimize extraction and recognition could perform better. Third, multilingual scenarios present special challenges. When speakers use different languages, their speech patterns differ significantly, making extraction harder. Fourth, as TSE systems become more widely deployed, security concerns grow. Attackers may try to fool these systems with specially crafted audio. Understanding and defending against such attacks is necessary for sensitive applications. Finally, real environments contain not just competing speakers but also music, traffic noise, and other sounds. While Libri2Vox (Chapter 5) incorporates naturally occurring background sounds from VoxCeleb2, future work may further extend TSE systems to explicitly model and handle these diverse interference types as distinct sources rather than treating them as undifferentiated background noise.

7.2.6 Emotional speech and other expressive speaking styles.

A limitation of this study is that we do not explicitly evaluate TSE under emotional or highly expressive speaking styles. Both the speaker encoder used for enrollment conditioning and the TSE models studied in this thesis are trained primarily on neutral speech, and our training data do not intentionally include emotional speech conditions. As a result, when the reference utterance or the mixture contains strong emotion (or other style shifts such as shouting, crying, laughter, or dramatic prosody), the learned speaker representations and the separation model may face a style-induced mismatch, potentially degrading extraction performance. These emotional-speech scenarios are therefore considered outside the scope of the present work. As future work, we plan to benchmark on emotional corpora and investigate style-robust speaker encoders and TSE training strategies, including curriculum designs that explicitly incorporate speaking-style variation.

Bibliography

[1] Yang Liu, Yueyue Fan, Marc Delcroix, Keisuke Kinoshita, and Tomohiro Nakatani. Target speaker extraction with multi-scale conformer and streaming local attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[2] Steven F Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.

[3] Jian Chen, Jacob Benesty, and Yiteng Huang. A new wiener filtering method for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1218–1234, 2006.

[4] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[5] Yariv Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1526–1555, 1992.

[6] Y. Xu, J. Du, L. Dai, and C. H. Lee. Regression-based speech enhancement with noise-specific deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1740–1744. IEEE, 2014.

[7] D. P. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):483–492, 2016.

[8] Hakan Erdogan, John R. Hershey, Shinji Watanabe, and Jonathan Le Roux. Phase-sensitive and recognition-boosted speech enhancement. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 708–712, 2015.

[9] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.

[10] Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3229–3233, 2018.

[11] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2472–2476, 2020.

[12] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 46–50, 2020.

[13] Haoyu Li, Szu-Wei Fu, Yu Tsao, and Junichi Yamagishi. iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2662–2666, 2020.

[14] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech enhancement generative adversarial network. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 2017.

[15] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. MetricGAN: Generative adversarial networks based Black-Box metric scores optimization for speech

enhancement. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2031–2041, 2019.

[16] Yihui Fu, Yun Liu, Jingdong Li, Dawei Luo, Shubo Lv, Yukai Jv, and Lei Xie. Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation. *arXiv preprint arXiv:2111.06015*, 2021.

[17] Yi Li, Yang Sun, and Syed Mohsen Naqvi. U-shaped transformer with frequency-band aware attention for speech enhancement. *arXiv preprint arXiv:2112.06052*, 2021.

[18] Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2928–2932, 2022.

[19] M Berouti, R Schwartz, and J Makhoul. Enhancement of speech corrupted by acoustic noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 208–211. IEEE, 1979.

[20] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[21] DeLiang Wang. Ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, pages 181–197. Springer, 2005.

[22] Yuxuan Wang, A. Narayanan, and Deliang Wang. On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22:1849–1858, 2014.

[23] Se Rim Park and Jinwon Lee. Fully convolutional neural networks for speech enhancement. In *Interspeech 2016*, pages 1993–1997, 2016.

[24] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5059–5063. IEEE, 2015.

[25] K. Tan and D. Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech 2018*, pages 3229–3233, 2018.

[26] Daniel Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[27] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.

[28] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[29] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. Single channel target speaker extraction and recognition with speaker beam. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5554–5558. IEEE, 2018.

[30] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey. Alternative objective functions for deep clustering. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 686–690. IEEE, 2018.

[31] Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021.

[32] Zhuo Chen, Yi Luo, and Nima Mesgarani. Deep attractor network for single-microphone speaker separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250. IEEE, 2017.

[33] E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[34] Adelbert W Bronkhorst. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, 77(5):1465–1487, 2015.

[35] James L Flanagan, James D Johnston, Robert Zahn, and Gary W Elko. Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78(5):1508–1518, 1985.

[36] Sharon Gannot, Emmanuel Vincent, Sharon Markovich-Golan, and Alexey Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017.

[37] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf. *APSIPA Transactions on Signal and Information Processing*, 8:e12, 2019.

[38] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Blind source separation exploiting higher-order frequency dependencies. *IEEE transactions on audio, speech, and language processing*, 15(1):70–79, 2006.

[39] Nobutaka Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 189–192. IEEE, 2011.

[40] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134, 2014.

[41] Jakub Jansky, Jan Málek, Jaroslav Cmejla, Tomas Kounovsky, Zbynek Koldovsky, and Jiri Zdansky. Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 676–680, 2020.

[42] John Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001.

[43] John R Hershey, Steven J Rennie, Peder A Olsen, and Trausti T Kristjansson. Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech and Language*, 24(1):45–66, 2010.

[44] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[45] Alexey Ozerov and Emmanuel Vincent. Using the fasst source separation toolbox for noise robust speech recognition. In *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, 2011.

[46] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[47] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814, 2019.

[48] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, 2018.

[49] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, pages 3244–3248, 2018.

[50] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jimeng Zheng, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu. Neural spatial filter: Target speaker speech separation assisted with directional information. In *Interspeech*, pages 4290–4294, 2019.

[51] Jens Heitkämper, Thilo Feher, Martin Freitag, and Reinhold Haeb-Umbach. A study on online source extraction in the presence of changing speaker positions. In *International Conference on Statistical Language and Speech Processing*, pages 198–209. Springer, 2019.

[52] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu. Multi-modal multi-channel target speech separation. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):530–541, 2020.

[53] Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues. In *Interspeech*, pages 2718–2722, 2019.

[54] Hiroki Sato, Tsubasa Ochiai, Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Shoko Araki. Multimodal attention fusion for target speaker extraction. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 778–784. IEEE, 2021.

[55] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021.

[56] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Interspeech*, pages 2728–2732, 2018.

[57] Ke Zhang, Junjie Li, Shuai Wang, Yangjie Wei, Yi Wang, Yannan Wang, and Haizhou Li. Multi-level speaker representation for target speaker extraction, 2024.

[58] Ke Xue, Rongfei Fan, Shanping Yu, Chang Sun, and Jianping An. Dualstream contextual fusion network: Efficient target speaker extraction by leveraging mixture and enrollment interactions. *arXiv preprint arXiv:2502.08191*, 2025.

[59] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech*, pages 3830–3834, 2020.

[60] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752, 2001.

[61] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010.

[62] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE, 2021.

[63] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr – half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[64] Ashish Alex, Lin Wang, Paolo Gastaldo, and Andrea Cavallaro. Data augmentation for speech separation. *Speech Communication*, 152:102949, 2023.

[65] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[66] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech*, pages 2613–2617, 2019.

[67] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[68] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):7–19, 2015.

[69] Ryo Arakawa, Shinnosuke Takamichi, and Hiroshi Saruwatari. Implementation of dnn-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device. In *Proc. ISCA Workshop on Speech Synthesis*, pages 93–98, 2019.

[70] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25, 2021.

[71] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[72] Ashish Alex, Lin Wang, Paolo Gastaldo, and Andrea Cavallaro. Mixup augmentation of generalizable speech separation. In *Proc. International Workshop on Multimedia Signal Processing*, pages 1–6. IEEE, 2021.

[73] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers

with localizable features. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[74] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. Librimix: An open-source dataset for generalizable speech separation, 2020.

[75] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Timit acoustic-phonetic continuous speech corpus, 1993.

[76] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit, 2019.

[77] Hongyi Zhang, Moustapha Cissé, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[78] Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021.

[79] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Fmix: Enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2021.

[80] Dawei Liang, Fuxun Yang, Tao Zhang, and Peter Yang. Understanding mixup training methods. *IEEE Access*, 6:58774–58783, 2018.

[81] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proc. International Conference on Learning Representations*, 2020.

[82] Shenshen Wei, Shan Zou, Fengnian Liao, et al. A comparison on data augmentation methods based on deep learning for audio classification. *Journal of Physics: Conference Series*, 1453(1):012085, 2020.

[83] Douglas B Paul and Janet M Baker. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[84] LibriVox. Librivox: Free public domain audiobooks. https://librivox.org, 2005. Accessed: 2023-05-01.

[85] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[86] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from librispeech for text-to-speech. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1526–1530, 2019.

[87] Miao Xu and Eng-Siong Chng. Target speaker extraction for overlapped multi-talker speaker verification. *Proc. Interspeech*, pages 1354–1358, 2021.

[88] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech*, pages 2616–2620, 2017.

[89] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, pages 1086–1090, 2018.

[90] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang. Cn-celeb: a challenging chinese speaker recognition dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7604–7608. IEEE, 2020.

[91] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang. Cn-celeb: multi-genre speaker recognition, 2020.

[92] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner,

and Sriram Srinivasan. Interspeech 2021 deep noise suppression challenge. In *INTERSPEECH*, 2021.

[93] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[94] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 411–412, 2013.

[95] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of the 21st International Congress on Acoustics*, pages 1–6, Montreal, Canada, June 2013.

[96] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

[97] Robert F Krueger and Wendy Johnson. Flexible attentional control in young children with learning problems. *Journal of Experimental Child Psychology*, 102(3):265–279, 2009.

[98] Eugene L Allgower and Kurt Georg. *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media, 2003.

[99] Layne T Watson. Globally convergent homotopy algorithms for nonlinear systems of equations. *Nonlinear Dynamics*, 1(2):143–191, 1979.

[100] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. *Proceedings of the 35th International Conference on Machine Learning*, pages 5238–5246, 2018.

[101] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48. ACM, 2009.

[102] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

[103] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1162–1172, 2019.

[104] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongyuan Xie, and Yongdong Zhang. Curriculum learning for natural language understanding. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, 2020.

[105] Zheng Lian Liu, Jianhua Tao, Zheng Wen, Bin Liu, and Carlos Busso. Curriculum learning for speech emotion recognition from crowdsourced labels. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6494–6498. IEEE, 2020.

[106] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(1):7382–7431, 2020.

[107] Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments. In *Conference on Robot Learning*, pages 835–853. PMLR, 2020.

[108] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[109] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani. Speaker-aware neural network adaptation for speaker extraction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6346–6350. IEEE, 2019.

[110] Yun Liu, Xuechen Liu, and Junichi Yamagishi. Improving curriculum learning for target speaker extraction with synthetic speakers. *2024 IEEE Spoken Language Technology Workshop*, 12 2024.

[111] John S. Garofolo et al. Csr-i (wsj0) complete, 1993. LDC93S6A, Linguistic Data Consortium, 1993.

[112] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090, 2018.

[113] Owais Nazir, Hira Taharat, Muhammad Waqas Ahmed, and Ghulam Muhammad Khan. Multi-scale time-domain tasnet with synthetic data augmentation for target speaker extraction. In *International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE, 2024.

[114] Anastasia Kuznetsova, Aswin Sivaraman, and Minje Kim. The potential of neural speech synthesis-based data augmentation for personalized speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[115] Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors. In *Interspeech*, 2023.

[116] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1940–1964, 2020.

[117] Berrak Sisman, K. Vijayan, M. Dong, and Haizhou Li. Singan: Singing voice conversion with generative adversarial networks. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 112–118, 2019.

[118] Hirokazu Kameoka, Kou Tanaka, Damian Kwaśny, Takuhiro Kaneko, and Noritaka Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE, 2018.

[119] Takuhiro Kaneko and Hirokazu Kameoka. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *European Signal Processing Conference (EUSIPCO)*, pages 2100–2104. IEEE, 2018.

[120] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice conversion with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

[121] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

[122] Yuanjun Lv, Jixun Yao, Peikun Chen, Hongbin Zhou, Heng Lu, and Lei Xie. SALT: Distinguishable speaker anonymization through latent space transformation. In *2023 IEEE Automatic Speech Recognition and Understanding (ASRU)*, 2023.

[123] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Andreas Nautsch, Jose Patino, Pravin Srivastava, Paul-Gauthier Noé, Nicholas Evans, Junichi Yamagishi, Jean-François Bonastre, et al. The voiceprivacy 2022 challenge: Preserving privacy in speaker and speech characterization. *Computer Speech & Language*, 81:101439, 2023.

[124] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-François Bonastre. Speaker anonymization using x-vector and neural waveform models. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1278–1282, 2019.

[125] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5329–5333. IEEE, 2018.

[126] Julius Meyer, Nicolas Kourtellis, and Jorge Soria. Speaker anonymization with phonetic intermediate representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE, 2022.

[127] Henry Turner, Shinji Takaki, and Junichi Yamagishi. Neural anonymization and resynthesis in complex frequency domain for speaker de-identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[128] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

[129] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Cheng Zhang, Yanqing Liu, Yuchao Chai, Xu Zhao, Yu Huang, Wei Zhang, et al. Naturalspeech: End-to-end text to speech synthesis with human-level quality. In *Advances in Neural Information Processing Systems*, pages 37653–37670, 2022.

[130] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[131] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, 2020.

[132] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Van Den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[133] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and Others. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In *Proc. ICASSP*, pages 4779–4783, 2018.

[134] Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech. *arXiv preprint arXiv:2006.04558*, 2020.

[135] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[136] Qiantong Hu, Graham Cohn, Sung-Feng Chien, Yu Cao, Shengyao Li, Xinjian Song, Ye Wang, and Rick Legetter. Synt++: Utilizing enhanced synthetic data for end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6092–6096. IEEE, 2022.

[137] Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. SynthASR: Unlocking synthetic data for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 899–903, 2021.

[138] Jinghua Ding, Yujie Zhao, Siqi Cui, Xiaowei Chi, Feng Wang, Wei Liu, and Qingyun Zhao. Data augmentation with linear filtering for automatic speaker verification. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4061–4065, 2020.

[139] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François

Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language*, 64:101114, 2020.

[140] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4604–4608. IEEE, 2022.

[141] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *International Conference on Machine Learning*, pages 1180–1189, 2015.

[142] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[143] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu. Neural target speech extraction: An overview. *IEEE Signal Processing Magazine*, 40(3):101–114, 2023.

[144] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1370–1384, 2020.

[145] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li. Spex+: A complete time domain speaker extraction network. In *Proc. of INTERSPEECH*, pages 1406–1410, 2020.

[146] Wupeng Wang, Zexu Pan, Xinke Li, Shuai Wang, and Haizhou Li. Speech separation with pretrained frontend to minimize domain mismatch. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4184–4198, 2024.

[147] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

[148] Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors. In *Interspeech*, 2023.

[149] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language*, 64:101114, 2020.

[150] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, Nicholas Evans, Massimiliano Todisco, Jean-François Bonastre, and Mickael Rouvier. Synvox2: Towards a privacy-friendly VoxCeleb2 dataset. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11421–11425, 2024.

[151] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko. Speaker anonymization using orthogonal householder neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3681–3695, 2023.

[152] Kavita Kasi. Yet another algorithm for pitch tracking (YAAPT). Master's thesis, Old Dominion University, 2002.

[153] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[154] Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. ICASSP 2021 deep noise suppression challenge. In *ICASSP*, 2021.

[155] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[156] Zeyu Gan and Yong Liu. Towards a theoretical understanding of synthetic data in LLM post-training: A reverse-bottleneck perspective. *arXiv preprint arXiv:2410.01720*, 2024.

[157] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2019.

[158] Sören Mindermann, Olivier Bachem, Sebastian Ruder, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. *International Conference on Machine Learning*, pages 15630–15649, 2022.

[159] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. In *arXiv preprint arXiv:1811.10959*, 2021.

[160] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021.

[161] Katerina Zmolikova, Marc Delcroix, K. Kinoshita, Tsubasa Ochiai, T. Nakatani, L. Burget, and Jan Honza Cernocky. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13:800–814, 2019.

[162] Jiahe Wang, Shuai Wang, Junjie Li, Ke Zhang, Yan min Qian, and Haizhou Li. Enhancing speaker extraction through rectifying target confusion. *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 349–356, 2024.

[163] Sanyuan Chen, Yu Wu, Zhuo Chen, Jinyu Li, Chengyi Wang, Shujie Liu, and M. Zhou. Continuous speech separation with conformer. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5749–5753, 2020.

[164] H. Taherian and Deliang Wang. Time-domain loss modulation based on overlap ratio for monaural conversational speaker separation. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5744–5748, 2021.

[165] Jun Yu Li, Ruijie Tao, Zexu Pan, Meng Ge, Shuai Wang, and Haizhou Li. Audio-visual active speaker extraction for sparsely overlapped multi-talker speech.

ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 10666–10670, 2023.

[166] Zexu Pan, Meng Ge, and Haizhou Li. Usev: Universal speaker extraction with visual cue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3032–3045, 2021.

[167] Yun Liu, Xuechen Liu, and Junichi Yamagishi. Improving curriculum learning for target speaker extraction with synthetic speakers. In *2024 IEEE Spoken Language Technology Workshop*, 2024.

[168] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, 2020. Association for Computational Linguistics.

[169] Yun Liu, Xuechen Liu, Xiaoxiao Miao, and Junichi Yamagishi. Target speaker extractor training with diverse speaker conditions and synthetic data. *APSIPA Transactions on Signal and Information Processing*, 14(1):e30, 2025.

[170] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *ArXiv Interspeech 2020*, pages 3830–3834, 2020.

[171] Yun Liu, Xuechen Liu, Xiaoxiao Miao, and Junichi Yamagishi. Target speaker extraction with curriculum learning. *ArXiv*, abs/2406.07845, 2024.

[172] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, pages 5998–6008, 2017.

[173] Songting Liu. Zero-shot voice conversion with diffusion transformers. *ArXiv*, abs/2411.09943, 2024.

[174] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *ArXiv*, abs/2410.06885, 2024.

[175] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

[176] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *International Conference on Learning Representations*, 2021.

[177] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608, 2021.

A List of Publications

A.1 Publications For This Thesis

The following publications have been produced as part of this thesis research:

- **Yun Liu**, Xuechen Liu, Junichi Yamagishi, "Training Dynamics-Aware Multi-Factor Curriculum Learning for Target Speaker Extraction," Proc. ICASSP 2026 — IEEE International Conference on Acoustics, Speech and Signal Processing.
- **Yun Liu**, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi, "Target Speaker Extractor Training with Diverse Speaker Conditions and Synthetic Data," APSIPA Transactions on Signal and Information Processing: Vol. 14, No. 1, e30, 2025. DOI: 10.1561/116.20250054.
- **Yun Liu**, Xuechen Liu, Junichi Yamagishi, "Improving Curriculum Learning For Target Speaker Extraction With Synthetic Speakers," 2024 IEEE Spoken Language Technology Workshop (SLT), Macao, 2024, pp. 364-370, doi: 10.1109/SLT61566.2024.10832348.

- **Yun Liu**, Xuechen Liu, Xiaoxiao Miao, Junichi Yamagishi, “Target Speaker Extraction with Curriculum Learning,” Proc. Interspeech 2024, pp.4348-4352, doi:10.21437/Interspeech.2024-1375
- Haoyu Li, **Yun Liu**, Junichi Yamagishi, "Joint Noise Reduction and Listening Enhancement for Full-End Speech Enhancement". Proc. ICASSP 2023 — IEEE International Conference on Acoustics, Speech and Signal Processing. DOI: https://doi.org/10.1109/icassp49357.2023.10094974.

B

Online Resources

This appendix lists the resources attached to the thesis.

- **Libri2Vox Dataset**: The novel dataset introduced in Chapter 5 is available at: https://github.com/nii-yamagishilab/Libri2Vox