APSIPA Transactions on Signal and Information Processing, 2025, 14, e30

Original Paper

Target Speaker Extractor Training with Diverse Speaker Conditions and Synthetic Data

Yun Liu[1,2*], Xuechen Liu[1], Xiaoxiao Miao[3] and Junichi Yamagishi[1,2*]

[1] *National Institute of Informatics, Tokyo, Japan*
[2] *Sokendai, Kanagawa, Japan*
[3] *Division of Natural and Applied Sciences, Duke KunShan University, Kunshan, China*

ABSTRACT

Target speaker extraction (TSE) is essential for various speech processing applications, particularly with complex acoustic environments. However, current TSE systems lack robustness under real world conditions due to limited training data diversity and unrealistic noise. To address these challenges, we first constructed Libri2Vox, a new dataset combining clean target speech from LibriTTS with interference speech from VoxCeleb2 that contains real acoustic variations, channel effects, and ambient conditions. To increase speaker variability, we augmented Libri2Vox with synthetic speakers generated by developing two speech anonymization methods: SynVox2 and SALT (speaker anonymization through latent space transformation). Further, we propose a three-stage curriculum learning approach that progressively introduces synthetic speakers after training a seed TSE model on real data with varying speaker similarity levels. Experiments with four different neural TSE models show that Libri2Vox's rich acoustic variations

*Corresponding author: yunliu@nii.ac.jp. This study is partially supported by MEXT KAKENHI Grants (24K21324) and JST, the establishment of university fellowships towards the creation of science technology innovation (JPMJFS2136).

Received 26 June 2025; accepted 12 August 2025
ISSN 2048-7703; DOI 10.1561/116.20250054

and synthetic speaker integration through curriculum learning consistently improve performance across common evaluation metrics. We also confirmed that the proper ratio of synthetic speakers to real speakers is important for improving the performance.[1]

Keywords: Target speaker extraction, curriculum learning, synthetic data, speech dataset

1 Introduction

Target speaker extraction (TSE) [55] is a key task in speech processing, focusing on isolating the voice of a desired speaker from complex acoustic environments. This capability is valuable in applications such as voice-controlled systems, teleconferencing, and hearing aids, where extracting clear speech signals directly impacts system performance and user experience. Despite notable advances, TSE still faces multiple challenges, particularly related to limited data diversity and a lack of robustness under real-world conditions [55].

A significant issue with TSE is the mismatch between training and deployment environments. Current models are typically trained on artificially mixed speech, which is controlled but fails to capture the complexity of real-world conditions [48]. The diverse nature of noise, including spatial configurations, reverberations, and dynamic changes, leads to significant performance degradation during deployment. The limitations of current TSE datasets (highlighted in Table 1) are evident in both restricted speaker diversity and the controlled, synthetic nature of mixtures. For instance, datasets such as WSJ0-2mix-extr [50, 15] and Libri2talker [51] are limited regarding the number of speakers, 101 and 1172, respectively. These datasets possess limited variability in terms of acoustic and speaker conditions. These limitations can lead to models that generalize poorly to unseen speakers and complicated real-world acoustic environments. Addressing these challenges and improving the robustness of TSE systems thus requires incorporating more variations in terms of noise conditions and speakers.

To address these limitations and enhance TSE system generalization, we propose a novel data integration approach focusing on two critical aspects: speaker diversity and acoustic variability. We leverage the VoxCeleb2 dataset [4], which encompasses more than 6,000 speakers recorded in diverse acoustic environments, providing a rich source of real-world variations. However, directly using the noisy recordings from VoxCeleb2 as target speech contradicts

[1]We will make the Libri2Vox dataset and code be public upon the acceptance of the paper.

Table 1: Comparison of WSJ0-2mix-extr, Libri2mix, and Libri2talker datasets.

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

the objective with TSE, which is to extract clean speech, thus degrading the effectiveness of the training data. To resolve this constraint while maintaining data diversity, we implement a strategic combination: employing the high-fidelity LibriTTS dataset [53], derived from LibriSpeech [34], as our target speaker source while using VoxCeleb2 for interference speaker source.

Another solution to address such challenges is the acquisition of synthetic data, which has demonstrated remarkable efficacy across various alternative tasks [26, 32], such as computer vision [28, 9] and natural language processing [10, 12, 46]. It has been a promising solution on data scarcity and model robustness. Synthetic data have also proven highly effective in various speech-related tasks. In automatic speech recognition (ASR), for example, the use of synthetic speech data generated using text-to-speech (TTS) models has demonstrated considerable improvements [18, 11]. Similarly, research on TTS systems has shown that training robust models on synthetic data produced using less stable systems can enhance transfer stability, delivering high-quality transfers while retaining speaker characteristics [40, 2]. For speaker-related tasks, synthetic data also enables TTS systems to synthesize speech from new, unseen speakers by sampling from a learned latent distribution [22, 54]. Building upon these advances, we extend synthetic data approaches to TSE. Our goal is to enable TSE models to reliably extract clean speech in diverse, real-world environments. Incorporating synthetic data helps address challenges of data scarcity, especially in handling new speakers and complex acoustic conditions.

Our two previous studies laid the foundation for this study. In our initial study [24], we implemented a curriculum learning (CL) approach [41] that progressively trains the model with increasingly complex scenarios, demonstrating significant performance improvements in challenging speaker extraction tasks. In our subsequent research [25], we introduced the use of synthetic speakers to the learning scheme. The synthetic speakers were generated through voice conversion and speaker anonymization [27], which led to

substantial improvements in TSE performance. Voice conversion transforms a source speaker into a target speaker, producing speech with a different identity. Speaker anonymization, similarly, takes speech in but removes the speaker's identity, creating a new, anonymized speaker that does not exist in reality. For this study, we created Libri2Vox, a dataset combining clean target speech from LibriTTS with interference speech from VoxCeleb2 (a larger dataset with more speakers than the current TSE dataset), that enables the generation model to generate a greater number and variety of synthetic speakers.

The main contributions of this study are threefold:

- **New dataset called Libri2Vox**: Libri2Vox includes a large and diverse set of speakers with realistic noisy interference. The target speakers are sourced from the cleaner LibriTTS, while interference speakers are derived from VoxCeleb2. This combination allows for better representation of real-world acoustic conditions while maintaining clean target speech signals. Focusing specifically on 2-speaker scenarios, Libri2Vox bridges the gap between idealized synthetic mixtures and uncontrolled recordings by incorporating VoxCeleb2's natural acoustic variations, channel effects, and ambient conditions. This design choice provides a controlled yet more realistic environment for developing and evaluating TSE systems, maintaining the necessary ground truth references while introducing more challenging and varied acoustic conditions than conventional TSE datasets.

- **Libri2Vox variants with synthetic speakers**: We further enrich the diversity of Libri2Vox by introducing two specialized synthetic speaker generation techniques, each designed to improve the robustness of TSE. Unlike generic data augmentation techniques that simply manipulate existing data, our approaches generate entirely new synthetic speakers that are acoustically distinct yet complementary to real speakers. These synthetic speakers strategically fill gaps in the speaker representation space, introducing variability patterns not present in the original data. Our previous study demonstrated that increasing the diversity of interference speakers through synthetic generation can significantly enhance TSE performance, particularly when combined with CL [25]. In this study, we built upon these findings by leveraging the more diverse VoxCeleb2 dataset, which provides a richer pool of speakers as the foundation for our synthetic speaker generation framework. This approach represents a methodological innovation in how synthetic data can be specifically optimized for speaker extraction tasks rather than general speech applications.

- **Investigating effectiveness of synthetic speakers in curriculum learning**: We propose a novel three-stage CL framework that strate-

gically integrates synthetic data to maximize TSE performance. Our approach first establishes model robustness on real data with varying speaker similarity levels before gradually introducing synthetic speakers in a three-stage learning curriculum. Experiments show this curriculum-based integration significantly boosts performance while avoiding the severe degradation observed when training directly on synthetic data alone.

2 Related Work

This section defines TSE, discusses deep neural network (DNN)-based TSE models, provides an overview of common datasets while identifying their shortcomings, and presents effective training strategies.

2.1 Definition of Target Speaker Extraction

The basic framework TSE is illustrated in Figure 1. Mathematically, it can be expressed as:

$$\hat{\mathbf{w}}^{(t)} = \text{TSE}(m, e_t = E(\mathbf{w}^{(r_t)}); \theta),$$

where given a enrollment waveform $\mathbf{w}^{(r_t)}$ containing speech signals of the target speaker t, this waveform is used to extract a speaker embedding $e_t = E(\mathbf{w}^{(r_t)})$ via a neural speaker encoder E. TSE aims to output the estimated clean speech signals $\hat{\mathbf{w}}^{(t)}$ of the target speaker from a given mixture m, where $m = s + s'$ contains the target speaker's clean speech s and interference speakers' speech s'. The notation θ represents the model parameters of the extraction framework.

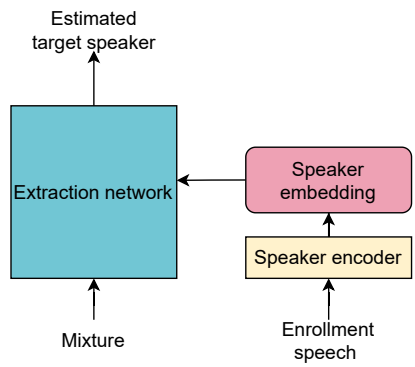

Figure 1: Basic conceptual framework of TSE.

2.2 Deep Neural Networks for Target Speaker Extraction

DNNs have enabled significant advances in developing single-channel TSE systems. Early approaches primarily used a reference signal from the target speaker to guide the extraction process, differing in network architectures and the manner in which speaker embeddings were used. The following DNNs are the most common for TSE.

SpeakerBeam [56] was the first model specifically designed for TSE. Unlike other speech separation models that attempt to determine the number of speakers in a mixture, SpeakerBeam focuses exclusively on extracting the target speaker. By leveraging speaker information through embeddings, it effectively isolates the desired speaker's voice, overcoming common issues such as label permutation and speaker tracing. Notably, SpeakerBeam's speaker embedding is extracted using a simple speaker encoder trained jointly with the TSE network.

VoiceFilter [47] integrates convolutional layers, long-short term memory (LSTM), and fully connected layers (FCs). Unlike SpeakerBeam, speaker embeddings used with VoiceFilter are extracted from a pre-trained speaker encoder [43], providing fixed guidance.

The Conformer [16] architecture combines convolutional layers and self-attention mechanisms to simultaneously capture local and global dependencies in speech input. The Conformer-based TSE model [24] processes the time-frequency domain short-term Fourier transform (STFT) spectrum of the input mixture. By using a Conformer blocka hybrid of multi-head self-attention and full convolutionConformer generates the real and imaginary parts of the target speech signal's STFT. Conformer blocks also integrate STFT features with speaker embeddings extracted from reference utterances using a pre-trained speaker encoder [7]. Unlike SpeakerBeam and VoiceFilter, which enhance only the magnitude and use the noisy phase for reconstruction, Conformer predicts the complex spectrum mask [49].

We use all of these models to prove the effectiveness of Libri2Vox and our training strategy.

2.3 Datasets for Target Speaker Extraction

Table 1 presents three major datasets for TSE: WSJ0-2mix-extr, Libri2mix, and Libri2talker.

WSJ0-2mix-extr [50] is derived from the WSJ0 [14] corpus, which consists of clean, read speech from the Wall Street Journal. For each 2-talker mixture audio in WSJ0-2mix-extr, two randomly selected utterances from different speakers in WSJ0 are mixed. This dataset is composed of training, development, and evaluation sets, with the training set including 20,000 mixtures generated from 101 speakers (50 male and 51 female), development set con-

taining 5,000 mixtures, and evaluation set containing 3,000 mixtures involving 18 different speakers not seen during training. The signal-to-noise ratio (SNR) of these mixtures was chosen between 0 and 5 dB. Most of the 2-talker speech samples are heavily overlapped, creating a challenging environment for TSE models.

Libri2mix [5] is derived from LibriSpeech [34]. It creates clean 2-talker mixtures by randomly selecting and mixing two utterances from its predecessor. It consists of several subsets, including train-360 with 50,800 utterances (212 hours, 921 speakers) and train-100 with 13,900 utterances (58 hours, 251 speakers), as well as dev and test sets, each containing 3,000 utterances (11 hours, 40 speakers). The dataset follows a minimum duration protocol, trimming the longer utterance in a pair to match the shorter one, resulting in a 100% overlap rate. This setup is designed to provide challenging conditions for speech separation models.

Libri2talker, an extended version of Libri2mix, reuses 2-talker mixtures by swapping the roles of the target and interference speakers, effectively doubling the available data. It includes a training set with 127,056 examples from 1,172 speakers, validation set with 2,344 examples, and evaluation set with 2,260 enrollment utterances from the LibriSpeech test-clean set and 6,000 test utterances from the Libri2mix test set. Like Libri2mix, Libri2talker applies the minimum duration protocol and retains a 100% overlap rate, making it suitable for both speaker verification and TSE tasks.

These datasets are composed of artificially mixed speech, where both speakers in each mixture were recorded under studio conditions, resulting in clean speech. This setup leads to a significant mismatch with real-world data, which typically contains more variability and noise. Since our ultimate goal is to handle real-world scenarios, using real-world data as interference sources is a more suitable approach to bridge this gap and enhance model robustness.

2.4 Training Strategies for Separation-related Task

There are predominately two categories of state-of-the-art training strategies for TSE: data simulation and optimization strategies. Data simulation is crucial for augmenting mixture training data, particularly when real-world labeled datasets are limited. Common techniques include the following:

- **Data Augmentation** [1]: This involves generating new training data by modifying existing datasets using techniques such as adding Gaussian noise, pitch shifting, or time stretching. Data augmentation helps train more robust models that generalize better to different environments.
- **Dynamic Mixing**: This strategy dynamically generates new mixtures of target and interference speakers during training. By continuously

varying the conditions in which the target speaker is extracted, dynamic mixing improves the model's generalization to different noise and interference scenarios.

Optimization strategies have also demonstrated significant promise in improving TSE performance. Common strategies include the following:

- **Metric-based Method**: To address the training/inference mismatch in deep noise suppression models, real data can be implemented using either generative models or reference-free loss without clean speech access [52]. With this strategy, an end-to-end non-intrusive DNN, called PESQ-DNN, is used to estimate the perceptual evaluation of speech quality (PESQ) [39] scores, providing a reference-free perceptual loss during training. An alternating training protocol is applied in which the DNN model is updated on real data, followed by PESQ-DNN updates on synthetic data. This strategy significantly improves the performance compared with models trained solely on simulated data.

- **Curriculum Learning**: CL is a strategy with which training data is introduced progressively, starting with easier examples and moving toward more complex ones. CL has been implemented in TSE by sorting training samples on the basis of predefined difficulty measures such as gender, speaker similarity, signal-to-distortion ratio (SDR), and SNR. Initially, easier samples in which the target and interfering speakers are more distinct are used, and progressively harder cases are introduced as training advances. CL has been shown to improve model convergence and performance [25, 24] and was used in our experiments to optimize TSE training.

3 Libri2Vox Dataset

TSE systems face two critical challenges: limited speaker diversity and artificial acoustic environments. Current TSE datasets, as shown in Table 1, typically contain only hundreds to a few thousand speakers, significantly constraining model generalization capabilities. This limitation becomes particularly apparent when systems encounter speakers or acoustic conditions outside the training distribution.

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

Apart from speaker diversity, acoustic condition is another intriguing factor. Most existed datasets for standard TSE training and evaluation are generated by artificially adding target and interference speech with noise, which does not 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 difficult 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 used 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 leverage VoxCeleb2 recordings as interference sources. These recordings inherently contain diverse, real-world acoustic conditions, providing more real training scenarios for TSE models compared with 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.

3.1 Dataset Construction

In constructing the *Libri2Vox* database, as shown in Figure 2, we used the following key steps.

3.1.1 Pre-processing and Mixing

Each audio segment is randomly split into 6 second(s) segments. Segments shorter than 6s are zero-padded at the end. The dataset is first processed in the following steps:

- All audio data is pre-processed using sv56[2] with a scale factor of -26 dB.
- In *LibriTTS*, target speech shorter than 2s is deleted. Then, speakers with fewer than 3 utterances are removed, resulting in the deletion of 31 speakers out of 1,151, leaving 1,120 speakers in total.
- For reference speech handling, 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

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

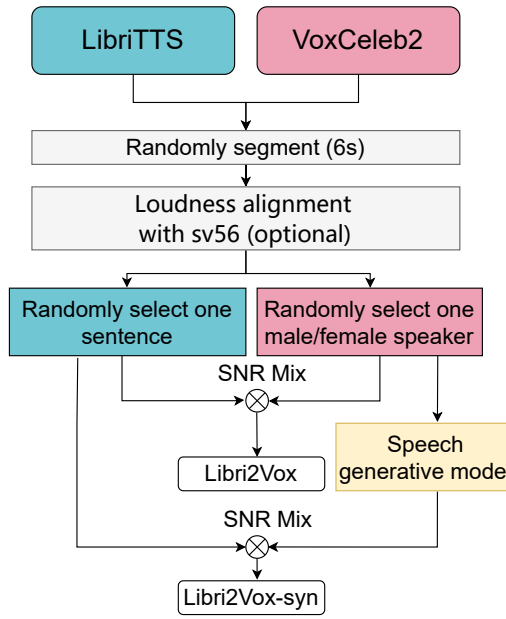

Figure 2: Data generation framework of Libri2Vox and its synthetic version.

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

For the dataset composition, we use 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", meaning that the speech segments are naturally corrupted by real-world noise such as laughter, cross-talk, channel effects, music, and other background sounds. This adds 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 wide range of accents, ages, and languages.

Regarding speaker diversity characteristics, Libri2Vox inherits rich demographic variety from both source datasets. From VoxCeleb2, we retain 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 strong 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 stylesranging 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 problem with VoxCeleb2 for this study is that the speech segments are predominantly recorded in noisy conditions. However, on the other hand, the inherent noise in VoxCeleb2 makes it an ideal source for interference speakers. By using VoxCeleb2 as the interference speaker source, we can take advantage of 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 leave 5,900 speakers for training.

The other data source acquired for this study is LibriTTS [53]. It is a refined version of LibriSpeech [34], in which a small portion of the noisy utterances from the latter was removed 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 means 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 created 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.

3.2 Statistics of Libri2Vox

Libri2Vox was pre-split after collecting and processing, as shown in Table 2 For the training set, *LibriTTS* provides 1,151 speakers, with 8.97 hours of data. Each *LibriTTS* utterance generates one corresponding data point (mixture, reference, target). VoxCeleb2 contributes 5,900 interference speakers for this partition, making the training set contains 149,691 triplets, equivalent to 250 hours of mixture data. The validation set consists of 40 *LibriTTS* speakers (approximately 8.97 hours) and 94 VoxCeleb2 speakers, totaling 134 speakers and 7,200 utterances. The test set includes 39 *LibriTTS* speakers (approximately 8.56 hours) and 118 VoxCeleb2 speakers, with a total of 157 speakers and 6,000 utterances.

Table 2: Libri2Vox statistics. "Total" column shows total number of utterances combined with corresponding LibriTTS and VoxCeleb2 sets.

Set	**# of Speakers**		**# of Utterances**	**Duration (h)**
	LibriTTS	**VoxCeleb2**	**Total**	
Training	1,151	5,900	149,691	250
Validation	40	94	7,200	8.97
Testing	39	118	6,000	8.56

4 Synthetic Libri2Vox Dataset

4.1 Why Synthetic Speakers?

Enhancing the diversity of training data can significantly improve the performance of TSE models. Conventionally, this has been achieved by applying data augmentation to real data [1]. However, the diversity provided by data augmentation is limited in terms of the range of data distribution it can cover. Another approach to generating large amounts of data with diverse speaker characteristics is to use speech generative models [19, 44, 38]. These recent models enable speech generation with a high level of naturalness and speaker similarity. This development brings up a key question: can these generative models be used to generate specialized training data for TSE?

There are several strategies to generate training data with generative models, one of which involves producing diverse synthetic interference speakers from the existing interference ones, ensuring the defined difference [29, 27]. This is the strategy we investigated.

4.2 Two Types of Synthetic Interference Speaker Generation Methods

We developed two different generation methods used to generate synthetic interference speakers that are distinct from the real speakers selected from VoxCeleb dataset, i.e. SynVox2 [29] and speaker anonymization through latent transformation (SALT) [27]. These two methods are used to increase the diversity of interference speakers used in TSE tasks. Below is a description of **SALT** .

4.2.1 SALT

SALT generates synthetic interference speakers by manipulating the latent space of pre-trained speaker representations. In the context of speaker extraction, let s_i represent the given interference speaker's audio. The steps to generate a synthetic speaker with SALT are as follows:

- **WavLM Representation Extraction**: We first extract the WavLM [3] representation, WavLM(s_i), of the input interference speaker speech s_i. This representation captures both the content and speaker-related features, as evident in that previous study [3]. Then, reference speaker (s_r) representations, WavLM(s_r^j), $j \in [1, N]$, are also extracted from a pool of N speakers by WavLM.
- **k-nearest neighbor (k-NN) Search**: Given WavLM(s_i) as the query, a k-nearest neighbor (k-NN) [6] search is conducted for each frame of the query representation. At each time step, the k most similar frames are selected from the reference speaker representations of all N speakers. This process yields a set of closest representations $\mathbf{D}_j$ for each of the N selected reference speakers.
- **Weighted Summation**: After selecting k-NN representations, random weights w_j are assigned to each of the reference speaker representation sets, $\mathbf{D}_j$. These weights are sampled from a normal distribution and normalized to sum to one. The weighted sum of the selected representations is then combined with the original interference speaker representation

via linear interpolation, as follows:

$$(\mathbf{D}_1, \cdots, \mathbf{D}_N) = \text{kNN}(\text{WavLM}(s_i), \text{WavLM}(s_r^j))$$
$$\mathbf{O} = (1-p)\sum_{j=1}^{N} w_j \mathbf{D}_j + p \cdot \text{WavLM}(s_i)$$

where p is a parameter that controls the balance between the original and synthetic representations.

- **Vocoder-based Reconstruction**: Finally, the interpolated representation $\mathbf{O}$ is passed through the HiFi-GAN vocoder [21] to generate a waveform of the synthetic interference speaker.

For this study, the number of nearest neighbors considered for interpolation was $k = 4$, and the interpolation weight between the original speaker representation and synthetic speaker representation was $p = 0.5$. We used the `WavLM-Base`[3] model trained on LibriSpeech, to extract latent space representations from the third layer. For the interpolation method, 50 speakers are randomly chosen from the LibriSpeech train-clean-100 dataset. For each of these speakers, 50 audio samples are selected at random to extract their features. The number of random reference speakers, N, is fixed at 4. The setting is the same as in the original paper [27].

By using SALT, we can generate a large variety of synthetic interference speakers that are different from real speakers, providing a more challenging dataset for TSE model training. This method enables for a balance between speaker similarity and diversity through parameters k and p, which enables better control over the generated synthetic speakers.

4.2.2 SynVox2

SynVox2 was designed for speaker anonymization using an orthogonal Householder neural network (OHNN) [30, 29]. The framework operates through the following three essential components:

- **Disentanglement:** Speech characteristics are derived using specialized encoders. The Yet Another Algorithm for Pitch Tracking (YAAPT) [20] extracts the fundamental frequency (F0), while the ECAPA-TDNN [7] speaker encoder, trained on VoxCeleb2, generates 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 [17], captures linguistic content information.

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

- **Anonymization:** The system use an OHNN-based anonymizer [30] that transforms original speaker embeddings into anonymized representations through multiple orthogonal Householder transformation layers. The network uses 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 integrates the extracted content features, F0 information, and anonymized speaker embeddings into a HiFi-GAN model trained on LibriTTS-train-clean-100. This integration produces high-quality anonymized speech waveforms that preserve natural speech characteristics while ensuring distinct speaker identities.

4.3 Synthetic Data and Statistics

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

5 Constructing TSE Models on Libri2Vox

5.1 Architecture of Different Target Speaker Extraction Models

This section describes the four TSE neural networks we used for the experiments: Conformer [16], VoiceFilter [47], SpeakerBeam [56], and bidirectional LSTM (BLSTM) (see Figure 3).

5.1.1 Conformer

The application of the acquisition of time-frequency representation for Conformer-based TSE was proposed [24]. Conformer has a hybrid architecture, combining convolutional layers and multi-head attention mechanisms. This makes it effective in capturing both local and global features in the input audio.

Input Process: The input consists of a 256-dimensional real part and 256-dimensional imaginary part of the STFT (obtained from a 512-point FFT with the direct current component removed), along with a 192-dimensional x-vector (a speaker embedding extracted using ECAPA-TDNN [7] and concatenated along the time dimension).

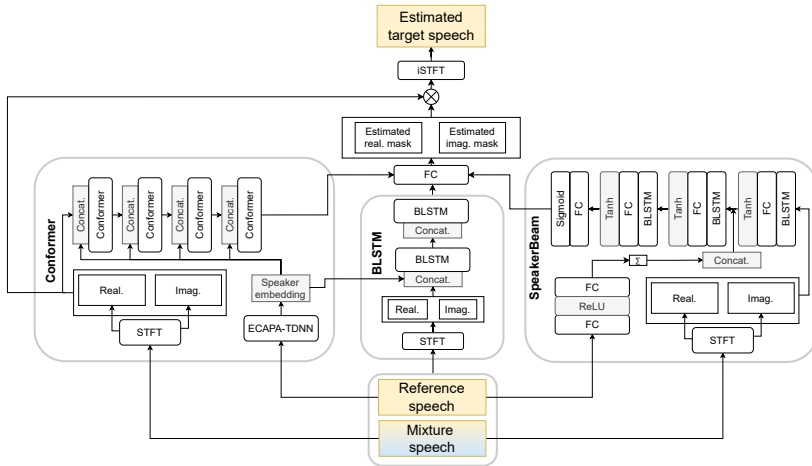

Figure 3: Details of Conformer, BLSTM, and SpeakerBeam TSE models.

Extraction Network (Conformer blocks): The model consists of four stacked Conformer blocks:

- Feedforward Layers: Each block begins with a feedforward layer of size 1024, with dropout set to 0.2.
- Multi-Head Attention: Multi-head attention is applied with four heads and a dropout rate of 0.2.
- Convolutional Layers: A convolution layer with a kernel size of 3, followed by batch normalization and the Swish activation function, is applied with dropout.
- Residual Connections: Residual connections with half-scaling are applied throughout the network.
- Final Linear Layer: The final output is a 512-dimensional feature vector, which is split into real and imaginary parts to compute the complex mask for STFT reconstruction.

Output Process: The model outputs a complex ratio mask [49] to apply to the complex-valued mixture STFT for reconstructing the target speaker's waveform.

5.1.2 BLSTM

BLSTM is a variation of the Conformer, in which the Conformer blocks have been replaced with BLSTM layers. Other aspects remain unchanged, except for the final FC, which is adjusted from 1024 to 512 dimensions to ensure consistent output.

5.1.3 SpeakerBeam

SpeakerBeam [56] is based on a BLSTM speaker encoder that extracts the target speakers voice using the speaker embedding as a guide. Different from the original SpeakerBeam , which uses magnitude spectra as input, we use complex spectra, as it has shown significant improvements over magnitude spectra. We use the same network architecture from the original version to extract speaker information. In such a case, only the input and output have been changed. This was done to evaluate whether our method is effective for the inner speaker information extraction network as well. Further details can be found in the complete architecture description in Appendix A.1.

5.1.4 VoiceFilter

Analogous to SpeakerBeam, only the input and output have been changed to complex spectra for VoiceFilter, while the pre-trained d-vector used in the original version was replaced with the same x-vector as Conformer. The rest of the structure remains the same. Further details can be found in the complete architecture description in Appendix A.2.

5.2 Speaker Information Extraction Model

ECAPA-TDNN [7] is a state-of-the-art speaker encoder. It uses convolutional and residual blocks for feature extraction, followed by attentive statistical pooling [33] to generate speaker embeddings. The model is trained with additive angular margin softmax [45]. The original model, available via SpeechBrain [35], was trained on VoxCeleb1 [31] and VoxCeleb2 [4].

For this study, we used the training framework from SpeechBrain, while retraining the model on the CN-Celeb dataset [23]. Although our target dataset contains English speech recordings, the model trained on CN-Celeb demonstrated superior performance compared with training on the original VoxCeleb 1+2 datasets. This improvement could be attributed to the increased speaker diversity and noise present in CN-Celeb.

5.3 Loss Function and Evaluation Metrics

We use both SNR as the loss function and SDR as the evaluation metric to assess the quality of separated or enhanced speech in TSE tasks.

5.3.1 Loss Function

The loss function used in this study is based on a negative SNR, calculated as the ratio between the power of the target speech and error (difference between the target speech and predicted target speech) in the time domain. It is expressed as:

$$\text{Loss}_{snr} = -10 \log_{10} \frac{s^2}{(s - \hat{s})^2}, \tag{1}$$

A higher SNR indicates better reconstruction of the target speech. where s represents the ground truth target speaker's speech, and $\hat{s}$ is the network's estimated target speaker's speech. The negative sign is used to convert the SNR into a loss value that can be minimized during training.

5.3.2 Evaluation Metric

To evaluate our TSE system performance, we used the widely-used SDR metric [50, 51], which measures the ratio of the target speech's power to the power of distortions introduced by the extraction system. For SDR computation, we use the implementation provided by torchmetrics [8]. For evaluation, improvement is measured using the improvement in SDR (iSDR), defined as the relative increase in SDR compared to the mixture. Specifically, iSDR is calculated as the difference between the SDR of the extracted target speech relative to the clean target speech, and the SDR of the original mixture relative to the clean target speech.

In addition to SDR, we use several perceptual quality metrics. The Perceptual Evaluation of Speech Quality (PESQ) [39] provides a mean opinion score (MOS) that correlates with subjective quality assessments, with values typically ranging from -0.5 to 4.5. The Short-Time Objective Intelligibility (STOI) [42] measure evaluates speech intelligibility by comparing the temporal envelopes of clean and processed speech, with values ranging from 0 to 1. Furthermore, we use deep noise suppression mean opinion scores (DNSMOS) [37], a non-intrusive speech quality assessment model trained to predict human ratings of speech quality without requiring a reference signal, with scores ranging from 1 to 5. For all these evaluation metrics, higher values indicate better performance.

6 Experiments

6.1 Experimental Setup and Model Configurations

All TSE models were trained using a custom learning rate scheduler, designed to adjust the learning rate dynamically on the basis of 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 Œ 10^{-3}, with a minimum threshold of 1 Œ 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 on the basis of 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 enabled for smooth transitions during training while avoiding rapid drops in learning rate that could destabilize the optimization process.

The Adam optimizer was used 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 conducted 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.

6.2 Training Strategy

6.2.1 Noisy Data Augmentation

We used noise from the deep noise suppression(DNS) Challenge dataset [36]. During training, there was a 50% chance that a randomly selected 6-second noise segment would be dynamically mixed with the target speakers audio. The SNR for this dynamic mixing was uniformly sampled from the range of $[-5, 10]$ dB. The purpose of applying noise augmentation is to explore whether it can further enhance the model's performance, especially since VoxCeleb2 already contains some real-world noise. We chose to implement augmentation during training rather than embedding it directly in the Libri2Vox dataset to maintain dataset integrity and flexibility. This approach allows us to systematically investigate how additional noise affects performance while preserving the natural acoustic variations already present in the VoxCeleb2 recordings.

6.2.2 Curriculum Learning

To enhance the models ability to distinguish between target and interference speakers with varying degrees of similarity, we implemented a CL strategy. Following our previous research[25], the training process is divided into three stages, as illustrated in Figure 4:

- **Stage 1:** In the first stage, the training data consists primarily of target and interference speaker pairs with low similarity. The goal at this stage is to enable the model to focus on simpler tasks, thus establishing a solid foundation for learning speaker characteristics.
- **Stage 2:** In the second stage, the model is exposed to speaker pairs with higher similarity, which gradually increases the complexity of the task.
- **Stage 3:** Finally, in the third 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.

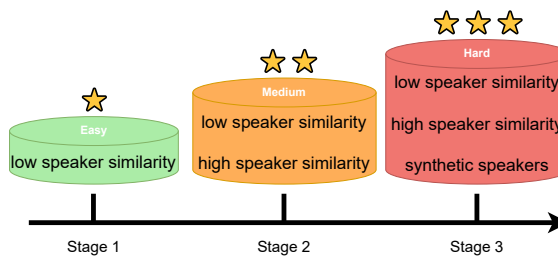

Figure 4: Three-stage CL.

7 Main Results

Using Libri2Vox and its synthetic version, we conducted extensive experiments to evaluate the impact of synthetic data integration with CL on TSE performance, with main results presented in Table 3. We implemented 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 CL with real data, and two variants

of three-stage CL incorporating synthetic data (SynVox2 or SALT). For the "(Real only)" configuration, Stage 2 incorporates the complete real dataset, and Stage 3 continues with the same complete real dataset for additional training epochs. This design allows us to isolate the impact of CL from potential benefits of extended training epochs.

Table 3: iSDR(0.07dB), iPESQ(1.514), iSTOI(0.734), iDNSMOS(2.356), results of three stage methods for Libri2Vox test set. 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	1.325	0.198	0.650
	w/o CL (Real+SynVox2)	✓			15.46	1.295	0.196	0.640
	w/o CL (Real+SALT)	✓			15.43	1.287	0.194	0.635
	w/ 1-stage CL (Real only)	✓			15.39	1.401	0.203	0.664
	w/ 2-stage CL (Real only)	✓	✓		15.89	1.424	0.205	0.664
	w/ 3-stage CL (Real only)	✓	✓	✓	16.01	1.420	0.205	0.664
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	**16.20**	**1.470**	0.206	**0.692**
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	**16.20**	1.466	**0.207**	0.683
BLSTM	w/o CL (Real only)	✓			12.34	0.852	0.157	0.495
	w/o CL (Real+SynVox2)	✓			12.53	0.903	0.164	0.501
	w/o CL (Real+SALT)	✓			12.54	0.885	0.161	0.491
	w/ 1-stage CL (Real only)	✓			11.73	0.794	0.149	0.473
	w/ 2-stage CL (Real only)	✓	✓		12.50	0.922	0.162	0.520
	w/ 3-stage CL (Real only)	✓	✓	✓	12.65	0.932	0.165	0.531
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	**13.07**	**0.997**	**0.171**	**0.554**
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	13.06	0.986	0.170	0.544
SpeakerBeam	w/o CL (Real only)	✓			11.62	0.804	0.150	0.489
	w/o CL (Real+SynVox2)	✓			12.01	0.915	0.156	0.503
	w/o CL (Real+SALT)	✓			12.07	0.887	0.153	0.501
	w/ 1-stage CL (Real only)	✓			11.17	0.753	0.144	0.473
	w/ 2-stage CL (Real only)	✓	✓		11.76	0.869	0.159	0.512
	w/ 3-stage CL (Real only)	✓	✓	✓	11.88	0.899	0.163	0.519
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	**12.35**	**0.974**	**0.171**	**0.547**
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	12.34	0.959	0.168	0.533
Voicefilter	w/o CL (Real only)	✓			11.92	0.818	0.158	0.453
	w/o CL (Real+SynVox2)	✓			11.41	0.747	0.149	0.402
	w/o CL (Real+SALT)	✓			11.46	0.728	0.147	0.398
	w/ 1-stage CL (Real only)	✓			11.34	0.709	0.149	0.414
	w/ 2-stage CL (Real only)	✓	✓		12.10	0.847	0.161	0.463
	w/ 3-stage CL (Real only)	✓	✓	✓	12.15	0.856	0.164	0.468
	w/ 3-stage CL (Real + SynVox2)	✓	✓	✓	**12.39**	**0.927**	**0.169**	**0.493**
	w/ 3-stage CL (Real + SALT)	✓	✓	✓	12.26	0.908	0.166	0.476

The results indicate consistent performance improvements across all models when implementing CL strategies. Conformer showed 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 surpassed the control condition using only real data (16.01 dB), confirming that the benefits stem 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 indicate 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 demonstrate 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 show 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 showed proportionally larger relative gains by incorprating CL. BLSTM achieved a 0.73 dB improvement (from 12.34 to 13.07 dB), while SpeakerBeam showed 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 demonstrated 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) yielded comparable improvements, likely due to their shared use of the HiFi-GAN vocoder architecture. These results collectively indicate the effectiveness of combining synthetic data augmentation with CL for enhancing TSE system performance.

The experimental results also reveal improvements across perceptual metrics. While intelligibility metrics showed modest improvements, with iSTOI increasing only marginally from 0.198 to 0.207 for Conformer with SALT data, quality-related metrics exhibited substantially more significant gains. The iPESQ improvements were particularly notable, with Conformer showing 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 pattern indicates that our approach predominantly enhances perceptual quality aspects of extracted speech rather than intelligibility. This suggest that the synthetic data integration through CL particularly addresses distortions and artifacts that affect perceived speech quality, which is often more critical in real-world applications.

8 Evaluation on Real-world Recordings

To further validate the advantages of TSE models trained on Libri2Vox in realistic environments, we conducted additional experiments using the ICASSP 2021 Deep Noise Suppression Challenge test set [36]. This test set contains 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 contains original reverberant speech samples, and our research does not focus on far-field speech and reverberation issues, we first removed 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 used DNSMOS since it does not require reference signals. In addition to the TSE models trained on the Libri2Vox dataset, we also used TSE models trained using the Libri2Talker dataset for comparisons.

Table 4 presents the DNSMOS results for different model configurations when tested on real-world recordings. The results of BLSTM clearly indicate that models trained solely on Libri2Talker perform poorly on real recordings, with a negative DNSMOS gain of -0.666. This poor performance persists even with noise augmentation (-0.421). In contrast, models trained with Libri2Vox achieved a substantial improvement with a DNSMOS gain of -0.081. This enhancement can be attributed to Libri2Vox's incorporation of real noisy speech from VoxCeleb2. Adding noise augmentation to Libri2Vox further improved performance (0.139), while our three-stage CL approach with SynVox2-generated synthetic speakers achieved the best results with a DNSMOS gain of 0.190. Similar trends were observed across all architectures (Conformer, SpeakerBeam, and VoiceFilter), with the combination of Libri2Vox and 3-stage CL with synthetic speakers consistently achieving the best performance.

9 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 were designed to isolate and quantify the individual impact of key components in our approach, providing deeper insights into how different elements contribute to overall system performance.

9.0.1 Comparison of Different Synthetic Data Ratios within Mini-batch

The goal of this experiment was to determine the optimal ratio of synthetic speakers within each mini-batch at Stage 3 of CL, finding the balance that maximizes performance while avoiding degradation from excessive synthetic

Table 4: 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**

data. The experiments involved the Stage 3 with SALT-generated synthetic speakers, following the configuration "w/ 3-stage CL (Real + SALT)" of Conformer shown in Table 3. As shown in Figure 5, optimal performance was achieved with synthetic speaker ratios of 0.2 and 0.5, both yielding an iSDR of 16.20 dB. These configuration outperformed the baseline configuration using only real data (ratio = 0.0), which achieved an iSDR of 16.01 dB. However, increasing the synthetic speaker ratio beyond these optimal values led to performance degradation. Note that when synthetic speakers comprised 90% of the training data, the iSDR decreased to 15.82 dB. Such degradation was furthered with the ratio being 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 is due to the significant mismatch between the synthetic training data and real test data. Therefore, it is necessary to train the model with real data first then incorporate synthetic data step by step.

9.0.2 Comparison of Different Amount of Synthetic Data

In the previous experiments, the number of synthetic speakers used for TSE model training was 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 be used to generate a

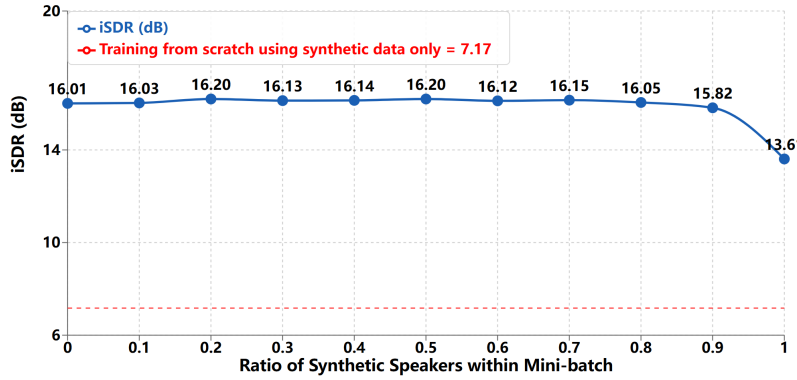

Figure 5: Impact of synthetic speaker ratio within one mini-batch at Stage 3 of configuration "w/ 3-stage CL (Real + SALT)" on Conformer. Red dashed line corresponds to performance (7.17 dB) where training started from scratch using synthetic data only.

different set of speakers by changing the parameter configuration.[4] Since synthetic data can be generated infinitely, we aim to 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 now exposes the model to several times more synthetic data than before. The goal of this study was thus to determine if increasing the number of synthetic speakers would improve the performance of BLSTM.

Two types of scenarios were considered: a scenario in which the number of speakers is increased in Stage 3, and another scenario in which Stage 4 is introduced after Stage 3 to increase the total number of synthetic speakers in Stages 3 and 4. They use consistent configurations for Stages 1 and 2 as specified in Table 3 under "w/ 2-stage CL (Real only)". Specifically, Stage 1 used data with cosine similarity below 0.5, and Stage 2 incorporated the complete dataset.

Table 5 presents the experimental results across different synthetic speaker configurations. For example, Experiment No. 8 involved using SALT for Stage 3 and additionally SynVox2 for Stage 4.

When both SALT-based and SynVox2-based synthetic datasets were used simultaneously in Stage 3 (No. 4), we observed an improvement to 13.16 dB, indicating the potential benefits of increased speaker diversity. We also observed incremental gains when adding SALT' and SALT" in Stage 3 (13.18 and 13.31 dB, respectively), and finally integrating all four datasets (13.36

[4]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: 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	-	13.06
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

dB).

However, the most significant performance gain emerged from CL training strategies. No. 8 and 9 implemented alternating synthetic datasets between Stage 3 and 4, achieving the highest iSDR (13.44 and 13.43 dB respectively). These CL training strategies are more effective than simultaneous dataset utilization (No. 7, 13.36 dB).

10 Ablation Study Regarding Datasets

Two ablation studies were conducted regarding the datasets. This analysis only focused on datasets, therefore synthetic speakers were not used.

10.1 Cross-dataset Evaluation

Using the existing Libri2Talker and WSJ0-2mix-extr and proposed Libri2Vox, we conducted a cross-dataset evaluation. We trained TSE models using one of the datasets with and one without noise argumentation and tests the models on test sets of the remaining datasets. On both Libri2Talker and Lirbri2Vox , the target speakers were selected from Librispeech/LibriTTS, but the interfering speakers are selected from different databases. Because these datasets have different characteristics, the model trained from each dataset is not well generalizable to the test set of the other datasets, and domain mismatch is likely to occur.

The experimental results are presented in Table 6. As expected, each model performed best when evaluated on the test set corresponding to its training dataset, with significant performance degradation observed in cross-dataset scenarios. When training on Libri2Vox, we observed optimal performance on the Libri2Vox test set, achieving an iSDR of 15.42 dB with the Con-

former model. However, this same configuration showed severe performance degradation (iSDR: -0.16 dB and 0.18 dB, respectively) when evaluated on the Libri2Talker or WSJ0-2mix-extr test sets. Similarly, models trained only on Libri2Talker exhibited degraded performance when tested on Libri2Vox or WSJ0-2mix-extr test sets. Models trained only on WSJ0-2mix-extr struggled significantly with other datasets, with the BLSTM model achieving only -1.72 dB on the Libri2Talker test set and 1.49 dB on the Libri2Vox test set. This pattern suggests a consistent limitation in the generalization capability of models trained on single datasets, regardless of which specific dataset is used. These results indicate that these datasets could be complementary to each other. This will be investigated in the next subsection.

Table 6: Experimental results of cross-dataset comparisons in iSDR(dB).

Model	**Libri2Talker**	**Libri2Vox**	**WSJ0-2mix**	**Noise aug.**	**Libri2Talker**	**Libri2Vox**	**WSJ0-2mix**
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

Noise augmentation appears most beneficial for models trained on WSJ0-2mix, likely because this dataset is relatively small (about 30 h) and entirely clean, making it more sensitive to additional variability. In contrast, both Libri2Talker and Libri2Vox already provide hundreds of hours of training data, and Libri2Vox further incorporates real acoustic variability from VoxCeleb2. As a result, noise augmentation yields only limited or mixed effects on these larger datasets, and in some cases may even be counterproductive.

10.2 Two-stage CL with Multiple Datasets

Finally, we demonstrated that the use of multiple datasets in CL can effectively reduce domain mismatch. Unlike previous experiments, the second stage of CL involved a training procedure that strategically uses multiple datasets simultaneously.

Table 7: Two-stage curriculum learning performance (iSDR in dB) on different test sets using various training dataset combinations. The first two columns show datasets used in Stage 1 and Stage 2 for training, and the last three columns show iSDR performance on each test set.

Stage 1	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

Table 7 presents performance comparisons across different two-stage CL configurations with the BLSTM model. When Libri2Vox was used in Stage 1 followed by multi-dataset training in Stage 2, we achieved the best overall cross-dataset performance (10.36 dB on Libri2Talker, 13.02 dB on Libri2Vox, and 10.55 dB on WSJ0-2mix-extr). This significantly outperformed other configurations, including when Libri2Talker or WSJ0-2mix-extr was used in Stage 1. Notably, despite having less data than Libri2Talker, Libri2Vox proved more effective as the foundation for CL, suggesting that its realistic acoustic variations provide a stronger initial representation that can be effectively refined with additional data sources.

Performance would probably be further improved if synthetic versions of Libri2talker and WSJ0-2mix-extr datasets are created and used in Stage 3 of CL, as shown in Table 3. However, this is beyond the scope of this paper and is a topic for future work.

11 Conclusion

In this paper, we first introduced Libri2Vox, a novel dataset designed to address the challenges of TSE in real-world acoustic environments. The dataset combines clean speech from LibriTTS with naturally noisy interference from VoxCeleb2, creating a diverse training environment with over 7,000 speakers. This approach bridges the gap between idealized synthetic mixtures and uncontrolled recordings by incorporating VoxCeleb2's natural acoustic variations, channel effects, and ambient conditions. Most notably, our evaluations on real-world recordings confirm that models trained on Libri2Vox achieve positive quality gains, while models trained on conventional artificial mixtures significantly underperform, demonstrating that the realistic acoustic variations in our dataset translate directly to improved performance in genuine real-world applications. We further enhanced the dataset's utility through synthetic data generation, developing two complementary methods, SynVox2 and SALT, to expand speaker diversity. Our experiments revealed that progressively adding more synthetic speakers with three-stage CL continues to yield performance

improvements, with sequential introduction of different synthetic speaker sets providing additional gains. This finding is particularly promising as synthetic data generation offers unlimited potential for further scaling speaker diversity beyond what is possible with real-world recordings alone. Unlike real recordings which are constrained by available speakers and recording conditions, synthetic speakers can be generated in virtually unlimited quantities, allowing for continuous expansion of training data diversity without the practical limitations of real-world data collection.

While synthetic speaker data offers valuable diversity for training TSE models, it is important to recognize the inherent limitations of using speech generative models. The useful distribution provided with these generative models is often constrained, and understanding which specific distributions are most beneficial for training remains a challenge. In some cases, only a small portion of the generated data may be truly useful for covering the necessary distribution. Generating more data beyond this point might not provide additional value, as certain types of data become redundant, offering no new information for the model [13]. The "new knowledge" provided with these speech generative models can quickly become repetitive, and the model may not need to repeatedly learn from similar data.

For future work, we will explore what types of synthetic data are most beneficial for TSE, potentially through data selection methods. For example, tracking the gradient changes of a network at each step could help determine whether the current data are useful for training. We will also explore using dataset distillation methods to generate synthetic data that represents real data characteristics. These synthetic data points could help us better understand the types of representations needed for TSE tasks.

A Details of Network Architecture

A.1 Architecture of SpeakerBeam

Input Process: The real and imaginary parts of the STFT are concatenated, resulting in a 512-dimensional input. The speaker embedding is extracted by processing the magnitude of the STFT of the reference speech through a fully connected layer then concatenated along the time dimension, resulting in an input of $T \times (256 + 256 + 192)$.

BLSTM Layers: This model uses 3 BLSTM layers, each with 512 hidden units per direction. The first layer processes the concatenated input.

Fully Connected Layers: Each BLSTM layer's output is passed through fully connected layers, reducing the output back to 512 dimensions.

A.2 Architecture of Voicefilter

Convolutional Layers: This model comprises 8 convolutional layers:

- Layers are zero-padded with kernel sizes varying from 1 to 7 and 64 output channels.
- Dilation factors increase from 1 to 16, and each layer is followed by batch normalization and ReLU activation.
- The final layer is a 2-channel convolution with a kernel size of 1.

LSTM Layer: A BLSTM with 400 hidden units per direction processes the concatenated convolution outputs and the 192-dimensional x-vector, which, similar to Conformer, is concatenated along the time dimension before being fed into the LSTM.

Fully Connected Layers: The first fully connected layer reduces the LSTM output to 600 dimensions. The second fully connected layer maps the result to 512 dimensions (for real and imaginary parts of the STFT).

References

[1] A. Alex, L. Wang, P. Gastaldo, and A. Cavallaro, "Data Augmentation for Speech Separation", *Speech Communication*, 152, 2023, 102949, DOI: 10.1016/j.specom.2023.05.009.

[2] M. Chen, Z. Wu, X. Wang, T. Lee, and H. Meng, "TTS-by-TTS 2: Data-selective augmentation for neural speech synthesis using ranking support vector machine with variational autoencoder", in *ICASSP 2020*, IEEE, 2020, 6699–703.

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing", *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 2022, 1505–18, DOI: 10.1109/JSTSP.2022.3188113.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition", in *INTERSPEECH 2018*, 2018, 1086–90, DOI: 10.21437/Interspeech.2018-1929.

[5] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation", *arXiv preprint arXiv:2005.11262*, 2020.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, 13(1), 1967, 21–7, DOI: 10.1109/TIT.1967.1053964.

[7] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification", in *INTERSPEECH 2020*, 2020, 3830–4, DOI: 10.21437/Interspeech.2020-2650.

[8] N. S. Detlefsen, J. Borovec, J. Schock, A. Harsh, T. Koker, L. D. Liello, D. Stancl, C. Quan, M. Grechkin, and W. Falcon, *TorchMetrics - Measuring Reproducibility in PyTorch*, version 1.2.0, February 2022, DOI: 10.21105/joss.04101, https://github.com/Lightning-AI/torchmetrics.

[9] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision", *NPJ Digital Medicine*, 4, 2021, DOI: 10.1038/s41746-020-00376-2.

[10] M. Fadaee, A. Bisazza, and C. Monz, "Data Augmentation for Low-Resource Neural Machine Translation", 2017, 567–73, DOI: 10.18653/v1/P17-2090.

[11] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: Unlocking Synthetic Data for Speech Recognition", in *INTERSPEECH 2021*, 2021, 899–903, DOI: 10.21437/Interspeech.2021-1882.

[12] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A Survey of Data Augmentation Approaches for NLP", in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, ed. C. Zong, F. Xia, W. Li, and R. Navigli, Online: Association for Computational Linguistics, August 2021, 968–88, DOI: 10.18653/v1/2021.findings-acl.84, https://aclanthology.org/2021.findings-acl.84/.

[13] Z. Gan and Y. Liu, "Towards a Theoretical Understanding of Synthetic Data in LLM Post-Training: A Reverse-Bottleneck Perspective", *arXiv preprint arXiv:2410.01720*, 2024, https://arxiv.org/abs/2410.01720.

[14] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete", LDC93S6A, 1993.

[15] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network", in *Proc. of INTERSPEECH*, 2020, 1406–10.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition", in *INTERSPEECH 2020*, 2020, 5036–40, DOI: 10.21437/Interspeech.2020-3015.

[17] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[18] T.-Y. Hu, M. Armandpour, A. Shrivastava, J.-H. R. Chang, H. Koppula, and O. Tuzel, "SYNT++: Utilizing Imperfect Synthetic Data to Improve Speech Recognition", in *ICASSP 2022*, 2021, 7682–6, https://api.semanticscholar.org/CorpusID:239615990.

[19] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, E. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models", in *Proceedings of the 41st International Conference on Machine Learning*, ed. R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Vol. 235, *Proceedings of Machine Learning Research*, PMLR, 21–27 Jul 2024, 22605–23, https://proceedings.mlr.press/v235/ju24b.html.

[20] K. Kasi, "Yet Another Algorithm for Pitch Tracking (YAAPT)", *MA thesis*, Old Dominion University, 2002, https://digitalcommons.odu.edu/ece_etds/388/.

[21] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis", in *Advances in Neural Information Processing Systems*, 2020.

[22] B. Li, X. Zhang, M. Liu, Y. Liu, Y. Gong, and J. Zhou, "Speaker Augmentation for Low Resource Speech Recognition", in *ICASSP 2018*, IEEE, 2018, 5044–8.

[23] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: multi-genre speaker recognition", *Speech Communication*, 137, 2022, 1–13, DOI: 10.1016/j.specom.2021.12.002.

[24] Y. Liu, X. Liu, X. Miao, and J. Yamagishi, "Target Speaker Extraction with Curriculum Learning", in *INTERSPEECH2024*, 2024, 4348–52, DOI: 10.21437/Interspeech.2024-1929.

[25] Y. Liu, X. Liu, and J. Yamagishi, "Improving curriculum learning for target speaker extraction with synthetic speakers", *2024 IEEE Spoken Language Technology Workshop*, December, December 2024, https://arxiv.org/abs/2410.00811.

[26] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, "Machine Learning for Synthetic Data Generation: A Review", *arXiv preprint arXiv:2302.04062*, 2023.

[27] Y. Lv, J. Yao, P. Chen, H. Zhou, H. Lu, and L. Xie, "SALT: Distinguishable Speaker Anonymization Through Latent Space Transformation", in *2023 IEEE Automatic Speech Recognition and Understanding (ASRU)*, 2023.

[28] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox, "What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation?", *International Journal of Computer Vision*, 126, 2018, 942–60, DOI: 10.1007/s11263-018-1082-6.

[29] X. Miao, X. Wang, E. Cooper, J. Yamagishi, N. Evans, M. Todisco, J.-F. Bonastre, and M. Rouvier, "Synvox2: Towards A Privacy-Friendly VoxCeleb2 Dataset", in *ICASSP 2024*, 2024, 11421–5, DOI: 10.1109/ICASSP48485.2024.10446513.

[30] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Speaker Anonymization Using Orthogonal Householder Neural Network", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2023, 3681–95, DOI: 10.1109/TASLP.2023.3313429.

[31] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset", in *INTERSPEECH 2017*, 2017, 2616–20, DOI: 10.21437/Interspeech.2017-950.

[32] S. I. Nikolenko, *Synthetic Data for Deep Learning*, Cham, Switzerland: Springer, 2021, ISBN: 978-3-030-75178-4, DOI: 10.1007/978-3-030-75178-4.

[33] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding", in *Proc. Interspeech*, 2018, 2252–6, DOI: 10.21437/Interspeech.2018-993.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books", in *ICASSP 2015*, IEEE, 2015, 5206–10, DOI: 10.1109/ICASSP.2015.7178964.

[35] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L. D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S. M. Mousavi, A. Nautsch, H. Nguyen, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperriere, M. Rouvier, R. D. Mori, and Y. Esteve, "Open-Source Conversational AI with SpeechBrain 1.0", 2024, arXiv: 2407.00463 [cs.LG], https://arxiv.org/abs/2407.00463.

[36] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge", in *ICASSP*, 2021.

[37] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors", in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6493–7.

[38] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech", in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021, https://arxiv.org/abs/2006.04558.

[39] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)A New Method for Speech Quality Assessment of Telephone Networks and Codecs", *ICASSP 2001*, 2, 2001, 749–52, DOI: 10.1109/ICASSP.2001.941023.

[40] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, *et al.*, "Effective Data Augmentation Methods for Neural Text-to-Speech Systems", in *ICASSP 2019*, IEEE, 2019, 6001–5.

[41] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum Learning: A Survey", *International Journal of Computer Vision*, 130(6), 2022, 1526–65, DOI: 10.1007/s11263-022-01611-x.

[42] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech", *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2011, 2125–36.

[43] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification", in *ICASSP 2014*, 2014, 4052–6, DOI: 10.1109/ICASSP.2014.6854363.

[44] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Z. Chen, S. Wang, Z. Liu, S. Ren, J. Liu, Z. Chen, Y. Q. Wu, J. Li, and F. Wei, "VALL-E: Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers", *arXiv preprint arXiv:2301.02111*, 2023.

[45] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive Margin Softmax for Face Verification", *IEEE Signal Processing Letters*, 25(7), 2018, 926–30, DOI: 10.1109/LSP.2018.2822810.

[46] K. Wang, J. Zhu, M. Ren, Z. Liu, S. Li, Z. Zhang, C. Zhang, X. Wu, Q. Zhan, Q. Liu, and Y. Wang, "A Survey on Data Synthesis and Augmentation for Large Language Models", 2024, arXiv: 2410.12896 [cs.CL], https://arxiv.org/abs/2410.12896.

[47] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking", in *INTERSPEECH 2019*, 2019, 2728–32, DOI: 10.21437/Interspeech.2019-1101.

[48] W. Wang, Z. Pan, X. Li, S. Wang, and H. Li, "Speech Separation with Pretrained Frontend to Minimize Domain Mismatch", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2024, 4184–98.

[49] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3), 2016, 483–92, DOI: 10.1109/TASLP.2015.2512042.

[50] C. Xu, W. Rao, E. S. Chng, and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2020, 1370–84.

[51] C. Xu, W. Rao, J. Wu, and H. Li, "Target Speaker Verification with Selective Auditory Attention for Single and Multi-talker Speech", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2021, 2696–709.

[52] Z. xu, M. Sach, J. Pirklbauer, and T. Fingscheidt, "Employing Real Training Data for Deep Noise Suppression", in *ICASSP 2024*, 2024, 10731–5, DOI: 10.1109/ICASSP48485.2024.10448333.

[53] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech", in *INTERSPEECH 2019*, 2019, 1526–30, DOI: 10.21437/Interspeech.2019-2441.

[54] X. Zhang, B. Li, M. Liu, Y. Liu, Y. Gong, and J. Zhou, "Overcoming Data Scarcity in Speaker Identification: Dataset Augmentation with Synthetic MFCCs via Character-level RNN", in *ICASSP 2018*, IEEE, 2018, 5049–53.

[55] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. ernocký, and D. Yu, "Neural Target Speech Extraction: An Overview", *IEEE Signal Processing Magazine*, 40(3), 2023, 101–14.

[56] K. molíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. ernocký, "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures", *IEEE Journal of Selected Topics in Signal Processing*, 13(4), 2019, 800–14, DOI: 10.1109/JSTSP.2019.2922820.