Archives of Acoustics Online First June 23, 2025 https://doi.org/10.24425/aoa.2025.154816

Research Paper

Benchmarking the First Realistic Dataset for Speech Separation

Rawad MELHEM*, Oumayma AL DAKKAK, Assef JAFAR

Higher Institute for Applied Sciences and Technology
Damascus, Syria

*Corresponding Author e-mail: rawad.melhem@hiast.edu.sy

Received February 4, 2025; revised May 22, 2025; accepted June 27, 2025;
published online July 23, 2025.

This paper presents a thorough benchmarking analysis of a recently introduced realistic dataset for speech separation tasks. The dataset contains audio mixtures that replicate real-life scenarios and is accompanied by ground truths, making it a valuable resource for researchers. Although the dataset construction methodology was recently disclosed, its benchmarking and detailed performance analysis have not yet been conducted. In this study, we evaluate the performance of four speech separation models using two distinct testing sets, ensuring a robust evaluation. Our findings underscore the dataset's efficacy to advance speech separation research within authentic environments. Furthermore, we propose a novel approach for assessing metrics in real-world speech separation systems, where ground truths are unavailable. This method aims to improve accuracy evaluations and refine models for practical applications. We make the dataset publicly available to encourage innovation and collaboration in the field.

Keywords: single-channel; speech separation; deep learning; corpus; datasets.

1. Introduction

Speech separation remains an active research area, with the primary challenge being the separation of speech mixtures in realistic environments. This field has numerous applications, including automatic speech recognition (ASR), speaker verification, automatic captioning for audio and video recordings, human-machine interaction, and hearing aid devices. Traditional methods, such as non-negative matrix factorization (NMF) (Ismael, Kadhim, 2024), complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) (Melhem *et al.*, 2024a), and independent component analysis (ICA) (Kariyappa *et al.*, 2023) have attempted to address these challenges, but their effectiveness has often been limited, particularly since they typically require prior knowledge of the speakers' data.

Recently, deep learning has significantly advanced speech separation techniques. Many studies have employed supervised learning with clean datasets containing ground truths; however, performance tends to decline in real-world scenarios. To enhance robustness, some researchers have explored training supervised models using noisy datasets that better reflect actual recordings. Additionally, there has been a shift toward unsupervised learning to improve separation accuracy when dealing with real mixtures.

Creating a realistic dataset with ground truths for speech separation is particularly challenging, as it is impossible to record the same utterance twice, once for a clean ground truth and then in a mixture with another speaker. In (Melhem *et al.*, 2024b), there was introduced the first realistic dataset for speech separation Realistic_TIMIT_2mix, which includes ground truths. We detailed the methodology for its construction and compared it with a synthetic dataset to demonstrate its effectiveness.

In this study, we extend our earlier efforts by conducting an analysis of the realistic dataset, examining its characteristics and evaluating its potential to enhance the efficacy of speech separation models and make it available online for public[1]. In addition,

[1] Avaiable at https://drive.google.com/drive/folders/1ViMQBN04ct0sKw66hSZytQIo89INahx-?usp=sharing.

we show how to enhance metric assessments in speech separation. The key contributions of our study include:

– analyzing the specifications of the realistic dataset;
– benchmarking the dataset against state-of-the-art deep learning models;
– publicly releasing the realistic dataset for broader use;
– introducing a novel methodology for evaluating speech separation metrics in real-world environments based on the construction of the realistic dataset.

The paper is organized as follows: Sec. 2 reviews related work on realistic datasets for speech separation; Sec. 3 provides a description of the dataset, Sec. 4 presents experiments for benchmarking the dataset; Sec. 5 discusses the obtained results, Sec. 6 explores how leveraging our dataset can enhance metric assessment in speech separation; and finally, Sec. 7 concludes the paper.

2. Related work

Most proposed solutions in speech separation have mainly employed supervised learning methodologies, as demonstrated in studies by Saijo *et al.* (2024) and Wang (2024). These approaches typically utilize a synthetic corpus for training, such as WSJ0_2mix dataset (Hershey *et al.*, 2016), which consists of clean, read speech in near-field conditions. While the results in idealized environments were commendable, the accuracy of these methods tends to degrade when confronted with more realistic scenarios. In response, some researchers have ventured into unsupervised techniques, directly addressing realistic mixtures to enhance separation accuracy in practical settings, as noted in (Wang, Watanabe, 2023; Han, Long, 2023). However, tackling realistic mixtures in an unsupervised manner presents significant challenges, necessitating extensive analyses of the auditory scene.

Numerous studies have sought to find effective solutions through supervised methods, leading to the development of specialized training sets. For instance, in (Wichern *et al.*, 2019; Maciejewski *et al.*, 2020; Cosentino *et al.*, 2020), the authors constructed a synthetic noisy datasets by separately recording noise signals and clean speech. These recordings were then mathematically combined to create noisy mixtures, ensuring that clean ground truths were available for the corresponding clean mixtures. Conversely, some researchers opted to capture speech signals in conjunction with noise, resulting in authentic noisy speech signals. This approach involved the addition of pairs of signals to generate noisy mixtures, as seen in datasets such as Mixer6 (Brandschain *et al.*, 2010) and VoxCeleb (Nagrani *et al.*, 2017). In these cases, the ground truths do not necessarily correspond to the mixtures, complicating the training process.

The CHiME datasets represent a comprehensive collection of speech data carefully designed for research in speech processing, particularly focusing on robust speech separation and recognition in challenging acoustic environments. These datasets include speech recordings in various noisy conditions, including background chatter, music, and meeting room sounds, thereby simulating real-world scenarios where speech processing systems may encounter difficulties. The CHiME datasets are available in multiple versions, including CHiME-3, CHiME-5 (Barker *et al.*, 2015; 2018), CHiME-7, CHiME-8 (Cornell *et al.*, 2023; 2024), and are frequently employed to evaluate algorithms designed to improve the performance of speech processing systems in noisy environments. Notably, the CHiME datasets feature multichannel recordings and lack ground truths, rendering them unsuitable for single-channel supervised learning approaches.

Subakan *et al.* (2021) introduced a realistic dataset called REAL-M for speech separation. This dataset comprises utterances collected from contributors who simultaneously read predefined sentences from the LibriSpeech dataset (Panayotov *et al.*, 2015) across various acoustic environments and utilizing different recording devices to replicate real-world conditions. Although REAL-M effectively captures authentic scenarios, it does not provide ground truths, which limits its applicability to unsupervised learning methodologies.

3. Dataset description

The dataset in this work is referred to as 'Realistic_TIMIT_2mix', named for its construction based on the TIMIT corpus. This process can be replicated with other corpora to generate similarly realistic datasets. Using a corpus such as TIMIT is essential for creating realistic audio mixtures. Ground truths are required for each mixture, and it is not feasible for a single person to record a sentence in isolation (to establish the ground truth) and then repeat it while another person speaks concurrently to create a mixture. The realistic dataset along with the codes utilized for recording and synchronizing the ground truth signals, is publicly accessible online under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

3.1. Dataset properties

The Realistic_TIMIT_2mix dataset is organized into four main folders: training, validation, testing, and ground_truths. The training folder contains 55 000 mixtures, totaling over 30 h of speech. The validation folder includes around 6000 mixtures, which is equivalent to approximately 5 h of speech, while the testing folder holds around 5000 mixtures, corresponding to nearly 3 h of speech. The mixtures are named according to a common format, such as 'speaker1-speaker2.wav'. Overall, the Realistic_TIMIT_2mix dataset comprises 66 000 mono-channel audio files, each sampled at 16 KHz with 16 bits per sample, resulting in a total size exceeding 3.5 GB.

3.2. Recording procedure

The primary tool utilized for building the Realistic_TIMIT_2mix dataset is the AudioPlayerRecorder (APR) MATLAB function. This function interfaces with the sound card to both play and record audio on the left and right channels simultaneously. It specifically necessitates the use of an ASIO driver for the sound device on Windows operating systems.

During the recording process for ground truths, APR is used for a single speaker to play and record their audio. However, for mixtures, both audio files corresponding to the speakers are played simultaneously, one on the left channel and the other on the right channel of the output audio device. These mixed audio streams are then recorded into a single WAV file. The detailed algorithm is elaborated in Algorithm 1 in (Melhem *et al.*, 2024b).

The recording process consumed a total of 45 working hours, making it more time-intensive compared to constructing a synthetic dataset. In synthetic datasets, mixtures are generated programmatically by sequentially combining samples from two speakers, a process that depends on PC configurations, and it can be accomplished in under 2 h. In contrast, for the Realistic_TIMIT_2mix dataset, each audio file must be played through the PC to allow the microphone to capture it, which contributes to the increased time investment.

The recordings were performed on a PC featuring the following hardware configuration: ASUS ROG STRIX Z390-F GAMING motherboard, Intel Core i9 CPU, 64 GB of RAM, NVIDIA GeForce RTX 2080 Ti GPU, and an ROG SupremeFX 8-Channel High-Definition Audio CODEC S1220A for audio processing. The distance between the microphone and audio output device was about 2 m, with the right and left channels placed 50 cm apart. The recording took place in a lab, far away from noise and disturbances.

4. Experiments

4.1. Dataset preprocessing

The TIMIT corpus consists of clean, read speech samples recorded at 16 KHz, featuring 630 speakers from eight distinct American English dialects. Each speaker utters ten unique sentences. During this experiment, the recording process was conducted in a quiet, noise-minimized environment. Within TIMIT dataset, each file is renamed to reflect the speaker's dialect, ID, and sentence type, as depicted in Fig. 1. This naming convention facilitates the creation of diverse audio mixes by selecting speakers of various genders, dialects, and speech patterns.

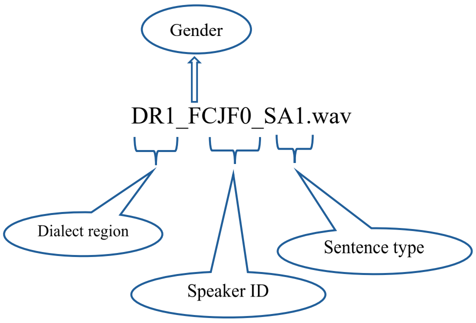

Fig. 1. File naming convention in the dataset.

4.2. Distortion and noise characterization

Each clean file within the TIMIT corpus undergoes playback via an output audio device, subsequently captured by a microphone, serving as the ground truth. As the clean file traverses a channel, comprising the output audio device, air, and microphone, distortions occur, due to attenuation and delay. To quantify these distortions, we compute the signal-to-noise ratio (SNR), where the signal is the original TIMIT file, and the noise is the difference between original file and the one captured by the microphone. Additionally, we compute the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) to measure the quality and intelligibility of the signals yielding an average metrics presented in Table 1. These metrics were computed after applying RMS scaling (Eqs. (1) and (2)) and cross-correlation time alignment.

Table 1. SNR of ground truths in the realistic dataset.

	SNR	PESQ	STOI
Ground truths	13.34 dB	2.06	0.56

To estimate the background noise, we recorded a 30-second silent segment in the lab environment and computed its root-mean-square (RMS) level, yielding a noise floor of −62 dBFS (decibels relative to Full Scale). The lab measures 10 m × 6 m × 4 m (L × W × H). During recordings, the space is unoccupied by people but contains tables, equipment, and curtains. Under test conditions, the lab exhibits no perceptible echo.

The total loudspeaker-to-microphone distortion was quantified by averaging the difference between the original and recorded signals across 4531 files (the number of ground truths) and computing the RMS of the averaged difference (RMS_Distortion), giving −40.6 dBFS. The procedure for calculating the RMS_Distortion is as follows:

- time alignment for each file pair using cross correlation;
- RMS-scaling for each recorded signal using the following formula:

$$\text{Scaling\_Factor} = \frac{\text{RMS(Original)}}{\text{RMS(Recorded)}}, \tag{1}$$

$$\text{Recorded} = \text{Recorded} * \text{Scaling}_{\text{Factor}}; \tag{2}$$

- compute the difference (distortion) signal for each file;
- calculate RMS of each distortion signal;
- average the RMS values across files;
- convert the RMS distortion to dBFS using the following formula:

$$\text{distortion (dBFS)} = 20\log_{10}\left(\frac{\text{RMS\_Distortion}}{\text{Full\_Scale}}\right), \tag{3}$$

where Full_Scale = 1 in MATLAB.

The time shift between the recorded and original signals of ground truths was determined through cross-correlation analysis, yielding an average shift of 1200 samples (approximately 75 ms at our sampling rate of 16 KHz).

For spectral analysis, we calculated the power spectral density (PSD) using a 1024-sample window (for high-frequency resolution) with 50 % overlap for each file pair (original and recorded signals). Before calculating the PSDs, we first applied RMS-scaling and time-alignment using cross correlation. The resulting spectrum, shown in Fig. 2, illustrates the distortion characteristics. The recorded and original signals match closely, except below 150 Hz, where microphone may increase low-frequency power. The recorded PSD shows higher baseline power across all frequencies because of the noise floor.

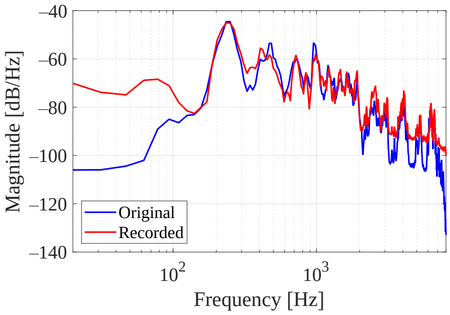

Fig. 2. PSD of the total distortion and the original signal.

The ground truths for Realistic_TIMIT_2mix must be clear and understandable. However, the primary focus is to ensure that each ground truth closely resembles itself in the mixed files. This requirement precludes using clean files as ground truths. In speech separation methods, the process depends on generating masks for each speaker and then applying these masks to the mixtures to isolate individual speakers. Any differences between the estimated masks and the corresponding ground truths can introduce distortions in the reconstructed speech of the separated speaker.

4.3. Deep learning models used

We chose four deep learning models for speech separation to benchmark the realistic dataset. TF-GridNet (Wang *et al.*, 2023), Sepformer (Subakan *et al.*, 2020), deep attractor network (DANet) (Zhuo *et al.*, 2017), and DANet with bidirectional gated recurrent units (DANet-BGRU) (Melhem *et al.*, 2021). The reasons behind choosing them is, first, to check the dataset on both modern and old versions of speech separation models, and second, to test the dataset on various architectures of deep models (e.g., transformer, long short-term memory (LSTM), and gated recurrent units (GRU)). For training Sepformer, we utilized the mixtures and ground truths in the temporal domain. However, for training TF-GridNet, DANet, and DANet-BGRU models, we employed the log magnitude of the spectrogram. The spectrogram was computed using a Hanning window of 32 ms length and an 8 ms hop size.

4.3.1. TF-GridNet

The architecture of this model consists of stacked blocks, each comprising three modules. The first module is an intra-frame spectral module, which is implemented as a single-layer bidirectional long short-term memory (BLSTM) network. The second module is a sub-band temporal module, which is also a single-layer BLSTM. The third module is a full-band self-attention mechanism implemented as a transformer, enabling the model to capture global sequence information (Wang *et al.*, 2023). For the input, TF-GridNet uses the log magnitude of the spectrum of the mixture.

4.3.2. Sepformer

It is a RNN-free model designed for speech separation. It consists of three components: the encoder, the masking network, and the decoder. The encoder consists of a single convolutional neural network (CNN) layer that processes the temporal mixtures, while the decoder utilizes a transposed convolution layer. The masking network is structured as a block that is iterated twice. Within this block, there are also two consecutive transformers: an intra-transformer, which captures relationships among samples within the same frame, and an inter-transformer, which captures relationships across frames (Subakan *et al.*, 2020). The parameters for the SepFormer architecture are provided in Table 2.

Table 2. Parameters for SepFormer architecture.

Parameter	Value
Number of convolutional filters	256
Kernel size	16
Number of attention heads	8
Chunk size	250
Number of DualPathBlocks	2
Number of transformers in each DualPathBlock	4

4.3.3. DANet

This model consists of four BLSTM layers followed by a fully connected layer that estimates masks for speakers (Zhuo et al., 2017). As input, the model takes the logarithm of the amplitude of the spectrogram. The BLSTM, a non-causal network, aims to utilize information from both past and future frames of the speech signal. To reconstruct the separated speech signals for each speaker, the phase of the mixture is combined with inverse short-time Fourier transform (ISTFT).

4.3.4. DANet-BGRU

It is similar to the previous model but it replaces the BLSTM layers with BGRU (Melhem et al., 2021). BGRUs are simpler and faster to train compared to BLTSMs. BGRU, with fewer gates compared to LSTM, streamlines the architecture, resulting in improved training efficiency and computational effectiveness.

All previous models were trained using the parameters presented in Table 3.

Table 3. Parameters for training AI models.

Parameter	Value
Learning rate	(10^{-5})
Batch size	8
Stop criterion	No improvement in validation error for three epochs
Optimizer	Adam
Normalization	z-score

The loss function for all models is the scale-invariant signal-to-distortion ratio (SI-SDR), defined as follows:

$$\begin{aligned} s_{\text{target}} &= \frac{\langle \hat{s}, s \rangle \, s}{\|s\|}, \\ e_{\text{noise}} &= \hat{s} - s_{\text{target}}, \\ \text{SI-SDR} &= 10\log\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2}. \end{aligned} \tag{4}$$

The metrics used for evaluation are: SI-SDR (Le Roux et al., 2019), PESQ (Rix et al., 2001), STOI (Taal et al., 2010). We evaluated the trained models on two distinct datasets. The first dataset, Realistic_TIMIT_test, was constructed similarly to Realistic_TIMIT_2mix but with files that were not part of the training set. The second testing set is the noisy Libri2Mix (Cosentino et al., 2020), which is a synthetic dataset.

5. Discussion

Figure 3 illustrates the effectiveness of using Realistic_TIMIT_2mix for training deep learning models. The models – TF-GridNet, Sepformer, DANet, and DANet-BGRU – all achieved convergence during training on Realistic_TIMIT_2mix. Specifically, Sepformer and TF-GridNet reached their minimum loss value around epoch 70, while DANet and DANet-BGRU did so around epoch 200. This discrepancy in convergence speed is attributed to the architectural differences: Sepformer and TF-GridNet leverage transformer blocks for parallel context feature extraction, whereas DANet and DANet-BGRU rely on recurrent neural network (RNN) architectures, such as LSTM and BGRU.

Notably, TF-GridNet exhibited less stable learning compared to Sepformer, likely due to fluctuations in the loss function. This instability may be linked to TF-GridNet's architecture, which includes two layers of BLSTM before the transformer, potentially complicating convergence.

Table 4 displays the evaluation results on the Realistic_TIMIT_test dataset, comprising realistic speech mixtures. Notably, Sepformer excelled across all evaluation metrics. With an SI-SDR of 9.57 dB, a challenging feature under realistic settings, alongside high PESQ (2.94) and STOI (0.66) scores, the quality of the separated speech was remarkably high. These results highlight the advantages of training models on data that closely mirrors real-world conditions, enhancing their performance. These findings underscore the effectiveness of Realistic_TIMIT_2mix in proficiently training deep learning models for speech separation.

Table 4. Performance of the learned models on the Realistic_TIMIT_test dataset.

Model	SI-SDR [dB]	PESQ	STOI
TF-Gridnet	9.23	2.87	0.64
Sepformer	9.57	2.94	0.66
DANet	8.62	2.02	0.51
DANet-GRU	8.66	2.09	0.52

The models were further assessed on the noisy Libri2Mix dataset, a synthetic collection with added noise, and the results are summarized in Table 5. Impressively, all models demonstrated strong performance, affirming the dataset's value for training purposes. Sepformer once again stood out, achieving superior results with an SI-SDR of 13.07 dB, indicative of highly accurate separation. The quality of the separated speech was notably high, as evidenced by high

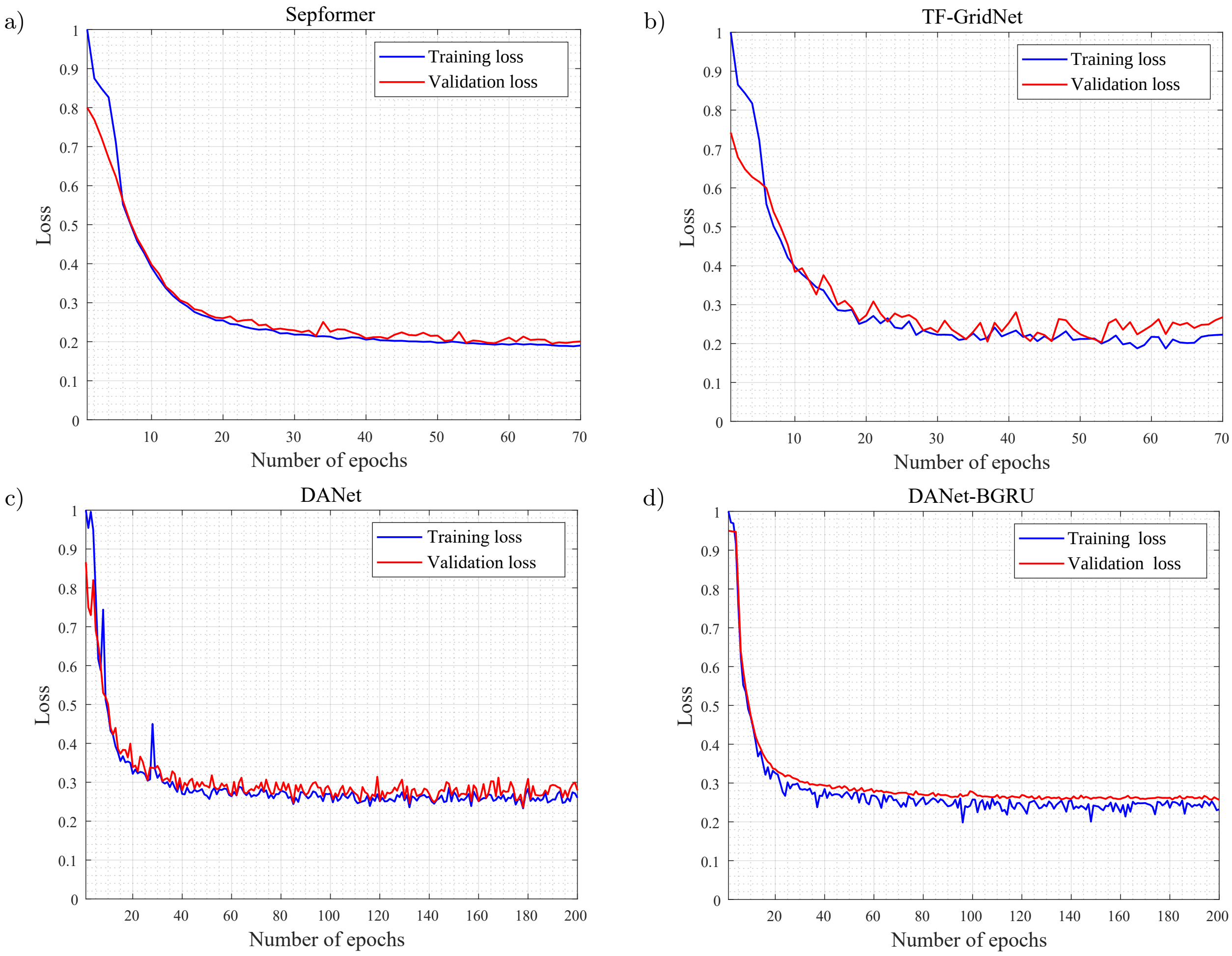

Fig. 3. Training and validation loss of the deep learning models.

Table 5. Performance of the learned models on the noisy Libri2Mix.

Model	SI-SDR [dB]	PESQ	STOI
TF-Gridnet	12.96	3.00	0.69
Sepformer	13.07	3.21	0.70
DANet	11.01	1.94	0.57
DANet-GRU	11.08	2.01	0.59

PESQ (3.21) and STOI (0.7) scores. These findings reinforce the notion that training models on realistic datasets enhances their ability to extract speech patterns from distorted, noisy mixtures.

6. Enhancing metric assessment with Realistic_TIMIT_2mix

Assessing the performance of trained models for separating speech in real-world mixtures presents a major challenge, as there are no reference signals for comparison. While synthetic datasets offer ground truth signals, they differ from real-world scenarios. Several researchers have approached this challenge by training neural networks to forecast SI-SNR by inputting speech mixtures into them (Subakan *et al.*, 2021), resulting in promising and predictive outcomes. Another alternative popular method for assessing the efficacy of speech separation models is the mean opinion score (MOS). However, MOS employment can be resource-intensive and time-consuming, as necessitates a specific demographic group with excellent hearing ability to hear separated signals and evaluate their quality.

To introduce a novel approach for measuring speech separation metrics in realistic environments, we propose leveraging the methodology employed in constructing the Realistic_TIMIT_2mix dataset. The process, illustrated in Fig. 4, consists of the following steps:

- record a speech signal for Speaker 1, saved as sp1.wav;
- record a speech signal for Speaker 2 saved asin sp2.wav;
- utilize the MATLAB function (APR) to play and independently record sp1.wav and sp2.wav, generating ground truths gt1 and gt2;
- use MATLAB function APR again to simultaneously play and record sp1 and sp2, producing a realistic speech mixture;

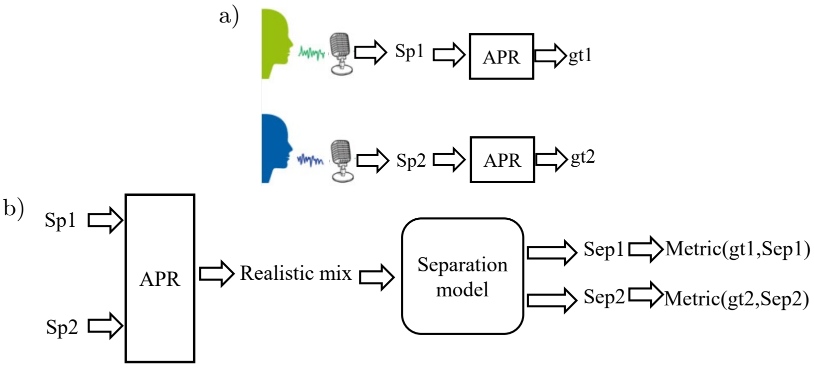

Fig. 4. Metric assessment algorithm with Realistic_TIMIT_2mix: a) for generating ground truths; b) for realistic speech mixture.

- apply the learned model (the one under evaluation) to separate the realistic mixture into Sep1 and Sep2;
- with the separated speakers and corresponding ground truths available, calculate desired metrics (such as SI-SNR, PESQ, STOI, SDR) to assess the model's performance.

7. Conclusion

This study presented an analysis and benchmarking of Realistic_TIMIT_2mix, the first realistic dataset designed for speech separation. We conducted an in-depth examination of its characteristics and quality, subsequently releasing it online along with accompanying codes. Furthermore, we trained four deep learning models on this dataset to demonstrate its effectiveness for model training.

Additionally, we proposed a novel approach to enhance the assessment of speech separation accuracy by leveraging the methodology employed in constructing Realistic_TIMIT_2mix. Moving forward, we intend to utilize this dataset in conjunction with robust input features to further enhance the performance of speech separation systems.

Fundings

This work was supported by the Higher Institute for Applied Sciences and Technology.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contribution

Oumayma Al Dakkak, as the main supervisor for the research, conceptualized the study, reviewed the first draft, and validated the findings. Rawad Melhem conducted the methodology and wrote the first draft. Assef Jafar contributed as a co-supervisor for the study. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors would like to thank the Department of Telecommunications at the Higher Institute for Applied Sciences and Technology, Damascus, Syria, for their full support.

References

1. Barker J., Marxer R., Vincent E., Watanabe S. (2015), The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, [in:] *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, https://doi.org/10.1109/ASRU.2015.7404837.
2. Barker J., Watanabe S., Vincent E., Trmal J. (2018), The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines, [in:] *Proceedings of Interspeech*, pp. 1561–1565, https://doi.org/10.21437/Interspeech.2018-1768.
3. Brandschain L., Graff D., Cieri C., Walker K., Caruso C., Neely A. (2010), Mixer 6, [in:] *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
4. Cornell S. *et al.* (2024), The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization, [in:] *8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*, https://doi.org/10.21437/CHiME.2024-1.
5. Cornell S. *et al.* (2023), The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios, [in:] *7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023)*, https://doi.org/10.21437/CHiME.2023-1.

6. Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. (2020), LibriMix: An open-source dataset for generalizable speech separation, arXiv, https://doi.org/10.48550/arXiv.2005.11262.

7. Han J., Long Y. (2023), Heterogeneous separation consistency training for adaptation of unsupervised speech separation, *EURASIP Journal on Audio, Speech, and Music Processing*, **2023**: 6, https://doi.org/10.1186/s13636-023-00273-y.

8. Hershey J.R., Chen Z., Le Roux J., Watanabe S. (2016), Deep clustering: Discriminative embeddings for segmentation and separation, [in:] *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, https://doi.org/10.1109/ICASSP.2016.7471631.

9. Ismael R.N., Kadhim H.M. (2024), NNMF with speaker clustering in a uniform filter-bank for blind speech separation, *Iraqi Journal for Electrical, Electronic Engineering*, **20**(1): 111–121, https://doi.org/10.37917/ijeee.20.1.12.

10. Kariyappa S. *et al.* (2023), Cocktail party attack: Breaking aggregation-based privacy in federated learning using independent component analysis, [in:] *Proceedings of the 40th International Conference on Machine Learning*, pp. 15884–15899.

11. Le Roux J., Wisdom S., Erdogan H., Hershey J.R. (2019), SDR – Half-baked or well done?, [in:] *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP.2019.8683855.

12. Maciejewski M., Wichern G., McQuinn E., Le Roux J. (2020), WHAMR!: Noisy and reverberant single-channel speech separation, [in:] *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP40776.2020.9053327.

13. Melhem R., Hamadeh R., Jafar A. (2024a), Study of the performance of CEEMDAN in underdetermined speech separation, arXiv, https://doi.org/10.48550/arXiv.2411.11312.

14. Melhem R., Jafar A., Dakkak O.A. (2024b), Towards solving cocktail-party: The first method to build a realistic dataset with ground truths for speech separation, *Romanian Journal of Acoustics and Vibration*, **20**(1): 103–113.

15. Melhem R., Jafar A., Hamadeh R. (2021), Improving deep attractor network by BGRU and GMM for speech separation, *Journal of Harbin Institute of Technology (New Series)*, **28**(3): 90–96, https://doi.org/10.11916/j.issn.1005-9113.2019044.

16. Nagrani A., Chung J.S., Zisserman A. (2017), VoxCeleb: A large-scale speaker identification dataset, [in:] *Proceedings of INTERSPEECH 2017*, pp. 2616–2620, https://doi.org/10.21437/Interspeech.2017-950.

17. Panayotov V., Chen G., Povey D., Khudanpur S. (2015), Librispeech: An ASR corpus based on public domain audio books, [in:] *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, https://doi.org/10.1109/ICASSP.2015.7178964.

18. Rix A. W., Beerends J.G., Hollier M.P., Hekstra A.P. (2001), 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*, pp. 749–752, https://doi.org/10.1109/ICASSP.2001.941023.

19. Saijo K., Wichern G., Germain F.G., Pan Z., Le Roux J. (2024), TF-Locoformer: Transformer with local modeling by convolution for speech separation and enhancement, [in:] *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, https://doi.org/10.1109/IWAENC61483.2024.10694313.

20. Subakan C., Ravanelli M., Cornell S., Bronzi M., Zhong J. (2020), Attention is all you need in speech separation, [in:] *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP39728.2021.9413901.

21. Subakan C., Ravanelli M., Cornell S., Grondin F. (2021), Real-M: Towards speech separation on real mixtures, [in:] *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP43922.2022.9746662.

22. Taal C.H., Hendriks R.C., Heusdens R., Jensen J. (2010), A short-time objective intelligibility measure for time-frequency weighted noisy speech, [in:] *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217, https://doi.org/10.1109/ICASSP.2010.5495701.

23. Wang Z.-Q. (2024), Mixture to mixture: Leveraging close-talk mixtures as weak-supervision for speech separation, [in:] *IEEE Signal Processing Letters*, **31**: 1715–1719, https://doi.org/10.1109/LSP.2024.3417284.

24. Wang Z.-Q., Cornell S., Choi S., Lee Y., Kim B.-Y., Watanabe S. (2023), TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation, [in:] *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP49357.2023.10094992.

25. Wang Z.-Q., Watanabe S. (2023), UNSSOR: Unsupervised neural speech separation by leveraging over-determined training mixtures, arXiv, https://doi.org/10.48550/arXiv.2305.20054.

26. Wichern G. *et al.* (2019), WHAM!: Extending speech separation to noisy environments, [in:] *Proceedings Interspeech 2019*, https://doi.org/10.21437/Interspeech.2019-2821.

27. Zhuo C., Luo Y., Mesgarani N. (2017), Deep attractor network for single-microphone speaker separation, [in:] *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/ICASSP.2017.7952155.