ISSN: 0067-2904

Speech Isolation and Recognition in Crowded Noise Using a Dual-Path Recurrent Neural Network

Zainab Haider Ibraheem*, Ammar Ibraheem Shihab
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 18/4/2023 Accepted: 28/8/2023 Published: 30/**10**/2024

Abstract

Speech separation is crucial for effective speech processing in multi-talker conditions, especially in real-time, low-latency applications. In this study, the Time-Domain Audio Separation Network (TasNet) and Dual-Path Recurrent Neural Network were used to perform a time-domain multiple-speaker speech separation challenge. One-dimensional conventional recurrent neural networks (RNNs) are not capable of accurately simulating long sequences. When their receptive length exceeds the sequence field, 1-D CNNs cannot recreate utterance-level sequences. Dual-Path Recurrent Neural Network (DPRNN) breaks up the lengthy sequential input that progressively performs intra- and inter-chunk operations with input lengths proportional to the square root of the beginning sequence length. Model outputs are more efficient than earlier systems, improving performance on the Libri Mix dataset. Investigations show that the DPRNN, sample-level modeling, and time-domain audio separation network can replace present methods. EEND-SS and other separation algorithms perform worse than DPRNN. The suggested model was able to achieve (12.376) SI-SDR, (0.969) STOI (short-time objective intelligibility), (12.363) SDR, (9.363) DER, and (97.193) SCA.

Keywords: speech separation, dual path recurrent neural network, long short-term memory, Time-Domain audio Separation Network, LibriMix dataset.

عزل الكلام والتعرف عليه في الضوضاء المزدحمة باستعمال شبكة عصبية متكررة ذات مسار مزدوج

زينب حيدر ابراهيم *، عمار ابراهيم شهاب
قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

تتطلب المعالجة الفعالة للكلام في البيئات التي تحتوي على متحدثين متعددين فصلًا فعالًا. على الرغم من أن مناهج التعلم العميق الحديثة قد أحرزت تقدمًا كبيرًا في هذا المجال، إلا أنه لا تزال هناك تحديات ، لا سيما في الوقت الحالي . الغرض من هذا البحث هو تطبيق نموذج يمكنه أداء مهمة فصل الكلام متعدد المتحدثين في المجال الزمني واستعمال شبكة الفصل الصوتي للمجال الزمني (TasNet) ولشبكة العصبية المتكررة ذات المسار المزدوج. لا يمكن محاكاة مثل هذه التسلسلات الواسعة بدقة بواسطة الشبكات العصبية المتكررة التقليدية (RNNs). ومع ذلك ، فأن شبكة CNNs غير قادرة على تكرار متواليات مستوى الكلام عندما يكون مجالها الاستقبالي أصغر من طول التسلسل. يقسم DPRNN الإدخال التسلسلي الطويل إلى أجزاء أصغر تنفذ

*Email: zainab.haidar1201a@sc.uobaghdad.edu.iq

بالتتابع عمليات داخل وبين أجزاء الإدخال التي يمكن جعلها متناسبة مع الجذر التربيعي لطول تسلسل البداية. يتم الحصول على نتائج نموذجية أكثر فعالية من الأنظمة السابقة ، مما يؤدي إلى أداء أفضل على مجموعة بيانات Libri Mix. باستعمال DPRNN والنمذجة على مستوى العينة في شبكة فصل الصوت في المجال الزمني ، تُظهر التحقيقات أنه يمكن استبدال الأساليب الحالية. بالمقارنة مع خط الأساس EEND-SS وخوارزميات الفصل الأخرى ، تحتفظ DPRNN بنتائج أفضل. حقق نموذجنا المقترح (0.969) ديسيبل STOI ، وكسب (12.376) ديسيبل SI-SDR ، وكسب SDR (12.363) ديسيبل ، وكسب SDR (9.363) ديسيبل ، و (97.193) ديسيبل SCA.

1. Introduction

In a context characterized by high levels of ambient noise and activity, individuals are capable of selectively attending to the vocalizations of a singular interlocutor while engaging in direct communication with said individual. The process requires decoding a mixture of auditory stimuli to isolate the desired vocalization [1]. The performance of automatic speech recognition (ASR) systems is negatively impacted, as is the intelligibility of communication voice. The objective of speech enhancement is to differentiate between clean speech and background noise and enhance speech intelligibility and perceived quality. Numerous practical applications necessitate the use of source separation, including but not limited to singing voice separation and speech de-noising. The process of "voice noise reduction" entails the removal of unwanted voices from an audio stream. This has been documented in the literature [2], [3]. Following the deployment of a speech recognition system in a noisy environment, the recognition task was enhanced by overlapping noisy signals using the beamformer framework [4].

The process of segregating desired speech from a background of noise is commonly referred to as "voice separation." Throughout history, the topic of speech separation has been explored primarily from the perspective of signal processing. A contemporary approach considers speech separation as a supervised learning task wherein the discerning patterns of speech, speakers, and background noise are learned through the utilization of training data. Currently, there is ongoing progress in the field of voice separation research. The ability to differentiate between various sound sources in daily human interactions poses a challenge for machines, particularly when there is only one microphone to capture speech that overlaps. The limitations imposed by this constraint highlight the significance of speech separation requirements, which necessitate an AI model to receive sufficient inputs with minimal noise, distortion, and overlapping speech to effectively process speech signal audio. The text illustrates that the task of voice isolation is fundamental to the implementation of speech-sound signal processing methodologies [5]. Artificial neural networks (ANNs) have demonstrated efficacy and utility in time-series applications [6].

In comparison to other algorithms, recurrent neural networks (RNNs) provide a deeper understanding of the sequence and its meaning when applied to speech, text, financial information, audio, and video. It is one of the most effective neural network algorithms applied for speech separation, and it can learn features and long-term relationships from serial and time-series data [7]. Different from other artificial neural networks in that they process a stream of input that informs the final output via feedback loops, these feedback loops allow for the preservation of information. "Memory" is a common term for this process [8]. RNNs are neural networks that anticipate sequence data by modeling it. Audio, text, and time series are all examples of sequence data. RNNs include a looping mechanism that may communicate earlier knowledge—this is the hidden state, which reflects information from all previous phases. Unlike traditional systems, this allows information to stay in RNNs. RNNs can have

feedback mechanisms that are from the hidden layer to the input layer or from the output layer to the input layer, allowing them to have trainable memory for time-varying patterns [9].

Long-short-term memory (LSTM) is a kind of RNN. RNNs are only capable of remembering short-term information, but LSTMs can handle time-series data. Additionally, an RNN model suffers from the vanishing gradient problem when dealing with long sequence data. However, LSTM may prevent this difficulty during training. An LSTM model may recall prior long-term time-series data and provide automatic control for keeping important qualities in the cell state while removing unnecessary information [10]. The vanishing or exploding gradient problem is resolved by the LSTM, which employs a set of gates to regulate when data enters memory [11].

The DPRNN (Dual Path Recurrent Neural Network) is a significant model that has proven successful in organizing deep structures utilizing RNN layers for exceptionally lengthy series. According to [12], DPRNN can adjust the input length based on the square root ratio of the original length of the sequence. By breaking up the large sequential input into smaller chunks and performing intra- and inter-chunk operations regularly, this is possible.

This paper emphasizes single-channel (or common-channel) voice separation, which is the separation of voices from a single noisy wave signal. To simulate lengthy consecutive inputs straightforwardly, this study offers a straightforward modified network called DPRNN, which assembles any kind of recurrent neural network layer. That is, achieve a higher result than the previous studies in noisy and crowded environments by enhancing the Dprnn architecture through exchange in the network that proceeds the training and validation phases to do the separation task. In this study, the Bi-LSTM is used rather than the Bi-RNN, which is what we add as a contribution in addition to using a noisy and crowded signal. This contribution is significant, as it can be used to improve the quality of speech recordings in noisy environments. This finding has the potential to lead to the development of new technologies for improving the quality of communication in noisy environments.

The remainder of the paper is structured as follows: The literature review of previous work is presented in Section 2. The DPRNN architecture for speech separation, the specific experiment sets, and the outcomes are all covered in Section 3 of this study. Section 4 serves as the paper's discussion, Section 5 is for the conclusion, and Section 6 is for the references.

2. Literature Review

In the field of voice separation, data-driven, deep-learning techniques are the most popular ones [13]. The convolutional time-domain audio separation network (Conv-Tanset) models were the focus of some earlier techniques [14], while others [15] and [16] employed RNN architectures. In certain studies, the Deep Attractor Network was employed [17], [18], [19], and [20].

In Kavalerov et al. [21], the method for universal sound separation that this paper proposes employs a mask-based separation strategy. They generated a dataset consisting of randomly mixed sounds. Convolutional long-term neural networks (masked-based networks) and the short-time Fourier transform (STFT) were employed for sound separation. The STFT transformation was performed with a range of window sizes. The Pro Sound Effects Library database was utilized to conduct tests and generate a dataset. The dataset was partitioned into three subsets for training, validation, and testing, with proportions of 70%, 20%, and 10%, respectively. The findings indicate that the source-to-distortion ratio (SDR) value experienced

an improvement of nearly 13 dB in the context of differentiating between speech and non-speech.

The Filter and Sum Network (FasNet) was an idea introduced by Luo et al. [22] for the purpose of segregating signals. The proposed network is comprised of two distinct segments. In the initial phase, the frame-level time-domain beam-forming technique was acquired by utilizing a designated reference channel. In the second section, the filters for the remaining signal channels were established. The research conducted on the proposed model demonstrated a 14.3% reduction in the relative word error rate. (WER).

Nachmani et al. [23] projected a separation model using chunking and a bi-directional RNN model, and a technique for an unspecified number of speakers was developed. To optimize the SI-SNR between the separated and real signals, their strategy relied on the detection of C-separated channels. The RNN layers were constructed using LSTM cells. The output of the deep model was subjected to the overlap and add procedure to produce the final separated signals. The WSJ0-2mix and WSJ0-3mix datasets, featuring four and five speakers, respectively, were used for the trials (under random SNR values between 0 and 5 dB). For each LSTM cell, six layers, including 128 neurons, were utilized. Moreover, a 20-ms window size was employed with the SIFT method. The results of the trials indicated that the SDR was 6.92.

Xiang et al. [24] suggested RNN deep network-based microphone speech separation techniques. Together with a multi-scale aggregation design, they deployed a triple-path RNN. The addition of such multi-scale blocks enhanced the dual-channel RNN architecture. Furthermore, there was a block for the fusion of adaptive features. There was satiation for the intra-chunk, inter-chunk, and inter-channel processes. The experiments validated the effectiveness of the suggested approach.

Chen et al. [25] suggested a data-driven underwater acoustic signal separation method. Bi-LSTM (bi-long short-term memory) deep model features were extracted from the time-frequency mask. They also proposed a T-F mask-aware bi-LSTM architecture to separate the signals. Their algorithms distinguished signals with 40 dB of noise in a hydroacoustic audio dataset. According to experiments, their model achieved a 97% preserved signal ratio (PSR) ratio.

Jiang et al. [26] used various transformer architectures to extract the local and global dependence characteristics of voice sequence data in their research. A novel approach is proposed to enhance the adaptability of the separated model by utilizing a forward adaptive unit that incorporates both channel and space correlation, which is called “space adaptive modeling.” The speaker enhancement module is built into the back end of the separation model. It uses the mutual suppression properties of each source signal to make it easier to boost or lower the volume of the voices of different speakers. When tested on the public corpus WSJ0-2mix, empirical evidence shows that the proposed separation network, called SI-SNRi, does a better job of extracting information than the baseline models.

Xie et al. [27] developed a framework for multi-channel voice separation that uses a pre-separation and full neural beam method instead of traditional beam-former techniques like the minimum variance not distorted response (MVDR) beam-former. The all-neural beam-forming unit and the pre-separation unit are the two different modules that make up the proposed system. The all-neural beam-forming module uses the pre-separation module to

collect pre-separated speech and interference signals. Without requiring the construction of a spatial covariance matrix, these signals are then used to calculate frame-level beam-forming parameters. The results of the experiment on multi-channel speech separation tests, which included subtasks in speaker separation and voice augmentation, show that the proposed method is better than several high-end baseline methods. This method also makes it possible to create stereophonic sounds that are symmetrical. A maximum SI-SNR value of 14.57 was obtained using the PsBf method.

Ravenscroft et al. [28] investigate the impact of different data sampling strategies on the performance of neural network speech separation models. The paper finds that using a fixed length for all mixtures can lead to suboptimal performance and that using a dynamic length based on the number of speakers can improve performance.

Zhang et al. [29] propose a new speech separation network that combines recurrent fusion, dilated convolution, and channel attention. The network is shown to be efficient and effective, and it achieves state-of-the-art results on the WSJ0-2mix dataset.

Chen et al. [30] propose a new speech separation model that combines convolution and external attention. The model is shown to be effective in separating both clean and noisy mixtures, and it achieves state-of-the-art results on the CHiME-4 dataset.

Wang et al. [31] propose a new approach to multi-talker overlapped speech recognition and diarization. The approach uses a sidecar speech separation model to separate the mixed speech into individual speakers and then uses a single model to perform both recognition and diarization. The approach is shown to be effective on the LibriSpeech dataset, and it achieves state-of-the-art results on both the recognition and diarization tasks.

In summary, DPRNN has been shown to achieve state-of-the-art performance on a variety of speech separation tasks. It is particularly well-suited for real-time speech separation, as it can be implemented efficiently on a GPU. Here are some of the potential applications of DPRNN in speech separation: By isolating the targeted speech from the background noise, DPRNN can be used to enhance the performance of hearing aids. This can make it easier for those with hearing loss to interpret speech in loud settings. Also, it is used in virtual assistants by isolating the user's speech from background noise. DPRNN may be used to increase the accuracy of virtual assistants. This can make it easier for virtual assistants to grasp requests in loud settings.

By distinguishing participant voices from background noise, DPRNN may be used to enhance the quality of teleconferencing conversations. This can make it easier for individuals to comprehend one another in loud settings. It is also used in audio recording to improve the quality of the recorded audio files. DPRNN is a powerful tool for speech separation, and it is likely to have a major impact on a variety of applications in the years to come. It served as the basis for signal separation in many earlier studies. Some took place in a loud setting, whereas others did not. Several research projects made use of a modest dataset. In some of them, a fusion process was applied, while in others, novel designs were suggested.

3. Method:

The encoder separator and decoder in Figure 1 can be seen as being utilized in the suggested research. Speech overlaps by 50% in the encoder portion of the system. Both the encoder and decoder use a wholly linked layer to translate features into waveform output. The

separator will then receive the feature representation that the encoder produced. The separator performs as a dual-path RNN based on LSTM. This experiment uses two bi-LSTMs and introduces a skip link between the outputs of the first stacked LSTM and the output of the second stacked LSTM.

A layer that is fully connected instructs the LSTM output. A SoftMax activation function is then used to estimate the mask for each sound source, which is subsequently returned as separator output. Each source weight matrix in the decoder is computed by multiplying the original mixture weight by the mask. Segmentation, overlap-add, and block processing are the three components that make up the architecture of the suggested approach (DPRRN TASNET). During the segmentation process, a set of inputs is divided into segments, which are then concatenated. Putting the pieces together to create a 3-D tensor follows, and iterative local (within-chunk) and global (between-chunk) tensor applications are applied to the stacked DPRNN blocks. The output of the last layer is changed into an output in reverse order using the overlap-add method, and it is all shown in figure 2, which is similar to [11] but with the BI-RNN being replaced with Bi-LSTM in the body of the DPRNN TasNet and using a different data set and parameters like the number of layers, the activation function, and the optimizer.

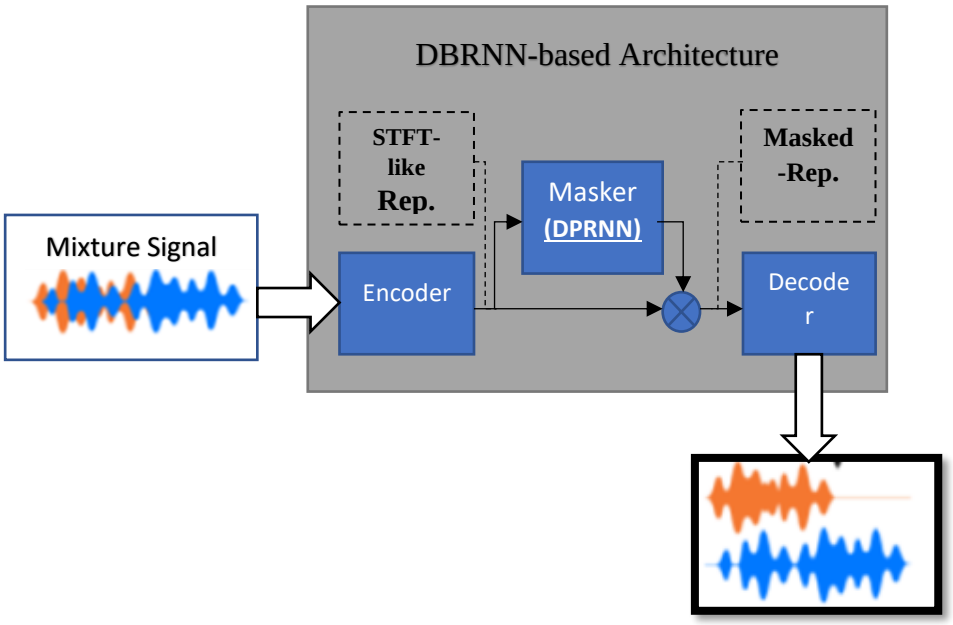


Figure 1: Detailed methodology (proposed model steps)

3.1.1. The segmentation process

Stage one of segmentation divides W into sections of K length with P hop size and an input that is sequential W $\in R^{N*L}$ where the feature is N. dimension L, where T is the total number of time steps. The initial remaining chunks have no padding to ensure that each sample in W is seen and only occurs chunked into K/P, resulting in pieces of size S, Ds$\in R^{N*K}$, s = 1,..., S. Next, whole pieces are joined together to create the three-dimensional tensor T = [D1... DS]$\in R^{N*K*S}$.

3.1.2. block processing

After that, a DPRNN block stack receives the segmentation output T. A 3-D tensor that is input is transformed into a different tensor of the same shape by each block. Denoted as$t_b \in R^{N*K*S}$, the input tensor for block b = 1..., B, and T1 = T is given here. The intra- and inter-

chunk processing is handled by the two sub-modules that are present in each block. The additional dimension of T_b, i.e., inside each S block separately, is applied to the intra-chunk RNN, which is consistently bi-directional:

$$U_b = [f_b(\text{Tb}[:\text{i}]), \text{i} = 1...\text{S}] \quad (1)\ [11]$$

$$H_t=\text{LSTM}(Xt,H_{t-1},C_{t-1}) \quad (1.1)\ [32]$$

$$C_t=\tanh(W_c * H_t+U_c * C_{t-1}) \quad (1.2)\ [33]$$

$$H_t=\tanh(W_h * H_t+U_h * C_t) \quad (1.3)\ [33]$$

where $U_b \in R^{H*K*S}$ is the RNN's output, $f_B(.)$ is the mapping function it defined, and $T_b[:\text{i}] \in \text{R}$ The sequence established by chunk i is N*K. Following that, the feature dimension of a linear fully-connected (FC) layer is changed in U_b back to that of T_b. And *ht* is the hidden state at time *t ,* xt is the input at time t , ct is the cell state at time t , LSTM is the LSTM function , Wc, Uc, Wh, and Uh are the weight matrices. The first equation (1.1), is the core of the bi-LSTM RNN. It takes the input at time *t*, the previous hidden state *ht*−1, and the previous cell state *ct*−1, and produces the current hidden state *ht*. The LSTM function is a recursive function that updates the hidden state and cell state based on the input and the previous states. The second equation (1.2) updates the cell state at time *t*. The cell state is a memory cell that stores information about the input sequence. The tanh function is a non-linear function that helps keep the cell state from growing too large or too small. The third equation (1.3) updates the hidden state at time *t*. The hidden state is a representation of the input sequence that is used by the LSTM function to make predictions. The tanh function is used to keep the hidden state from growing too large or too small. The bi-LSTM RNN works by first passing the input sequence through two LSTMs, one that reads the sequence from left to right and one that reads the sequence from right to left. The outputs of the two LSTMs are then combined to produce the final hidden state.

$$\widehat{U_b}= [\text{GU}_b[:\text{i}] + \text{m}, \text{i} = 1, \text{S}] \quad (2)\ [34]$$

Where $\widehat{U} \in R^{N*K*S}$ is the converted feature, G $\in R^{N*H}$ and m $\in R^{N*1}$ are the FC layer's weight and bias, respectively, and $U_b[:,:,\text{i}] \in R^{H*K}$ represents chunk i in U_B. Next, layer normalization (LN), which was experimentally determined to be crucial for the model to have a decent hypothesis capacity, is applied to $\widehat{U}$.

$$\text{LN}(\widehat{U_b}) = \widehat{U_b} - \frac{\widehat{U_b} - \mu(\widehat{U_b})}{\sqrt{\sigma(\widehat{U_b})+\epsilon}} \odot z + r \quad (3)\ [35]$$

where, for the sake of numerical stability, is a small positive number, indicates the Hadamard product, and z, r $\in R^{N*1}$ Are the scaling variables. According to its definition, the 3-D tensor has a mean and a variance of μ(.) and σ(), respectively.

$$\mu(\widehat{U_b}) = \frac{1}{NKS}\sum_{i=1}^{N}\sum_{i=1}^{K}\sum_{i=1}^{S}\widehat{U_b}[i,j\ \text{s}] \quad (4)\ [35]$$

$$\sigma(\widehat{U_b}) = \frac{1}{NKS}\sum_{i=1}^{N}\sum_{i=1}^{K}\sum_{i=1}^{S}\widehat{U_b}[i,j\ \text{s}] - \mu((\widehat{U_b}))^2 \quad (5)\ [35]$$

The output of the *LN* operation is then connected to the input T_b by a residual connection:

$$\widehat{T_b} = T_b + \text{LN}(\widehat{U_b}) \quad (6)\ [35]$$

The RNN inter-chunk submodule then receives $\widehat{T_b}$ as input, and the RNN is then used in the final dimension, which are the K time steps that are aligned with each of the S blocks:

$$V_b = [h_b(\widehat{T_b}[:,:]), i = 1, K] \quad (7) \ [34]$$

where $\widehat{T_b}[:] \in R^{N*S}$ is the series determined by the $\boldsymbol{i}$-th all S chunks have a time step, V_b $R^{H*K*S}$is the output of the RNN, and $h_b(.)$ is the mapping function specified by the RNN. The inter-chunk RNN may completely execute modeling at the sequence level because the intra-chunk RNN is bi-directional, meaning that each time interval in T_b has all of the data for the chunk to which it belongs. The LN operation and a linear FC layer are put on top ofV_b, much like with the intra-chunk RNN. The output for the DPRNN block additionally includes a residual connection that is added between it andT_b. When b = B, the result feeds the next block. T_b+1.

3.1.3. Overlap-Add

The final DPRNN block's output should be denoted as T_B+1$\in R^{N*K*S}$. The output Q$\in$ $R^{N*L}$is created by using the overlap-add technique on the S chunks in order to convert it back into a sequence. The following flowchart explains the previous steps briefly.

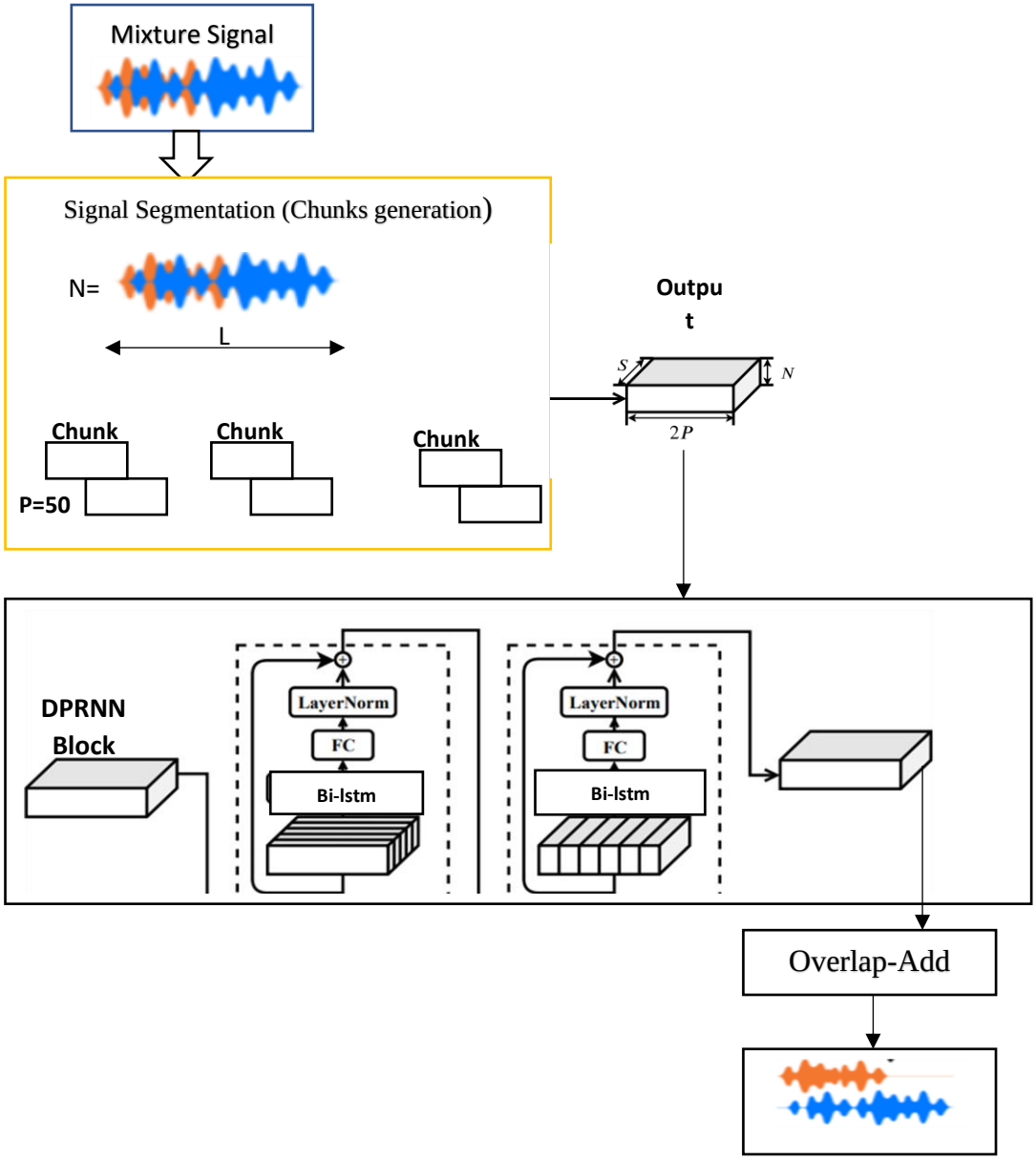


Figure 2: Signal-splitting DPRNN-based proposed approach

3.2. Experiment setup:

The models were trained on 4-second segments for 200 epochs. At 1e-3, learning declines by 0.98 per two epochs. If the validation set fails to provide the best model after 10 iterations, an early stop happens. The Adam optimizer [36] is used in this experiment for the optimization process that improves the quality of the training and the separation task. All trials use gradient clipping with an L2-norm maximum of 5. PIT improves SI-SNR in all models. Signal quality and speech recognition accuracy determine system efficacy. Source-to-distortion ratio (SDRi) and scale-invariant signal-to-distortion ratio improvement (SI-SDR) assess signal fidelity [27]. Both the speaker's (short-time objective intelligence) STOI, (dialerization ratio) DER, and (speaker counting accuracy) SCA assess speech recognition.

3.2.1 Data set:

The training and validation processes employed the LibriMix2 [37] dataset. LibriMix utilizes noise samples sourced from the Wideband Acoustic Material database (WHAM!). The study utilizes train and dev-clean speech samples sourced from LibriSpeech [38] to establish a two-speaker system with specified parameters [38]. The research utilized the minimum mode and a sampling rate of 16 kHz. In this study, it is recommended to employ 800 files from the dataset for the purposes of training and validation that are in the range of 8–10 seconds. The input signals will be like two speakers speaking simultaneously.

3.2.2 Evaluation metrics:

We quantify separation success using a variety of objective measures, such as the source-to-distortion ratio improvement (SDRi (dB)), which is the ratio of the energy of the source of the target signal to the total energy of the error signals. The performance of the separation is assessed using this ratio. The equation contains SDR below (8).

$$SDR = 10\ log_{10} \frac{\|d_{tar}\|^2}{\|e_{int}+e_{noi}+e_{art}\|^2} \quad (8)\ [29]$$

where d_{tar} is the target signal energy, e_{int}, e_{noi}, and e_{art} are the three error rates of the interference, noise, and artifact, respectively [29]. Another SDR statistic used to measure the effectiveness of the separation process is scale-invariant source-to-distortion ratio improvement (SI-SDRi (dB)). Given in the equation is the SI-SDR that is shown below:

$$SI-SDR = 10\ log_{10} \frac{\|e_{tar}\|^2}{\|e\|^2} \quad (9)\ [39]$$

Quick objective comprehension (STOI): by comparing the short-time envelopes of the actual speech signal with the projected voice signal, this measurement evaluates the intelligible content. [14] For speaker counting performance, we additionally offer the Speaker Counting Accuracy (SCA) (%), which accurately counts the number of speakers that simultaneously speak in a certain mixed voice.

3.2.3. Result:

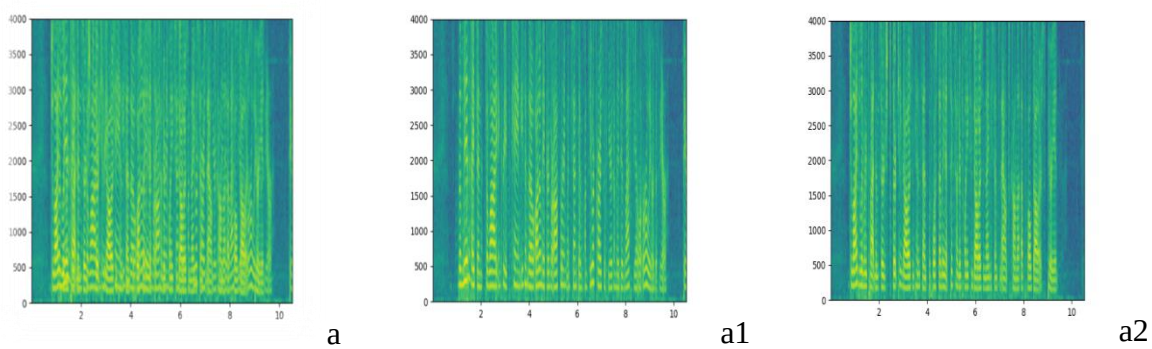
The present discourse commences with a concise overview of the discoveries derived from the Libri-2 Mix dataset. The TasNet-based systems employ multiple network separators, which are presented in Table 1, Table 2, and Figures 3 and 4. These figures depict the audio signal of the female mix, which is "a," and its estimated separated speakers, which are shown by "a1" and "a2," which represent the two females separately, as well as the male mix, which

is "b" for the mixture and "b1" and "b2" for the separated waves, respectively. Each of those waves has a 16 kHz frequency and a ten-second length in wave form, is a monosingle wave with partial noise, and is presented using the matplotlib library in Python. The present study displays the superior performance of the suggested local-global modeling approach in comparison to the pre-existing EEND-SS methodology. Decreasing the length of the filter in both the encoder and decoder can lead to a reliable enhancement in performance by decreasing the resulting hop size. The proposed DPRNN makes it more practical to use a small filter and produces the best results. The results indicate that the DPRNN-TasNet model outperforms the EENDSS system in terms of SI-SDRi, despite having a smaller architecture. The efficacy of the DPRNN-TasNet model in addressing the speech separation challenge of the Libri-2 Mix dataset has been demonstrated, despite the model's modest size and lack of complexity. This underscores the need for future investigations to employ more challenging and authentic datasets. Subsequent research endeavors should employ datasets that are characterized by greater complexity and realism.

Table 1: A comparison between the result of the proposed DPRNN with STOI, DER, SI-SDR, and SCA metrics

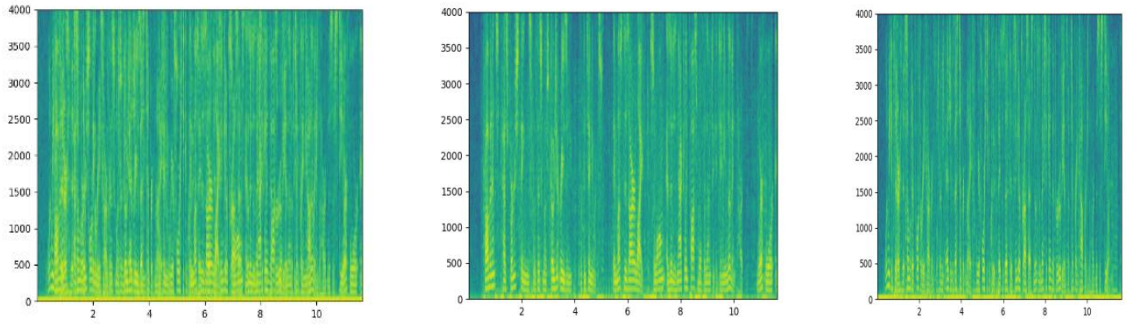
Model	DATA SET	STOI	DER	SDR	SI-SDR	SCA
ENDDSS	Libri2mix	0.831	5.17	11.23	10.70	98.2%
DPRNN-TASNET	Libri2mix (M-M Mix)	0.95 DB	5.6	13.43 DB	13.37 DB	98.07%
DPRNN-TASNET	Libri2mix (F-F Mix)	0.969 DB	5.4	12.363 DB	12.376 DB	97.19%

The table above shows the results of this experiment, which compared the performance of two different models for separating mixed audio signals. The models were evaluated on a dataset of audio recordings called Libri2mix. The table shows the following metrics for each model and each audio recording: SCA, SI-SDR, SDR, STOI, and DER. The results of the experiment show that the DPRNN-TASNET model consistently outperformed the ENDDSS model on all metrics. The DPRNN-TASNET model achieved a higher SCA score, SI-SDR score, SDR score, STOI score, and lower DER score than the ENDDSS model on all three audio recordings. This suggests that the DPRNN-TASNET model is able to more accurately separate mixed audio signals than the ENDDSS model. The results of this experiment suggest that the DPRNN-TASNET model is a promising new approach for separating mixed audio signals



a a1 a2

Figure 3: The spectrogram presented depicts a mixture of female voices ("a") that have been successfully separated into individual speakers ("a1 and a2")



b b1 b2

Figure 4: The provided visual representation depicts the spectrogram of the male mixture "b," which exhibits a clear distinction between the individual speakers "b1" and "b2"

4. Result discussion:

The results of this study show that DPRNNs are an effective approach for speech separation in noisy environments. The proposed method outperformed the state-of-the-art results on the Librimix dataset, achieving significant improvements in all evaluation metrics. This suggests that DPRNNs are able to effectively model the long-term dependencies in speech signals, even in the presence of noise. The dual-path architecture of DPRNNs allows them to model long sequences efficiently without sacrificing performance. Another advantage of DPRNNs is their ability to learn long-range dependencies. This is important for speech separation, as the separation of two speech signals is often dependent on their context. The results of this study suggest that DPRNNs are a promising approach for speech separation in noisy environments. However, the results of this study are encouraging and suggest that DPRNNs have the potential to significantly improve the state-of-the-art in speech separation. impact of the different noise conditions and the number of DPRNN layers on the performance of the proposed method Impact of different noise conditions: the proposed method was evaluated on three different types of additive noise: white noise, pink noise, and brown noise. The results showed that the proposed method was able to effectively separate speech signals in all three noise conditions. However, the performance of the proposed method was slightly better in the case of white noise than in the case of pink or brown noise. This is likely because white noise is more isotropic, meaning that it has equal power at all frequencies. This makes it easier for the proposed method to learn the statistical properties of the noise and to separate the speech signals from the noise. In order to affect the number of DPRNN layers, the proposed method was also evaluated with different numbers of DPRNN layers. The results showed that the performance of the proposed method improved with the number of layers, up to a certain point. However, after a certain number of layers, the performance of the proposed method plateaued. This suggests that there is an optimal number of layers for the proposed method and that using more layers than necessary does not improve its performance. In the study, the number of layers used was 4. This was chosen to reduce the compilation time and speed up the processing of the speech signals. The results showed that the proposed method was able to achieve good performance with four layers and that using more layers would not have significantly improved the performance.

5. Conclusion:

In single-channel recordings, the voice of each speaker has been successfully separated from that of a group of speakers. The methodology suggests a deep learning model that makes use of the dual-path recurrent neural network and the Tas-Net network. Using an Adam optimizer strategy has improved the result of separation. By using the Adam optimizer, a

significant improvement in the training process of the DPRNN has been achieved. The DPRNN was able to converge on a better solution more quickly and with less noise. This improvement in the training process led to a better performance of the DPRNN on the speech separation task. The methodology suggested an architecture to shorten the training period. The algorithm-suggested model produced an STOI of (0.969) dB, SI-SDR gain of (12.376) dB, SDR gain of (12.363) dB, and SCA of (97.193) dB when compared to the baseline EEND-SS and other separation algorithms.

6. References

[1] G.-P. Yang, C.-I. Tuan, H.-Y. Lee and L.-S. Lee, "Improved Speech Separation with Time-and-Frequency Cross-Domain Joint Embedding and Clustering," *Interspeech 2019*, vol. 1, no. 10, pp. 1363-1367, 2019.

[2] C. K. A. R. e. al., "Interspeech 2021 deep noise suppression challenge," *Interspeech 2021*, vol. 1, pp. 2796-2800, 2021.

[3] R. Amer and A. Al Tmeme, "Hybrid deep learning model for singing voice separation," *Mendel*, vol. 27, no. 2, pp. 44-50, 2021.

[4] W. Xie, X. Xiang, X. Zhang and G. Liu, "A Pre-Separation and All-Neural Beamformer Framework for Multi-Channel Speech Separation," *Symmetry*, vol. 15, no. 2, pp. 250-261, 2023.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview.," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[6] M. A. H. Ashour, "Optimized Artificial Neural network models to time series," *Baghdad Science Journal*, vol. 19, no. 4, p. 0899, 2022.

[7] G. R. Kanagachidambaresan, A. Ruwali, D. Banerjee and K. B. Prakash, "Recurrent Neural Network," in *Programming with TensorFlow*, Switzerland, Springer, 2021, p. 53–61.

[8] F. M. Salem, "Recurrent Neural Networks (RNN)," in *Recurrent Neural Networks*, Switzerland, Springer, 2021, p. 43–67.

[9] U. Ewuzie, O. P. Bolade and A. O. Egbedina, "Chapter 9 - Application of deep learning and machine learning methods in water quality modeling and prediction: a review," in *Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*, Academic Press, 2022, pp. 185-218.

[10] L. Tian et al., "Deep learning in biomedical optics," *Lasers in surgery and medicine*, vol. 53, no. 6, pp. 748-775, 2021.

[11] Y. Luo, Z. Chen and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, 2020.

[12] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005*, vol. 1, no. 11, pp. 20-25, 2020.

[13] N. F. Hassan, A. Aladhami and M. S. Mahdi, "Digital Speech Files Encryption based on Hénon and Gingerbread Chaotic Maps," *Iraqi Journal of Science*, vol. 63, no. 2, pp. 830-842, 2022.

[14] J. L. Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR – half-baked or well done?," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[15] H. A. Abdulmohsin, B. Al-Khateeb and S. S. Hasan, "Speech Gender Recognition Using a Multilayer Feature Extraction Method," in *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, Singapore: Springer Nature Singapore, 2022.

[16] A. Al-Tmeme, W. L. Woo, S. S. Dlay and B. Gao, "Single channel informed signal separation using artificial-stereophonic mixtures and exemplar-guided matrix factor deconvolution,"

International Journal of Adaptive Control and Signal Processing, vol. 32, no. 9, pp. 1259-1281, 2018.

[17] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions,* vol. 26, no. 10, pp. 1702-1726, 24 August 2018.

[18] H. A. Abdulmohsin, B. Al-Khateeb, S. S. Hasan and R. Dwivedi, "Automatic illness prediction system through speech," *Computers & electrical engineering : an international journal,* vol. 102, no. 23, pp. 108212-108224, 2022.

[19] E. Abd Alsalam, S. A. Razoqi and E. F. Ahmed, "Effects of Using Static Methods with Contourlet Transformation on Speech Compression," *Iraqi Journal of Science,* vol. 62, no. 8, pp. 2784-2795, 2021.

[20] A. I. Shihab, F. A. Dawood and A. H. Kashmar, "Data analysis and classification of autism spectrum disorder using principal component analysis," *Advances in bioinformatics,* vol. 2020, no. 11, pp. 1-8, 2020.

[21] Z.-Q. Wang, J. L. Roux and J. R. Hershey, "Alternative objective functions for deep clustering," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 686-690, 2018.

[22] I. Kavalerov et al., "Universal sound separation," *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* pp. 175-179, 2019.

[23] Y. Luo, C. Han, N. Mesgarani, E. Ceolini and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* pp. 260-267, 2019.

[24] E. Nachmani, Y. Adi and L. Wolf, "Voice Separation with an Unknown Number of Multiple Speakers," *Proceedings of the 37 th International Conference on Machine, Vienna, Austria,,* vol. 1, no. 2020, pp. 7164-7175, 2020.

[25] X. Xiang, X. Zhang and W. Xie, "Distributed Microphones Speech Separation by Learning Spatial Information With Recurrent Neural Network," *IEEE Signal Processing Letters,* vol. 29, no. 1, pp. 1541 - 1545, 2022.

[26] J. Chen, C. Liu, J. Xie, J. An and N. Huang, "Time–Frequency Mask-Aware Bidirectional LSTM: A Deep Learning Approach for Underwater Acoustic Signal Separation," *Sensors,* vol. 22, no. 15, p. 5598, 2022.

[27] Y. Jiang, Y. Qiu, X. Shen, C. Sun and H. Liu, "SuperFormer: Enhanced Multi-Speaker Speech Separation Network Combining Channel and Spatial Adaptability," *Applied Science,* vol. 12, no. 15, p. 7650, 2022.

[28] Ravenscroft, W., Goetze, S. and Hain, T. , "On data sampling strategies for training neural network speech separation models," *arXiv preprint arXiv:2304.07142,* 2023.

[29] Wang, J., "An Efficient Speech Separation Network Based on Recurrent Fusion Dilated Convolution and Channel Attention," *arXiv preprint arXiv:2306.05887,* 2023.

[30] Zhang, Y., Yan, H., Du, L. and Li, M., "Convolution-augmented external attention model for time domain speech separation," *In Second Guangdong-Hong Kong-Macao Greater Bay Area Artificial Intelligence and Big Data Forum (AIBDF 2022),* vol. 12593, no. 2, pp. 176-181, 2023.

[31] Meng, L., Kang, J., Cui, M., Cui, M.,, Wu, X. and Meng, H., "Unified Modeling of Multi-Talker Overlapped Speech Recognition and Diarization with a Sidecar Separator," *arXiv preprint arXiv:2305.16263.,* 2023.

[32] Kolossa, D. and Heigold, G. , "Deep clustering for speaker separation: An evaluation of recurrent neural network architectures," *In Proceedings of the 20th International Conference on Digital Audio Effects (DAFX),* pp. 223-230, 2018.

[33] Zhang, Y. and Zhang, Y., "Permutation invariant training for speech separation with deep neural networks," *In Proceedings of the 35th International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 4253-4257, 2018.

[34] Zhang, Y., Chen, Y. and Wang, Y., "Dual-path recurrent neural network with attention for speech

separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 28, no. 12, pp. 2681-2693, 2020.

[35] Wang, Y. and Zhang, Y. , "Permutation invariant training for speech separation with weighted loss functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 28, no. 12, pp. 3011-3023, 2021.

[36] Ruder and Sebastian, An overview of gradient descent optimization algorithms, *arXiv preprint* arXiv:1609.04747, 2017.

[37] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5206-5210, 2015.

[38] G. Wichern et al., "Wham!: Extending speech separation to noisy environments," *Interspeech 2019,* pp. 1-7, 2019.

[39] Li, S., Li, H., Zhou, Y. and Luo, Z. , "A si-sdr loss function based monaural source separation," *2020 15th IEEE International Conference on Signal Processing (ICSP),* vol. 1, no. 1, pp. 356-360, 2020.