# Literature Review

The sound source separation tasks focus on the separation of a particular intended sound separating the speech from the rest of sounds or separating the noises from the environment to have full cognitive processing of a specific sound. Recent research has been conducted for sound separation in deep learning networks by using multiple transformer layers which necessitate huge data, and computational capacity such as DPTNet [1], Sepformer [2] etc. Numerous neuroscience studies have suggested that in solving the cocktail party problem, the brain relies on a cognitive process called top-down attention which modulates cortical sensory responses to different sensory information. SuDORM-R [3] and A-FRCNN [4], encoder-decoder speech separation networks, neglect cortical areas. Existing DNN-based methods are categorized into two major groups: time-frequency domain and time domain. Recently, DPTNet and Sepformer have replaced LSTM with transformer layers to avoid performance degradation due to the long-term dependencies and process in parallel. However, this led to computational costs due to the large number of parameters and additional transformer layers. The encoder-decoder speech separation model SuDORM-RF achieved a certain extent of trade-off between performance and complexity. Development of RCNN model A-FRCNN obtained competitive results though it has room for improvement. Some earlier models uses top-down attention models only uses the attention layer to the top layer of multilayer LSTM and ignore lower layers which could not take full advantage of top-down attention associates large gap in performance and complexity including SuDORM-RF and A-FRCNN. The TDANet [5] works in three layers it embeds the features with the corresponding mask into individual target speech, and extracts the multiscale features at different temporal resolutions. The encoder in the bottom-up connections implemented downsampling layers consisting of a 1-D convolutional layer followed by GLN. Taking the multiscale features we use the average pooling layers and compress the features into different temporal dimensions to reduce the computational cost and these features are fused into the Global feature. The global feature is then fed into the transformer layer adds the positional information of the feature frame in it and then calculates the dot product with each attention layer in the multi-head attention layers which direct the model to focus on different aspects of information, finally the output with the applied residual connection are normalized. These normalized values sent to the second step of global attention signals where the normalized compressed values are upsampled to maintain coherence with the dimension of the input feature and the sigmoid function extracts the attention which is the achieved multiscale semantic information are elicited which improves the quality of separated audios. In comparison to other models, TDANet stands ahead in terms of training and inference time with less complexity and less number of parameters, multiply-accumulate complexity(MAC), and environmental requirement relative to other models such as Sepformer, A-FRCNN-16, DPTNet, Dual-PathRNN [6] etc. However, it has some limitations as it only reflects certain working principles of the auditory system and is not a faithful model of the auditory system. For example, it requires a bottom-up encoder and a top-down decoder, and it is unclear how such components are implemented in the brain. Model leverages the semantic information of a sound classifier for universal sound separation which aims to separate acoustic sources from an open domain regardless of their class. It conditions the separation network based on the semantic embedding extracted from a sound classifier. This approach is useful in the case of an iterative setup where the initial separation stage and their corresponding classifier-derived embedding are used in the following networks afterward. One major challenge in this task is to deal with the vast number of classes encountered in the world. This experimentation adopted different ways of conditioning a sound source separation network and reported SOTA for universal sound source separation. The model for separation was trained on Audioset dataset and, the model MobileNet-style architecture network gives a conditional embedding of $V \in \mathbb{R}^{F*C}$ where C is the number of classes. They used three types of semantic representations mixture embedding, all embedding and soft-OR embedding. For source separation, they used a time-dilated convolutional network(TDCN++) model and its iterative version iTDCN++. The TDCN++ model consists of analysis and synthesis basis transforms which encode and decode the signal, respectively, as well as a masking-based separation module that consists of stacked blocks of dilated convolutions and dense layers. In the iterative version, the process of estimating source signal is repeated twice. Both separation networks are trained using permutation-invariant negative signal-to-noise ratio (SNR). The embeddings are then integrated with the signal which requires at first resampling the signal to match the dimension with the embedding. The model uses various methods of extracting embeddings of only mixtures, embedding for both sources and the mixture and concatenating, using fine-tuned embedding to reduce the risk of losing significant information, fine-tuned embeddings for the iterative model of the mixtures as well as the clean sources. Lastly, the guided fine-tuned embeddings for the iterative model use a sigmoid cross entropy loss to guide the classifiers towards producing more similar representations where the oracle embeddings extracted from the clean sources of the training set are used as targets. The total loss in this case is calculated by adding the cross-entropy loss of each classifier in the iterative session. Another study, Band Split Recurrent Neural Network(BSRNN), is a frequency domain model proposed working with the spectrogram explicitly split into subbands and performing interleaved band level and sequence level modeling. It uses different modules to complete the full structure of the model starting from the band split module to the mask estimation module. The band split module splits the spectrograms into subbands with predefined bandwidth from expertise and then conducts a summation of the real and imaginary value for each subband followed by normalization and merges to a 3-D tensor. The merged tensor then goes through two different residual RNN layers: the sequence-level RNN and band-level RNN which go to the BLSTM followed by a fully connected layer. The main motivation for BSRNN [7] is to explicitly split the frequency components into subband features and use a sequential-order-aware module to capture the inter-band dependencies. BSRNN doesn't perform strict frequency-bin-level modeling as prior works on automatic speech recognition and full-band speech enhancement to save the model complexity, memory footprint and improve the processing speed. It also described a semi-supervised data sampling pipeline for fine-tuning the model trained on a small-scale labeled dataset on a large-scaled unlabeled dataset which followed an iterative process of creating new pseudo labels from the unlabeled data and adding the pseudo-labeled data to the existing data. It uses dataset MUSDB18-HQ [8] uses an threshold to segment the salient data. This model uses different bandwidth splits at various frequencies and found that lower frequency bands are important for the model to successfully estimate more accurate spectrograms also a small-scale grid search to determine the band split of bass, drum, and other tracks. There are two metrics have been used are uSDR corresponds utterance-level signal-to-distortion ratio and cSDR or chunk-level SDR and found that is resulted better output relative contemporary models in identifying sources of music. With a diverse set of research aspects in audio source separation it was found that there was no model or introduction of supervised general audio source separation which is source agnostic and trained on large datasets at scale without prior knowledge about the sources. The earlier studies or models were trained on source-specific tasks, or universal source separation works predominantly focused on separating mixes similar to field recordings with mostly sound events like dog barking or alarms. General Audio Source Separation(GASS) [9] is the first study method that does not require any prior information or assumptions about what specific sounds might be present in the recording. It introduces the transferability of the models, fine-tunes or modifies different already established models e.g., TDANet-Wav [5], TDANet-STFT, BSRNN [7], etc and trains the models for 10 million steps using the Adam optimizer with a batch size of 10 and a cyclical learning rate between $10^{-7}$ and $10^{-4}$ spanning 400k steps per cycle and minimized the logarithmic-MSE loss with a threshold T set to -30 dB. The trained models are propagated to different test or evaluation procedures and it appeared that in general, the fine-tuned models perform well relative to the non-tuned, or model tuned from scratch which reveals the transferability of the models. Models are trained. The in-distribution results show that the models can separate an unknown number of sources from a variate set of mixes that include speech, music, and sound events. Among the out-of-distribution results, the no-tuning models achieved competitive performance for sound event and speech separation, but we also noted that our models had challenges in generalizing to separate cinematic and music mixes. All Fine-tuned models (except the music separation one) obtain state-of-the-art results in their respective benchmarks. Data sets were made by combining from different public and licensed sources which are then further processed for ensuring compatibility. To prepare the data the compatibility concerning number of channels, number of sources between range 1 - 4, was resampled and created a uniform distribution of period. However, recent source separation was mainly motivated by other audio processing tasks or research fields such as speech enhancement and speech recognition while the intrinsic characteristics of much were not fully discovered. Another dimension of working in audio is working with audio events which include the classification of audio events. This classification method used the Mel-Spectrogram Separation with CNN which manipulates the signal's spectrogram in logarithm scale and attempts to identity the audio events recorded using unknown devices, this study attempts to reduce the distance of model performance on known data and unknown data, where it is seen that the model trained on known data perform relatively well due to its familiarity with the pattern of the spectrogram and the environmental and other complexities. Various devices convert or record the same input with identical frequencies their characteristics do not remain same which results in better performance on the acquainted features and lower performance on the unknown devices. The log mel-spectrogram separation algorithm proposed in this study is based on the method of McDonnell et al, who won second place in the DCASE 2019 Task 1B. The proposed CNN has three major changes from the VGGNet-based CNN [10]. The first is using an average value pooling layer instead of a max value pooling layer with a change in the stride size from 1 to 2. The global average value pooling layer uses the average value pooling of the first layer instead of the dense layer of three layers used in existing CNN. The CNN takes input as

(40, 40, 1) dictates the dimension of the spectrogram which passes through the lambda function split the input into two separate mel-spectrogram of size (40,20), divides into low and high-frequency bands and trains separately. Followed by the batch normalization and convolutional layers and at each layer prosecution further doubles the number of filter sizes which started at 18 initially. Different open-source datasets were used namely UrbanSound8K, BBC Sound FX, Google audio set etc. The datasets were folded into 5 groups and four folds among them were rerecorded using devices such as Samsung, iPhone SE, LG V50, and Google Pixel, split into validation and test datasets. Ultimately, the results of the five folds were averaged, and appeared that the performances of the devices improved from the base model in each fold. Among the different baseline methods without embedding and proposed methods with embeddings, the iterative models that use the embeddings perform best. Recent study, the Dual-Path Mamba (DPMamba) [11] replaces the transformer architecture for audio source separation with much efficiency and with less infrastructural requirement. The architecture of this model is very simple and easy to implement associated with greater performance for handling sequential data using state space model (SSM) mamba. Mamba models have matched and even surpassed transformers of comparable sizes in sequence modeling tasks of not only audio but also image, and genomics. Relatively smaller sized this model performed and compared with other models trained on WSJ0-2mix data. The dual path separation model Mamba incorporates an input-dependent selection mechanism sequence modeling performance but still enjoys linear complexity concerning the sequence length. The long sequence modeling method in dual-path RNN splits a long speech into multiple short chunks and applies Mamba models within each chunk, across all chunks, in the original direction, and in the reversed direction of time. The encoded waveform, in MambaMaskNet, a dual path network operates on three-dimensional data, chunks the data frames and iteratively processes the dataset with DP blocks, within each DP block, four SSMs process the features in four different ways: intra-chunk forward, intra-chunk backward, inter-chunk forward, and inter-chunk backward. Comprises a stack of R dual-path (DP) blocks and backward SSM processes in the opposite direction of the sequence. Both intra-chunk and inter-chunk units contain a normalization layer, a bidirectional Mamba (BiMamba), and a skip connection. It first processes the intra-chunk unit, processes frames within chunks, and then goes across chunks individually. Lastly, perform the forward and backward SSM in parallel by their own convolution followed by the sigmoid linear unit. Three different size models either meet or surpass the performance of existing CNN, RNN, and transformer models of similar or large sizes. Enhancing the efficiency of the Mamba separation model and improving the performance by integrating Mamba with other network layers is the direction of future works provided by the study. Mossformer2 is a hybrid model having the capabilities to model both long-range, coarse-scale dependencies and fine-scale recurrent patterns by integrating a recurrent module into the MossFormer framework. The recurrent module is based on a feedforward sequential memory network (FSMN), which is considered "RNN-free" recurrent network due to the ability to capture recurrent patterns without using recurrent connections. The dilated FSMN block in this study comprises by using gated convolutional units (GCU) and dense connections. In addition, a bottleneck layer and an output layer are also added to control information flow in the module. The reason for the addition of a recurrent module with earlier Mossformer so that the model can better handle the finer scale recurrent patterns in sequential information. Speech signals inherently exhibit recurrent patterns that manifest in phonetic structures, prosody, and semantic associations, all of which play a significant role in speech separation. The recurrent module models intricate temporal dependencies within speech signals. The dilated FSMN block in this model comprised of a feed-forward (FFN) and a memory layer. The FFN enhances the memory layer by employing stacked two-dimensional dilated convolutional blocks establish interconnections within the memory layer. To further improve the dilated FSMN this model allows for a decreased embedding dimensionality and uses convolutional units in place of linear units, implemented through a bottleneck layer followed by a PRelU activation and layerNorm layer. The bottleneck output follows the GCU layer and gives the ultimate results going through the output layer consisting of simple layer norm and 1*1 convolution. Mossformer2 [12] provides better output than Mossformer [13] and other state-of-the-art models on diverse benchmarks.

This review covers seven different articles for with comprehensive covering diverse aspects of audio separation including models that work only for isolating individual speech from a mixture of speeches or models that strive to perform on diverse sets of audio domains with generalized version of a single model, models crisscrossed with different datasets, difference in version where each version perform best over one another in specific datasets, some models are genuinely motivated by the biological functioning of the human auditory system, one uses the semantic information conveyed from other models. The information of the time domain over the periods, its changes, tracking the prosodic features, pitch, intonation, stress, rhythm, tempo, pauses, and loudness are the fundamentals or key for soundtracking, seeing the movement and constituents, derivatives of these features dependent or independent of the sound signal. The mathematics of the features are very important contain the variate information of sound features. The studies cover models uses the transformer architecture, for holding the sequential information of the signal. Researchers are trying to fetch some better models replacing the transformer architecture as it is computationally expensive and inefficient. Models like TDANet, Mamba, and Mossformer successfully replaces the transformer architecture with alternative methods and are still under development. Speech separation techniques are broadly categorized into two main approaches based on the input types: the time-frequency domain approach and the time-domain approach. The time domain input separation models are now the mainstream of research and the models from a high overview generally seen contain three parts: the encoder, separator, and decoder. The more challenging task in sound separation is the development or improvement of all hearing general source separation models that could separate the audio or sound regardless of the type or source of the sound, means a single out-of-distribution solution for all types of sound and even noise.

| Recent Research | Achievements | Limitations | Potential Gaps |
|---|---|---|---|
| Encoder decoder speech separation network with top-down global attention and local attention, inspired by brain's top-down attention for solving cocktail party problem | Reduced model complexity, parameters and inference speed, and significant performance improvement as well | Not a faithful model of the auditory system reflect certain working principle of auditory system, comparison with a black box | The functionality of auditory system, the biological simulation of brain sound processing and interpretation chain, is diminutively inclusive. |
| Leverage the semantic information of a sound classifier for universal sound separation | Utilization of conditioning information from other framework, guided optimization, training without ground truth class labels | Semantic embeddings weaker than mask in STFT, two times of iteration, reliability on a single metrics, dependency on other model for information, Non-democratic terminologies should have explained | Title Universal sound separation, can not be represented by separating only two sources |
| Music source separation using frequency-domain splits spectrogram into sub bands and performed interleaved band-level and sequence-level modeling | Discovery of intrinsic characteristics and patterns of the music signals, performance improvement through semi-supervised fine tuning | Prior expertise or knowledge required about the characteristics of the target source in most important band splitting module, focuses only on high-sample-rate source separation, transformer architecture could have approached for better capture long-range dependencies | Limited to model development purpose, did not serve its other purpose as discovery of intrinsic characteristics, it may not covered the subject experience of sound, timbre, diversification required in the band split schemes. |
| Single general audio source separation (GASS) model trained to separate speech, music, and sound events in a supervised fashion with a large scale dataset | First supervised general source separation trained on large-scale data, fine tuning or prospect of transfer learning performed have made it SOTA | Music and sound event detection faced challenges, criss-crossed combination of datasets and models | A single promising model yet not developed, fine tuning on a new use cases doesn't make the model generalized, ultimately source specific training have conducted, it have more to do with music and cinematic separation to work out. |

| | | | |
|---|---|---|---|
| Using log mel-spectrogram separation for audio event classification with unknown devices | Reduced parameters or model size, alleviate the performance difference between known and unknown devices significantly | Inefficient for larger large scale datasets, imbalanced and small datasets | This model can classify only 16 types of signal |
| Dual-path Mamba, replacing transformer models short-term and long-term forward and backward dependency of speech signals using selective state spaces reduces inefficiency of computation and memory | Surpassing transformers of comparable sizes in sequence modeling tasks, less memory, energy, computation and resources required, less data, easy architecture, first work that adopts Mamba in single channel speech separation, holds future prospects for better improvement | One dataset, single domain don't know how this model perform in music, sound event separation whereas other models is struggling for universal separation | Existing research for speech seperation has already provided better output reducing the complexity and computation of transformer. TDANet's MAC only 5 percent of Sepformer and DPT-Net, inference time is only 10 percent, large-size version 10 and 25 percent, this paper didn't mention inference time. |
| Using recurrent module based on a FSMN, which is considered rnn-free recurrent network due to the ability to capture finer scale recurrent patterns without using recurrent connections | Improvement in performance, better capturing of fine scale patterns, enhancement of attention mechanism and reliable combination | Increased size of parameters, increase in RTF, increased complexity | The enhancement of capabilities in modeling finer-scale recurrent patterns should have quantified with a detailed discussion of the impact on different factors affecting the recurrent patterns or patterns itself. |

# References

[1] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[2] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 21–25.

[3] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP).* IEEE, 2020, pp. 1–6.

[4] X. Hu, K. Li, W. Zhang, Y. Luo, J.-M. Lemercier, and T. Gerkmann, "Speech separation using an asynchronous fully recurrent convolutional neural network," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 509–22 522, 2021.

[5] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," *arXiv preprint arXiv:2209.15200*, 2022.

[6] 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 (ICASSP).* IEEE, 2020, pp. 46–50.

[7] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[8] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-hq-an uncompressed version of musdb18," *(No Title)*, 2019.

[9] J. Pons, X. Liu, S. Pascual, and J. Serrà, "Gass: Generalizing audio source separation with large-scale data," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2024, pp. 546–550.

[10] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J.-S. Park, G.-J. Jang, and J.-H. Kim, "Convolutional neural network based audio event classification," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 6, pp. 2748–2760, 2018.

[11] X. Jiang, C. Han, and N. Mesgarani, "Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation," *arXiv preprint arXiv:2403.18257*, 2024.

[12] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Q. Yip, D. Ng, and B. Ma, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2024, pp. 10 356–10 360.

[13] S. Zhao and B. Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.