TACIT: **T**CN **A**nd **C**onformer **I**ntegrated **T**wo-stream model for Background Music Removal

Abdus Saboor Gaffari, Latifa Alhosani and Solomon Ghebretatios
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates

Table of Contribution

Team member AUS ID	**Team member name**	**Team member's contribution to the project**	**Percent Effort (out of a total of 100%)**
b00105302	Abdus Saboor Gaffari	Coded the dataset generation script, proposed model, training script and evaluation script. Report writing: Dataset selection, dataset generation and other parts. Train the models and evaluate them. Recorded the training and evaluation video.	33.33%
g00104304	Latifa Alhosani	Coded the UI using Gradio. Report writing: description of the problem, previous approaches, and other parts. Recorded part of the video and edited it.	33.33%
b00104390	Solomon Ghebretatios	Coded the scripts for training and evaluating the baselines. Report writing: Architecture explanation and selection, results and discussion, and other parts	33.33%

I. Description of the Problem

With the rise of platforms like YouTube, Instagram, and Twitch, there has been a huge increase in user-generated video content. A lot of this content includes background music. The problem is that automated content detection systems are usually triggered when copyrighted music appears in videos without the appropriate licensing. As a result of this, Videos may get demonetized, muted, or even removed. Tools that can eliminate background music alone without altering other audio, such as speech or ambient sounds, are desperately needed by video makers, particularly those who are filming in public settings where it is difficult to avoid it. It's not only about avoiding copyright issues. Having clean speech without music interference can also help with tasks like speech recognition, emotion detection, and making content more accessible to people with hearing impairments [1].

Although some commercial tools provide audio separation, existing methods have three main challenges:

Balancing Music Removal and Audio Realism: Most traditional tools not only strip out music. They frequently also eliminate other background noises, which in turn can make the audio sound too clean, which is completely unnatural. This is an issue when you want to maintain that natural feel of the environment, like in a Vlog or documentary [1].

Complexity of Mixed Audio Environments: In real life, audio includes more than just speech and music. It might have traffic, birds, people talking, or other ambient sounds. On top of that, music often overlaps with speech, especially when it includes vocals. This overlap makes it difficult to separate them accurately, especially when working with monaural audio that doesn't provide spatial cues [1].

Presence of Vocals in Music Tracks: A lot of music used in online or public content has vocals, which makes the separation task harder. Since singing voices and speech overlap quite a bit in frequency, most models struggle to tell them apart. This often leads to either some of the music vocals leaking into the speech output or, in worse cases, parts of the actual spoken dialogue getting cut out by mistake.

Generalization and Scalability: A lot of deep learning systems separate the entire source and then attempt to retain only the non-musical portions. As a result, the procedure becomes more intricate and resource-intensive. Our ideal method would not require a lot of manual procedures and would be able to remove only the audio while leaving everything else intact [1].

This task is made even harder by how diverse music is, with different genres, languages, recording conditions, and background noises all adding to the challenge. There are also very few large and annotated datasets to train models on. So, any solution must generalize across all of these factors, which is difficult even for today's most advanced deep learning models [1]. To address these challenges, we propose a lightweight time-domain architecture called **TACIT** (**T**CN **A**nd **C**onformer **I**ntegrated **T**wo-stream model) that can effectively separate background music, including vocals, from speech while preserving ambient (environmental) noise. TACIT is designed to work on monaural mixtures and leverages a hybrid combination of convolution and attention to handle overlapping sources.

II. Previous Approaches

The most closely related previous work [1], which examines deep learning music removal from monophonic audio mixtures. Unlike traditional speech enhancement techniques which attenuate every background noise as well as sound source separation which retains the non-music parts leaving a responsibility on the user to reconstruct an audio, their model aims to makes the sound of only music disappear which retains speech and environmental information. Their model is developed using Conv-TasNet which is a time-domain convolutional network originally developed for speech separation tasks. Their model is trained using combinations of mixed speech and music selected from a combination of LibriSpeech[2] and FMA[3] datasets at various signal-to-noise ratios (SNRs) with loss determined by root-mean-square error (RMSE) between predicted speech and clean speech signals. Their evaluation is measured using spectrogram visualizations and human listening tests that confirm the resulting mixtures contain reduced music parts and contain speech and noise [1]. There are areas of improvement in this work while it does provide encouragement in preserving an original audio effect and has shown it is possible to remove music in a targeted way in monophonic spaces but there are very few definable metrics of performance, such as signal-to-distortion ratio (SDR) or SNR improvements. The evaluation based upon visual and auditory assessments has yet to be quantitative. In addition, the performance of the model decreases when the speech is a background voice to the music suggesting difficulties discerning spoken words from lyrics. Also, there are no comparisons with other architectural options or established baselines in the work.

In addition, [4] demonstrate a Convolutional Denoising Autoencoder (CDAE) as a strong approach to the problem of music-corrupted speech for augmenting automatic speech recognition (ASR) in uncontrolled real-world environments. While methods have taken a traditional feature approach by hand crafting music features, their CDAE takes a data-driven approach by learning from labelled data by training on speech corrupted by varied genres of music, including violin, piano, symphony, and rap, with English (Aurora4) and Chinese (863) speech corpora. This data-driven approach offers CDAE the opportunity to learn and store common spectral and temporal patterns of music to separate and remove, while maintaining intelligibility of the speech. Quantitative experiments yielded dramatic improvements in ASR Word Error Rate (WER). For example, WER dropped from approximately 60% to below 24% with violin speech, and to under 10% with piano background. CDAE generalizes to both language and genre, making it ideally suited for live streams or mobile content. Additionally, CDAE outperformed more simplistic denoising or classical source separation approaches that considered music as generic noise. Building on the hybrid models, [5] talks about the twin path models namely the Dual-Path RNN (DPRNN) [6], Dual-Path Transformer Network (DPTNet) [7], and SepFormer [8] can be thought of as significant configurations in monaural source separation. The twin-path models take the encoded mix and segment it into overlapping chunks so that the local and global dependencies can be attended to with a RNN or transformer layers of two types that overcome the challenge of long sequences that are often present in voice and music mixtures. These utilize both models better and obtained best performance in class on canonical speech and music separation benchmarks. DPTNet, rather than employing the recurrent component, replaces the recurrent component with a non-recurrent (but still context aware) transformer to enhance parallelization, and learn about order, without explicit position encoding. SepFormer and MossFormer build on the approach further and introduce multi-head self-attention and convolutional modules, or some form of pairing of full sequence attention with lightweight recurrent or convoluted blocks. Both DPTNet and SepFormer have so far best performance metrics (SI-SNRi up to 24 dB on hard datasets) and given the dual-path models were shown to allow highly selective, perceptually aware separation of music from voice at a level of noise that was robust to high variability and real-world noise.

In keeping with this direction, in this study [9], SE-Conformer presents an entirely new architecture applying the Conformer model specifically designed for speech recognition to denoising and speech enhancement (in fact, music can be a strong interference consistent with the noise sources). SE-Conformer combines multi-layer encoder/decoder convolutional architectures with the Conformer blocks (which are convolutions and self-attention similar to the Transformer architecture) to best model and recover both short-term, and long-term dependencies typically contained in raw time-domain audio. This paradigm is crucial in dealing with highly nonstationary noise context and with more predictable periodic variability such as music to manage better. Objective evaluations (PESQ, STOI) confirm the performance of the model compared to contemporaneous frequency-domain and time-domain baselines in realistic speech enhancement scenarios, in more realistic conditions. Most importantly, the effectiveness of its parallel layer structures combined with its natural time-domain will make SE-Conformer an attractive model to remove background music and other types of interference from video streams and user-generated content, where the considerations of likely real-time performance, and the realism of the audio is critical.

In order to further enhance monaural separation, [10] propose an advanced model, MossFormer which improves upon traditional approaches by using gated single-head transformer blocks with a self-attention paradigm that utilizes both local contextual attention and global contextual attention. This is a departure from dual-path transformer models which use local and extended (in chunks) attention in two separate paths. The key innovation of MossFormer is that it allows quadratic (local) self-attention and linearized (global) self-attention to be integrated directly, so global attention is simply modeled directly on the entire sequence instead of in subsets or classically. This means that global dependencies can be monitored directly on the entire audio signal for phrase details while taking computational constraints into account. To augment its local feature extraction for easier learning of complementary overlapping speech and music, MossFormer uses position-wise convolutional modules in the attentive gating block. Extensive experimental evaluation on WSJ0-2/3mix[11] and WHAM![12]/WHAMR![13] shows that MossFormer not only outperformed existing state-of-the-art models, including SepFormer, but also achieved separation quality approaching the limits for SI-SDRi, sampled at 22.8 dB on WSJ0-2mix and 21.2 dB on WSJ0-3mix. Further experiments using ablation studies showed that both the convolutional modules and the unconventional triple-gating model were essential for models' performance overall, even with more complex or realistic audio mixtures that models were tested on. MossFormer model design capability for capturing global context as well as detailed local features in monaural mixtures hold promise for applications that include selective background music removal for user-generated content, which need speech and environmental realism preserved.

Apart from these deep learning models, [14] provide a way of extracting speech from mixed audio that has background music - an area of particular interest to the type of user-generated video content that platforms like YouTube and Instagram allow. In their article, Nakano et al. propose two methods of removing background music from single-channel audio with little or no distortion of the speech; vector quantization (VQ) and non-negative matrix factorization (NMF). In the study it shows that both VQ and NMF improve speech recognition performance in the presence of musical interference. However, the VQ method was especially impressive, even in testing at signal-to-noise ratios that are normally regarded as challenging for recognition, giving very high word recognition rates. In terms of factor based methods that suppress music but retain speech intelligibility, this research adds further insights while also touching on practical solutions that could help creators avoid copyright problems and improve speech clarity in their videos.

Recently, several generative models have been proposed [14] [15] in the literature for the audio source separation task, including for speech and music disentanglement. Classical models such as Non-negative Matrix Factorization (NMF) and Variational Autoencoders (VAEs) laid the groundwork, but now the field is shifting to a focus on more expressive models. Generative Adversarial Networks (GANs) have performed very well in the audio source separation domain. Because GANs do not require pre-specified output distributions, they have greater flexibility and expressiveness than NMF or VAEs. Simple GAN architectures, such as a multi-layer perceptron trained with Wasserstein-GAN loss, have demonstrated better Source-to-Distortion Ratio (SDR) results than NMF and VAEs on datasets such as TIMIT. Diffusion-based models have emerged as an important alternative. Diffusion-based models can be trained with a likelihood-based or adversarial objective, and more often produce natural, realistic outputs with fewer artifacts than reconstruction-based models [15][16]. DiffWave and SEPDIFF are two notable examples demonstrating the promise of diffusion models for speech enhancement and source separation respectively.

Also, autoregressive end-to-end architectures, like SepFormer, have gained traction and work directly in the time domain without the phase reconstruction challenges associated with frequency-domain models. SepFormer methods, as well as approaches that use GANs in general, have demonstrated performance on supervised separation benchmark datasets [17][8].

Moreover, because there are few studies focused explicitly on background music removal, the broader domain of music source separation provides useful insight. Neural network-based approaches have emerged such as Hybrid Demucs and Hybrid Transformer Demucs. These state-of-the-art models process both waveform and spectrogram domain features by combining convolutional and transformer architectures. They achieve high performance in multi-source music separation tasks, such as isolating vocals, drums, bass, and other elements, especially when trained on large multi-track datasets using data augmentation and long-context modeling. These models offer superior accuracy and generalization to various music genres [18] [19]. However, they typically require considerable computational resources and annotated training data and are not directly designed for distinguishing music from non-music content such as speech [18] [19]. Another branch of recent work includes GAN-based source separation models. These leverage perceptual loss functions and operate in hybrid feature domains, improving audio realism and phase consistency, though they still face challenges with phase estimation and hyperparameter tuning in practical environments [18].

In music separation, conventional non-neural methods have also been investigated. The unsupervised spectral decomposition offered by Nonnegative Matrix Factorization (NMF) and its variations, such as NMFD, is appropriate for the isolation of monaural instruments in comparatively clean mixtures. They were widely used in early music separation research because of their main benefit, which is that they don't require labeled data [18]. However, their performance tends to degrade significantly in complex, noisy environments with overlapping sources. Independent Component Analysis (ICA) is another classical method, generally applied in multi-channel settings where it assumes statistical independence of sources. In consequence of this necessity, ICA hardly works well for challenging real corrupt monaural mixtures as those encountered in background music removal [18].

Other approaches have considered the singing voice separation via pixel-wise CNN masking. These models segment the time-frequency representation into speakers, with training based on supervised neural networks that utilize ideal binary masks. They achieved very good performance for vocal/accompaniment separation, but they are typically tailored to different styles of instruments or voice isolation and do not have general applicability across a range of music styles [5].

Speech separation work also advances our comprehension of this issue. Dual-Path RNNs for instance is a time-domain-based approach that have been shown to achieve high-performance source separation for overlapping speech, and in some cases, against background noise or music. Networks like these have been shown to generalise well across various recording conditions, speakers and languages. Yet such methods often have to be specially modified to capture the spectral details that are distinctive for music, and they do not naturally account for the delicate balance of perception as required for copyright-related contexts [5]. Models like attention-based generative networks, and Weighted-Factor Autoencoders (WFAE) have also been studied to model source-specific latent representations. These methods aim to increase the separation of overlapping audio sources by incorporating discriminative features and have achieved promising performance on monaural speech separation. But their simple use for background music suppression is still an emerging topic for research purpose [20].

III. Data Selection

Data selection is a key part of training any generative model, especially for the task of audio source separation with complicated mixtures. The objective of this work is to separate speech from background music while preserving any ambient environmental noise. To the best of our knowledge, there is no openly available dataset that has mixtures that contain music with vocals, speech, and environmental noise. To address this limitation, we construct a custom dataset by combining audio segments from three publicly available sources. Each one represents a different part of the mixture: one for clean speech, one for background music (including vocals), and one for ambient or environmental sounds.

A. Speech Dataset

For the clean speech component of our dataset, we used the CSTR VCTK Corpus [21], which is a well-known multi-speaker English dataset developed by the University of Edinburgh. It includes recordings from 110 native speakers with a wide range of accents. Each speaker reads around 400 sentences. All the recordings were made in a hemi-anechoic chamber using consistent hardware, so the audio quality is very clean.

TABLE I: Summary of Previous Approaches

Approach	Features	Strengths	Weaknesses	Best Results
Conv-TasNet [1]	Time-domain convolutional network	Effective targeted music removal preserving environmental sounds	Limited metrics; struggles distinguishing speech from vocals	Qualitative improvements in listening tests
CDAE [4]	Data-driven denoising autoencoder	Effective genre and language generalization	Limited real-time applicability, needs labeled data	WER ¡10% (piano), ¡24% (violin)
DPRNN, DPTNet, SepFormer [5], [7]	Dual-path RNN or Transformer models	High performance, good long-sequence handling	Computationally intensive	SI-SNRi up to 24 dB
SE-Conformer [9]	Hybrid convolution and self-attention	Good for real-time, natural audio	Still developing for complex environments	Superior PESQ, STOI scores
MossFormer [10]	Joint local-global attention model	Exceptional separation quality; computationally efficient	Complex gating structure	SI-SDRi up to 22.8 dB
VQ and NMF [14]	Classical factorization techniques	Good at speech clarity in presence of music	Struggle with complex mixtures; limited generalization	High word recognition at challenging SNR
GANs [14], [15]	Adversarial training, perceptual loss	High flexibility, good audio realism	Hyperparameter sensitivity; computational demands	Better SDR than NMF/VAEs
Diffusion models (DiffWave, SEPDIFF) [15], [16]	Likelihood/adversarial training, realistic output	Realistic audio quality with fewer artifacts	Computational complexity	Strong perceptual quality
Hybrid Demucs [18], [19]	Hybrid waveform/spectrogram transformer	High generalization and accuracy across genres	Resource-intensive; not designed for speech separation	State-of-the-art instrument isolation

The audio was originally recorded at 96 kHz and 24-bit resolution and later released in 48 kHz and 16-bit format after conversion.

We picked this dataset because it has a wide mix of speakers and accents, clean and consistent recordings, and a good amount of phonetic variety, all of which help when training a model to separate speech from noisy or complex mixtures. With over 44 hours of speech data and a wide range of English accents, the corpus provides sufficiently large and varied amount of samples to train a robust model that can generalize well to unknown speech.

B. Music Dataset

One of the few high-quality open-source datasets that includes music with vocals is MUSDB18 [22]. It is one of the most widely used dataset for music source separation research [23], [19], [24] and serves as a common benchmark for evaluating audio separation models. MUSDB18 contains 150 professionally produced full-length stereo tracks sampled at 44.1 kHz with isolated vocals and instrumental tracks, and covering a wide range of musical genres.

In this work, we use the full music mixture tracks from MUSDB18, which include both instrumental and vocals, without utilizing the isolated stems. This matches our goal of separating out speech from music with vocals, which reflects real-world conditions. Although the dataset contains approximately 10 hours of audio, we segment the tracks into shorter overlapping clips to obtain a sufficient number of training samples for the generative modeling.

C. Environmental Sound Dataset

As discussed earlier, this work aims to separate speech from music while keeping the ambient noise intact with the speech. For this purpose, similar to the work [1], we chose the ESC-50 [25] Environmental Sound Dataset, which was developed for the task of environmental sound classification and contains various environmental sounds such as animal sounds, water, human non-speech sounds, urban noises, etc. The corpus contains 2000 mono-channel recordings that are sampled at 44.1 kHz and distributed among 50 sound classes with 40 examples per class. We chose this dataset as the 5-second recordings include a wide range of ambient and background noises that closely mimic real-world interference in speech recordings, which makes it an ideal and sufficiently long corpus to train our model.

IV. GAI Architecture Selection

Largely based on our survey, various Generative AI architectures, including Autoencoders (AEs), GANs, VAEs, and diffusion models, were evaluated for their applicability to our specific and complex audio source separation problem. From these, autoencoder-based frameworks emerged as the most suitable choice. This preference is primarily due to their foundational role in source separation tasks; most current SOTA methods for both speech and music separation are inherently autoencoder based [26][17][8]. A survey of the literature has also revealed the superiority of autoencoder based methods when operating in the waveform domain. AE-style encoder-separator-decoder pipelines such as Wave-U-Net[26] and Conv-TasNet[17] dominate time-domain separation because they reconstruct waveforms end-to-end and retain phase information. AEs directly model the waveform and learn a compact latent space where masking or denoising operations can be carried out, thereby bypassing common issues in frequency-domain approaches like phase estimation and reconstruction. Given the complex nature of our mixture, we employ the encoder to find a representation where the distinct characteristics of ”music” versus ”speech + noise” becomes more apparent and therefore facilitates their separation by subsequent specialized modules.

Inside our encdoder-seperator-decoder network, we employ specialized neural network architectures that have demonstrated superior performance for specific tasks. More specifically these networks are chosen for dedicated processing streams - Temporal Convolutional Networks (TCNs) (Figure 1) for the ”music” component and Conformer blocks (Figure 2) for the ”speech” component. TCNs have successfully been used in highly influential models such as Conv-TasNet[17] and subsequent efficient variants SuDoRM-RF[27], where they show strong ability to model audio data while offering good performance and training efficiency. Our TCN blocks use pointwise and depthwise separable convolutions with increasing dilations, PReLU, and layer normalization. The Conformer blocks[28], which combine the transformer’s ability to capture global context (long-range dependencies via self-attention) with the CNN’s strength in extracting local features, are employed as they have shown excellent performance in speech separation. They are well suited for the task given the complexity of our mixture as conformer-based models such as TD-Conformer[29] and CoDPTNet[30] have outperformed traditional transformer based methods such as SepFormer[8] and achieved state-of-the-art results in noisy and reverberant speech separation conditions. The idea is to construct a hybrid, dual-stream approach which allows each architecture to leverage its inherent strengths for modeling the specific type of audio it’s tasked with separating.

Finally a masking mechanism, which is a prevalent and effective technique for source separation across the literature and especially in AE based waveform processing models, is applied to estimate a mask for each source from the outputs of the TCN and Conformer modules. We then proceed with an element-wise multiplication of the generated masks with the original encoded feature representation of the input mixture which effectively isolates the features deemed relevant to each respective source (music-related from the TCN and speech-related from the Conformer).The application of masking techniques contributes to enhanced model generalizability and promotes greater stability throughout the training process.

Our proposed approach builds upon the Conv-TasNet[17] architecture, replacing its simpler 1D-convolutional inner blocks with interconnected, task-specific modules to better handle complex data. Following is a snapshot description of our model’s architecture and

TABLE II: Architectural Components Overview

Architectural Component	**Role in Our System**	**Contribution**	**Fit to our data**	**Main References (Prior Work Evidence)**
Encoder–Separator–Decoder Autoencoder Backbone	End-to-end time-domain reconstruction for all streams	1-D conv encoder and decoder preserve phase; latent space suited for masking / denoising	Avoids STFT phase errors; handles raw 8 kHz mixture without information loss	Wave-U-Net[26], Conv-TasNet[17]
TCN Stack	Dedicated music stream	Depth-/point-wise 1-D convs with exponential dilation	Captures long rhythmic / harmonic patterns that define music; low-latency, parameter-light	Conv-TasNet[17], SuDoRM-RF[27]
Conformer Blocks	Dedicated speech stream	Self-attention (global) + depth-wise conv (local) in one layer	Separates unpredictable speech/noise from repetitive harmonics; keeps phonetics and prosody	Conformer[28], TD-Conformer [29], CoDPTNet[30]
HybridBlock	Cross-link TCN ↔ Conformer streams	Learned gates (α, β) pass a fraction of each stream to the other every layer	Lets rhythmic context guide speech modelling and vice-versa; balances dominance of music vs. speech on-the-fly	GALR[31], Sandglasset[32], CoDPTNet[30]
Mask-Head Mechanism	Produce one mask per source	Element-wise masks on latent features; no explicit phase estimation	Stable, widely proven; encourages generalisation and lets training focus on separation, not synthesis	Wave-U-Net[26], Conv-TasNet[17]

its operational flow as depicted in Figure 3.

The mixture waveform is first passed through a 1-D convolutional encoder that transforms raw audio into a sequence of latent feature frames. This sequence enters a stack of K hybrid separation blocks; each block contains a TCN stream and a Conformer stream that run in parallel, exchange their outputs through residual cross-connections, and forward a fused representation to the next block, while a skip connection routes

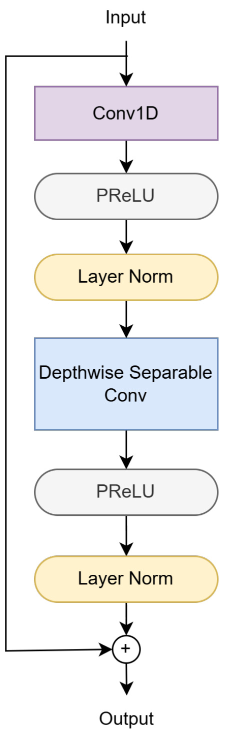


Fig. 1: Block diagram of TCN Block

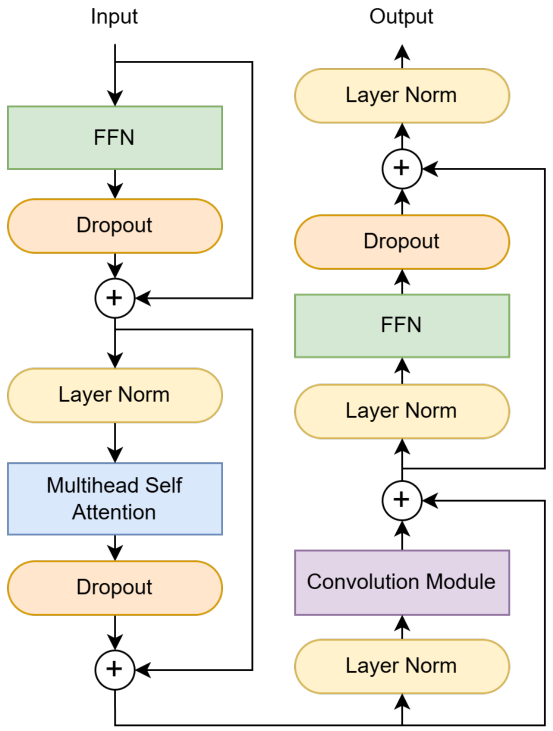


Fig. 2: Block diagram of Conformer block

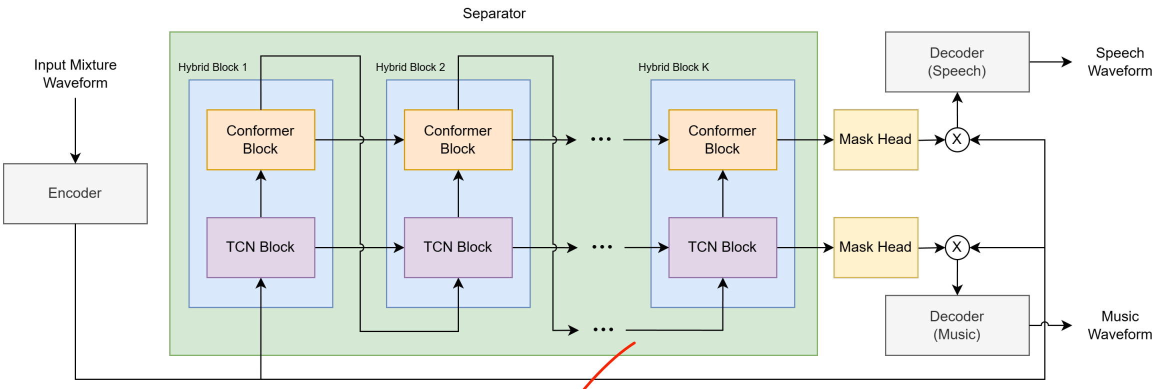


Fig. 3: Block diagram of TACIT

the encoder output around the entire stack. After the final block, a dual "mask head" projects the shared representation into two element-wise masks—one intended to retain speech (+ ambient noise) and the other to capture music. Each mask is multiplied with the original encoded features to yield source-specific latent embeddings, which are then fed into two decoders (using convtranspose adapted from conv-tasnet architechutre) that synthesize time-domain waveforms: the first decoder returns the cleaned speech track that the application keeps, and the second reconstructs the music stem that can be discarded or saved for supervision during training.

V. Why do chosen GAI Architectures Work?

Above we outlined the key components of our proposed architectures. We also discussed the rationale behind our selection criteria for using an autoencoder based model. An important factor worth reiterating that drove our decision is the ability to directly operate on the audio waveform, which as discussed autoencoder based models excel at. Working in the time-domain avoids potential artifacts and information loss (most commonly phase information) associated with time-frequency transformations, which is crucial for reconstructing high-quality speech. Below we continue the discussion with more in-depth look into the modules and their interactions to reveal why we believe they could work well for our task. Table III encapsulates our discussion.

TABLE III: Module-specific Architectural Justifications

Module (Stream)	**Core Idea & Mechanisms**	**Why It Suits the Task**	**Representative Evidence**
Temporal Convolutional Network (TCN) (Music stream)	1-D depth-wise convolutions with exponentially increasing dilations	Captures long musical events—beats, riffs, sustained vocals—over hundreds of milliseconds; handles them as one coherent "music" source for suppression; avoids RNN latency & gradient issues	Conv-TasNet [17], SuDoRM-RF [27], SepFormer ablation on RF [8]
Conformer (Speech stream)	Couples multi-head self-attention (global view) with depth-wise convolution (local phonetic cues) inside each block	Skips over stationary musical harmonics while locking on speech/noise details; keeps both fine-grain phonetics and long-range prosody, improving speech intelligibility in cluttered mixes	Conformer[28], TD-Conformer vs. SepFormer on noisy speech[29], CoDPTNet gains on WHAMR! [13]/MC-WSJ-2Mix[11]
HybridBlock (Dual-stream fusion)	Learned gates (α, β) pass a fraction of Conformer output to TCN and vice-versa at every layer $\rightarrow$ adaptive cross-pollination	Lets rhythmic/harmonic context guide speech modelling and speech cues refine music masking; balances streams on-the-fly (e.g., chorus vs. spoken line) for tighter separation	GALR[31], Sandglasset hybrid designs[32], CoDPTNet shows interaction $>$ isolation[30]

A. Temporal Convolutional Network (TCN) Blocks (for Music Stream)

Temporal Convolutional Networks are a natural fit for the music stream because they marry a very large effective receptive field with light-weight, fully parallel 1-D convolutions. By stacking depth-wise separable convolutions whose dilation factors grow exponentially, as popularised in Conv-TasNet[17] and further optimised in SuDoRM-RF[27], a TCN can "see" hundreds of milliseconds of context in a single forward pass, which is sufficient to capture the recurring rhythms, harmonic progressions, and sustained vocals that define musical structure. Moreover, It does this with far fewer parameters, and without the vanishing-gradient or sequential bottlenecks that plague RNNs[8]. This combination lets the network recognise a guitar riff that spans multiple bars, trace a melodic line through a chorus, and detect the quasi-stationary spectral envelope of backing vocals, and then treat all of those temporally coherent elements as a unified music source for suppression. In short, the TCN's dilation-driven field of view aligns perfectly with the long-range, patterned nature of music, allowing for precise and low-latency isolation of the entire music component while keeping computation and training stable.

B. Conformer Blocks (for Speech Stream)

Conformer layers mix self-attention, which looks across the whole signal, with a small depth-wise convolution, which focuses on nearby samples, so one block can "see" both the big picture and the tiny phonetic cues a listener needs to follow speech [28]. This blend turns out to be very good at ignoring steady musical harmonics - the attention part skips over them, while the convolution lock onto the less-predictable speech and background-noise patterns that remain. Empirically, Conformer-based separators such as TD-Conformer[29] outperform the Transformer baseline SepFormer on noisy-reverberant mixtures, demonstrating superior speech retention under complex acoustic conditions. When paired with convolutional streams the synergy is even clearer: CoDPTNet[30], which fuses Conformer and Dual-Path Transformer blocks, surpasses both of its parents on WHAMR! [13] and MC-WSJ-2MIX, confirming that Conformer complements the translation-equivariant yet locality-limited TCN. Because speech intelligibility depends simultaneously on fine-grain phonetic cues and long-range prosody or semantic flow, this capacity to model cross-time spectral relationships makes Conformers a natural choice for isolating the speech + noise track when it is embedded beneath a highly structured music background.

C. Hybrid Block (Dual-Stream with Interactive Processing)

While the TCN and Conformer blocks are specialized to handle different stream, they are not entirely isolated. A Hybrid Block in our architecture lets the two specialised streams talk to one another instead of working in alone. At each layer a small, learned gate mixes a fraction α of the Conformer output into the TCN path and a fraction β of the TCN output into the Conformer path, so the network can decide on-the-fly how much rhythmic or harmonic context (carried by the TCN) should guide speech modelling, and how much long-range prosody or noise detail (captured by the Conformer) should refine the music mask. This kind of cross-pollination echoes earlier hybrid separators — GALR (Globally Attentive Locally Recurrent Networks)[31], which swaps local RNN features and global attention features to cut memory without losing quality, and Sandglasset[32], which cascades multi-granularity self-attention with convolution to the same end . Moreover, CoDPTNet showed that explicitly fusing Conformer blocks with a complementary sequence model beats either branch alone on noisy-reverb benchmarks, confirming that interaction, not isolation, brings the gains. Intuitively, if a chorus dominates a moment in time, the gate can lean on the TCN's fine-grained view while letting the Conformer step back; when speech emerges, the balance can flip, giving the model an adaptive, context-aware way to keep spoken words and ambient noise while muting the music.

VI. Validation Methodology

In this section we outline our experimental setup and validation methodology. A discussion on all the different types of loss functions employed is provided. Following that, we discuss the composition of our dataset from the three datasets listed above. Finally we discuss the evaluation metrics we used to evaluate our model at differnt stages and against different baselines. Although this task is somewhat novel and slightly diverges from the traditional tasks of music and speech separtion, the metrics chosen are consistent with common benchmarks for similar tasks.

Experiment Setup

The validation methodology used in this work closely follows the standard supervised training setup. The used dataset is first split into three non-overlapping sets: train, validation, and test. This same data split is used across all experiments, ablations and baselines. The dataset generation procedure used to create the dataset is discussed in detail later in this section. All models are trained using the training set and validated using the validation set, at the end of every epoch. This allows us to monitor the learning progress and tune hyperparameters such as learning rate, gradient clipping, and loss weights.

The validation step computes the same combination of losses used during training and other perceptual evaluation metrics. These scores are tracked per epoch and are used to determine the best model checkpoint, to save it and use it later for the final testing.

For model selection, an ablation study was first conducted using a shallower version of our proposed architecture. We experiment with multiple combinations of loss functions to understand their individual and joint effects on separation quality. The experiments included using SI-SNR for both speech and music, SI-SNR with Mel loss for speech, and other combinations, including source-specific L1 losses. Our goal here was to find which loss combination provides the best trade-off between the quality of speech and music suppression. Once we identified the best-performing loss configuration, we used it to train a deeper version of the model with more hybrid blocks and conformer heads. This deeper model was then compared with two baseline methods: Conv-TasNet[17] and HTDemucs[19]. Both baselines were re-trained from scratch on our custom dataset using the same train and validation splits as our model and using the same loss functions used in the original works, i.e. SI-SNR per source for Conv-TasNet and per source L1 reconstruction loss for HTDemucs. This was done to ensure fairness and to make sure the models were exposed to the same type of mixtures and audio duration during training.

For the final evaluation, all models were tested on the same set of 10-second audio segments created from the test split mentioned earlier. The test evaluation computes multiple separation quality metrics, including overall SI-SNR, source-wise SI-SNR for both speech and music and two perceptual scores, PESQ and STOI. These scores are used to report the results presented in the next section. All reported results are from the best validation checkpoint for every model.

Loss Functions

As discussed in the previous section, we test out a combination of three loss functions: SI-SNR, L1 reconstruction, and Mel spectrogram loss, to ensure better perceptual quality of the speech and suppression of music. Each loss chosen captures a distinct aspect of reconstruction quality, or helps in earlier convergence. Combining these losses may enable the model to generalize well, and this is what we try to evaluate.

SI-SNR: Scale-Invariant Signal-to-Noise Ratio (SI-SNR) is a metric that evaluates the similarity between the predicted and the actual audio source signals in the time domain, independent of their amplitude or loudness scale. It is particularly effective for assessing the quality of audio source separation [33]. The SI-SNR we use can be defined as:

$$\text{SI-SNR}(s, \hat{s}) = 10 \cdot \log_{10} \left(\frac{\|\alpha s\|^2}{\|\hat{s} - \alpha s\|^2} \right), \quad \text{where} \quad \alpha = \frac{\langle \hat{s}, s \rangle}{\|s\|^2}$$

where, s is the actual audio signal and $\hat{s}$ the estimated audio signal The corresponding loss can then be calculated as the negative of SI-SNR:

$$\mathcal{L}_{\text{SI-SNR}}(s, \hat{s}) = -\text{SI-SNR}(s, \hat{s}).$$

Due to its nature of ignoring amplitude variations, this loss function is widely used in many state-of-the-art time-domain audio source separation models, such as Conv-TasNet[17] and DPTNet[7]. It is particularly useful when the goal is to directly reconstruct the individual source waveforms, as it is in our case.

L1 Reconstruction loss: The L1 reconstruction loss is the absolute difference between the estimated source and the actual source waveforms. Given the actual source s and the estimated source $\hat{s}$, the L1 loss is defined as:

$$\mathcal{L}_{\text{Recon}}(s, \hat{s}) = \|s - \hat{s}\|_1.$$

L1 loss is more commonly used in recent audio separation models, especially in waveform-based models like HTDemucs [19]. In HTDemucs, using L1 loss helped with training stability and tends to give better fidelity for both speech and music. It's also used in GAN-based models for music separation like [34], where it works as a stabilizing loss alongside adversarial loss. In our work, we experiment L1 loss in both these setups, in the first setup it's used as an extra term to help with the full mixture waveform reconstruction, and in the other setups it's either used alongside SI-SNR or on its own for each of the separate sources. Using the L1 reconstruction loss for the mixture alongside the individual losses per source might prevent noise artefacts from occurring in the estimates and promote early convergence.

Mel spectogram loss: Mel spectrogram loss compares the Mel spectrogram representations of the estimated source with the actual ground truth source, introducing perceptual supervision. The Mel scale improves the reconstructed audio's perceived quality because it is more in line with human hearing frequency range [35]. The Mel spectrogram loss can be computed as:

$$\mathcal{L}_{\text{Mel}}(s, \hat{s}) = \| \log(\text{Mel}(s) + \epsilon) - \log(\text{Mel}(\hat{s}) + \epsilon)\|_1,$$

where, $\text{Mel}(s)$ and $\text{Mel}(\hat{s})$ denote the Mel spectrograms of the reference and estimated sources, respectively, and ϵ is a small constant to avoid the logarithm of zero.

Recent research on improving speech, such CleanMel [35] and Mel-FullSubNet [36], shows that training models in the Mel-frequency domain improves speech quality and generalization. Mel-based objectives have been demonstrated to enhance the perceptual realism of downstream tasks such as automated speech recognition (ASR) and improved audio. We represent the loss as a function of the Mel spectograms in our work rather than modeling in the Mel-frequency domain, which can also help in early convergence when used as an additional loss component.

Dataset Generation

As we don't have readily available datasets for our task, we constructed a custom dataset by combining samples from the previously mentioned datasets: MUSDB18 [22] for music with vocals, VCTK [21] for clean speech, and ESC-50 [25] for environmental noise. As the MUSDB18 dataset originally comes as a `.stem.mp4` format, which is a single file with all the separated stems, we first convert the stems to `.wav` files and only use the mixture track, which includes both instruments and vocals. We then resamples the audio files for all three datasets to a common sampling rate (SR) of 16kHz, for consistency, and converted them to mono-channel to better align with our aim. Each of the three datasets were then independently split into training, validation, and test subsets using a 70%–15%–15% ratio to ensure balanced combinations during model training-evaluation and to prevent any cross leakage.

Now moving on to actual dataset creation, each sample is constructed by first selecting a music track to define a target duration of the sample. Then, multiple speech and noise files are stitched together sequentially to match or exceed the defined duration. These speech and noise tracks are then mixed at a fixed signal-to-noise ratio (SNR) to create the resulting speech source ground truth. The SNR is calculated as

$$\text{SNR}_{\text{noise}} = 10 \log_{10} \left(\frac{\|s\|^2}{\|n\|^2} \right)$$

Where s and n represent the speech and noise signals, respectively. To simulate a moderately noisy environment we set this SNR value at a fixed 5dB.

Once the speech+noise source is ready, we add the previously selected music track as a background audio to create the final mixture. To simulate various levels of music interference, as in real-world scenarios, we mix the music at a randomly selected SNR, between +5dB to -5dB, relative to the speech-noise source. The music-to-speech+noise SNR is computed in a similar manner as before:

$$\text{SNR}_{\text{music}} = 10 \log_{10} \left(\frac{\|s+n\|^2}{\|m\|^2} \right)$$

where m is the music signal. To ensure perceptual clarity and avoid clipping, all signals are normalized after mixing. The resulting mixture and the speech+noise, and music sources are stored as separate components to be used for training and evaluation later.

Evaluation Metrics

SI-SNR: As already discussed in the section about losses, SI-SNR, or Scale-Invariant Signal-to-Noise Ratio, is a measure of how the predicted audio signal relates to the actual audio signal in the time domain, and does not modulate for amplitude scaling, meaning it can be utilized as-is to edit how well the model separates target sources by contrasting the clean waveform with the predicted waveform. SI-SNR is typically used as an evaluation metric for time-domain speech separation models, such as Conv-TasNet[17] or DPTNet [7].

PESQ: PESQ, or Perceptual Evaluation of Speech Quality, is an objective measure of the quality of the separated speech from a cognitive perspective – natural versus distorted. PESQ is an objective wrapper to emulate human judgement which can be an important measure for measuring the perceived quality of a speech signal can be between 1 (poor quality) and 4.5 (high quality). PESQ has been used widely in evaluation of telecommunication and speech enhancement systems [19].

STOI: STOI, or Short-Time Objective Intelligibility, isn't an objective measure of the intelligibility of the separated speech when competing or background noise is present. The short-time intelligibility denotes the similarity of the spectral envelope between clean speech and degraded speech over short time intervals. A STOI score can range from 0 to 1, with values closer to 1 being described as more intelligible or higher quality. This objective speech intelligibility nudges the scores to measure how separation models generalize to their performance across tasks ASR [19].

SI-SNRi: The Scale-Invariant Signal-to-Noise Ratio improvement (SI-SNRi) quantifies the enhancement in source separation performance as determined by comparing the SI-SNR of the model's output with the SI-SNR of the original mixture. SI-SNRi measures how much better the model reconstructs the source compared to the unmodified input. Improvement over the input means a higher SI-SNR. SI-SNRi is a standard metric in the research literature in speech separation, as in Conv-TasNet[17] and DPTNet [7].

VII. Results

A. Discussion on Loss function ablations

The choice of loss function for each constituent and mixture (speech, music, mixture) was not arbitrary. Because we add music at different volumes when constructing the mixture, and L1 reconstruction loss cares so much about the amplitude of the reconstruction, it did not make sense to have L1 loss for music. An experiment was carried out using L1 for music and the results supported our hypothesis. Upon seeing results for Speech - SISNR+Mel and Music - SISNR with and without L1 mixture loss, we observed that L1 mixture was not adding anything significant therefore it was excluded from further experiments. All losses for speech were kept as it is our ultimate objective. Table IV shows the ablation results on the experiments carried out for various loss function permutations.

TABLE IV: Ablation Results for Loss Function Combinations

Loss			**Metrics**					
Speech	**Music**	**Mixture**	**SI-SNR ↑ Speech**	**SI-SNRi ↑ Speech**	**PESQ ↑ Speech**	**STOI ↑ Speech**	**SI-SNR ↑ Music**	**SI-SNR ↑ Overall**
L1	L1	-	3.369	2.764	1.912	0.611	3.222	3.296
L1	SI-SNR	-	6.350	5.745	1.852	0.601	6.841	6.596
L1 + Mel	SI-SNR	-	6.493	5.888	1.894	0.604	6.763	6.628
Mel	SI-SNR	-	-6.675	-7.280	1.998	0.613	6.868	0.097
SI-SNR + Mel	SI-SNR	L1	6.844	6.238	2.092	**0.637**	6.831	6.837
SI-SNR + Mel	SI-SNR	-	7.229	6.624	2.068	0.631	6.943	7.086
SI-SNR + L1	SI-SNR	-	**7.241**	**6.636**	2.067	0.629	**6.948**	**7.095**
SI-SNR	SI-SNR	-	7.001	6.396	**2.101**	0.635	6.903	6.952

Several observations could be made from the loss function ablations. Pure SI-SNR optimization (“SI-SNR / SI-SNR ”) delivered a solid 6.95 dB overall performance but it leaves audible high-frequency artefacts reflected in lower PESQ. This suggests that Time-domain SI-SNR is necessary but not entirely sufficient. Adding a single auxiliary (L1 or Mel) term nudges different behaviours. SI-SNR + L1 nudges the estimator toward amplitude-faithful reconstructions, raising speech SI-SNR to 7.24 dB and slightly boosting PESQ without harming STOI, while SI-SNR + Mel gives similar SI-SNR but a larger PESQ (+0.026) and STOI (+0.006) gain, indicating better perceptual quality. The later plateaus on music SI-SNR.

An interesting observation was Mel-only supervision back-fires for speech separation. Using Mel / SI-SNR the model collapses, yielding -6.7 dB speech SI-SNR. We deduce sole focus on perceptual magnitude encourages the network to “explain away” speech energy as music because phase-aligned fine structure is unconstrained.

The clear winners combine structure-aware (SI-SNR) and perception-aware (L1 or Mel) losses. SI-SNR alone maximises SNR but overlooks psycho-acoustic masking; Mel-only maximises perceptual similarity but lets absolute phase drift. Their combination enforces both waveform fidelity and perceptual plausibility, yielding the best holistic quality.

B. Discussion on Overall Results and Baseline Comparison

TABLE V: Comparison of Source Separation Models

	SI-SNR-Based Metrics				**Perceptual Metrics**	
Model	**SI-SNR ↑ Speech**	**SI-SNR ↑ Music**	**SI-SNR ↑ Overall**	**SI-SNRi ↑ Speech**	**PESQ ↑**	**STOI ↑**
Conv-TasNet	7.282	7.476	7.379	6.676	2.135	0.642
HTDemucs	4.352	2.712	3.532	3.747	1.954	0.620
TACIT$_{Deep}$ **(Ours)**	**8.045**	**7.779**	**7.912**	**7.440**	**2.142**	**0.647**

Our model (TACIT) outperforms both Conv-TasNet and HT-Demucs on every metric we report (Table V). It gets a +0.63 dB boost in overall SI-SNR over Conv-TasNet, which comes out to around 9 percent relative error reduction. Compared to HT-Demucs, the gain is even bigger at +0.764 dB, which results in more than double the separation quality. These improvements also show up in intelligibility (STOI increases by 0.005) and perceptual quality (PESQ increases by 0.007). While the numbers are small, they are meaningful since those scales are narrow and usually change very little. This shows that improvements at the waveform level actually carry over to metrics that matter for human perception.

These results support our main idea, combining a TCN for music with a Conformer for speech, and letting them share information through learnable gates, gives us a feature space that captures both rhythm and speech structure. This hybrid setup seems to find a balance between two competing goals: removing structured, repetitive music while preserving the more unpredictable and detail-heavy speech and ambient signals. Table VI below summarizes the overall performance over the baseline models.

Finally, we conducted a listening survey in which 18 participants evaluated the speech-plus-noise outputs from all three models. As shown in Figure 4, our method (Model 2) achieved the highest perceptual ratings. Participants also rated the project's overall objective and its relevance to real-world applications; those results, displayed in Figures 5 and 6, reinforce the value of this research direction and encourage further investigation.

TABLE VI: Qualitative Observations and Explanations

Observation	Likely Cause	Evidence & Reasoning
Higher music-side SI-SNR than Conv-TasNet (+0.303 dB)	The larger receptive field from dilated convolutions and global gating helps the model see more context. Unlike Conv-TasNet, which also looks over long windows, our model can use global speech patterns to better tell vocals and speech apart.	Music SI-SNR goes up without hurting speech quality, suggesting the music mask learns to ignore energy that sounds like speech.
Large gap to HT-Demucs (+5 dB speech)	HT-Demucs is mainly built for music separation, so it treats the speech+noise mix as background ambience and doesn't give it enough filters, which hurts the separation.	Music SI-SNR drops to 2.7 dB while speech stays at 4.4 dB, which shows the model isn't removing enough music, not that it's over-suppressing speech.
Modest but consistent PESQ/STOI edge	Hybrid blocks minimize speech artefacts introduced by aggressive masking.	Ablations (previous section) show PESQ gains correlate with the presence of Mel or L1 losses for the speech source.

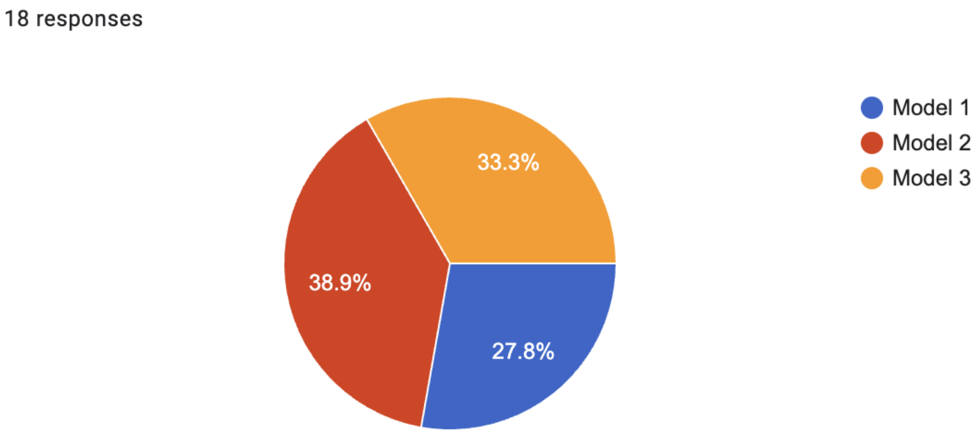
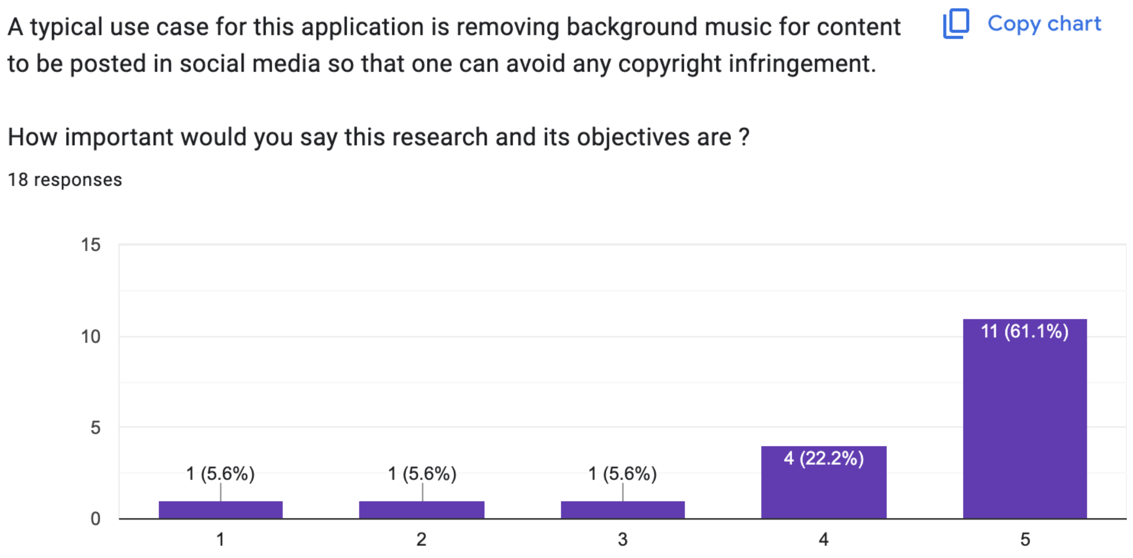


Fig. 4: Poll for HTDemucs (Model 1), TACIT (Model 2), and Conv-TasNet (Model 3)

VIII. Discussion and Future Work

This area of research is fairly new. To the best of our knowledge ours was preceded by only study [1] that had similar objective. Although the performance of the proposed model was not excellent, it was able to surpass the performance of existing SOTA in music and speech separation tasks. The model mostly gets confused between the vocals of the music

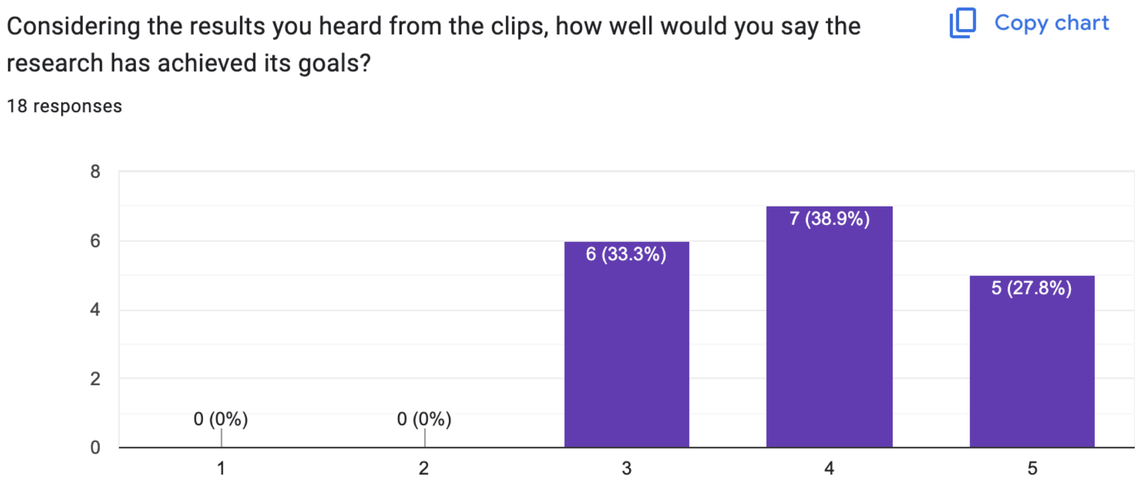


A typical use case for this application is removing background music for content to be posted in social media so that one can avoid any copyright infringement.

How important would you say this research and its objectives are ?

18 responses

Fig. 5: Research objective validation



Considering the results you heard from the clips, how well would you say the research has achieved its goals?

18 responses

Fig. 6: Reserach goal evaluation

and the speech, handling which need further research. Based on the promising outcomes and limitations we've observed we see some potential future directions that could offer better separation. The first is to integrate the conformer with WaveNet. WaveNet's autoregressive filters are excellent at fine temporal detail and Conformer blocks capture long-range context with fewer steps. Combining them could give the best of both causal precision and global awareness. Another avenue that could be interesting is using VAEs to learn source priors for the speech + noise and music sources.

Also variants such as CVAEs improve robustness to overlapping sources and noise which could be valuable since our mixture keeps speech and random noise. Finally, our survey highlights that diffusion and adversarial models (e.g., GANs) deliver superior perceptual quality; incorporating such techniques—for instance, by using TACIT as a generator trained with an appropriate discriminator—may lead to further gains.

Link to a YouTube Video

Youtube link for the demo (https://youtu.be/nY9m_evTR4E)

References

[1] S. Ozawa and T. Ogawa, “Background Music Removal Using Deep Learning,” in *2023 IEEE 3rd International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pp. 58–62. [Online]. Available: https://ieeexplore.ieee.org/document/10217480

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5206–5210. [Online]. Available: http://ieeexplore.ieee.org/document/7178964/

[3] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. [Online]. Available: https://arxiv.org/abs/1612.01840

[4] M. Zhao, D. Wang, Z. Zhang, and X. Zhang, “Music removal by convolutional denoising autoencoder in speech recognition,” in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 338–341. [Online]. Available: https://ieeexplore.ieee.org/document/7415289

[5] S. Ansari, K. A. Alnajjar, T. Khater, S. Mahmoud, and A. Hussain, “A Robust Hybrid Neural Network Architecture for Blind Source Separation of Speech Signals Exploiting Deep Learning,” vol. 11, pp. 100 414–100 437. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10247035

[6] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” Mar. 2020.

[7] J. Chen, Q. Mao, and D. Liu, “Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation,” in *Interspeech 2020*. ISCA, pp. 2642–2646. [Online]. Available: https://www.isca-archive.org/interspeech_2020/chen20l_interspeech.html

[8] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong. Attention is All You Need in Speech Separation. [Online]. Available: http://arxiv.org/abs/2010.13154

[9] E. Kim and H. Seo, “SE-Conformer: Time-Domain Speech Enhancement Using Conformer,” pp. 2736–2740. [Online]. Available: https://www.isca-archive.org/interspeech_2021/kim21h_interspeech.html

[10] 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)*, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10096646?casa_token=0TbUkYjpCAoAAAAA:MFmiFBsK1L0Dd-BDVh4yA_Bh-8Qc5qRv1paJgw2D6hCpji7wBz2xqSQqzT-sxjOgJ5_y7QEssw

[11] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2016.7471631

[12] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, Sep. 2019.

[13] M. Maciejewski, G. Wichern, and J. Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.

[14] S. Nakano, K. Yamamoto, and S. Nakagawa, "Speech recognition in mixed sound of speech and music based on vector quantization and non-negative matrix factorization," in *Interspeech 2011*. ISCA, pp. 1781–1784. [Online]. Available: https://www.isca-archive.org/interspeech_2011/nakano11_interspeech.html

[15] S. Pascual, A. Bonafonte, and J. Serrà. SEGAN: Speech Enhancement Generative Adversarial Network. [Online]. Available: http://arxiv.org/abs/1703.09452

[16] O. Mogren. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. [Online]. Available: http://arxiv.org/abs/1611.09904

[17] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," vol. 27, no. 8, pp. 1256–1266. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8707065

[18] Q. Wu, H. Deng, K. Hu, and Z. Wang, "Music source separation via hybrid waveform and spectrogram based generative adversarial network," *Multimedia Tools and Applications*, pp. 1–15, Aug. 2024.

[19] S. Rouard, F. Massa, and A. Défossez, "Hybrid Transformers for Music Source Separation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10096956?casa_token=IA1JU6zV-WAAAAAA:Oq5auqGrp6JSO0VRyawDTo7QXhqf52sumYG57LmZMjs3XH6rhV7RmMok4N8uZcy_e68zmts

[20] J.-j. Chen, Q.-r. Mao, Y.-c. Qin, S.-q. Qian, and Z.-s. Zheng, "Latent source-specific generative factor learning for monaural speech separation using weighted-factor autoencoder," vol. 21, no. 11, pp. 1639–1650. [Online]. Available: https://doi.org/10.1631/FITEE.2000019

[21] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)." [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[22] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18 - a corpus for music separation." [Online]. Available: https://zenodo.org/record/1117372

[23] G. Plaja-Roglans, M. Miron, and X. Serra, "A DIFFUSION-INSPIRED TRAINING STRATEGY FOR SINGING VOICE EXTRACTION IN THE WAVEFORM DOMAIN," in *ISMIR 2022*, 2022.

[24] C. Lan, J. Jiang, L. Zhang, and Z. Zeng, "Blind Source Separation Based on Improved Wave-U-Net Network," *IEEE Access*, vol. 11, pp. 125 951–125 958, 2023.

[25] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, pp. 1015–1018. [Online]. Available: https://dl.acm.org/doi/10.1145/2733373.2806390

[26] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end- to-End audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, Sep. 2018, pp. 334–340.

[27] 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)*, Sep. 2020, pp. 1–6.

[28] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," May 2020.

[29] W. Ravenscroft, S. Goetze, and T. Hain, "On Time Domain Conformer Models for Monaural Speech Separation in Noisy Reverberant Acoustic Environments," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2023, pp. 1–7.

[30] ——, "COMBINING CONFORMER AND DUAL-PATH-TRANSFORMER NETWORKS FOR SINGLE CHANNEL NOISY REVERBERANT SPEECH SEPARATION," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2024, pp. 11 491–11 495.

[31] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Effective Low-Cost Time-Domain Audio Separation Using Globally Attentive Locally Recurrent Networks," *CoRR*, vol. abs/2101.05014, 2021.

[32] ——, “Sandglasset: A Light Multi-Granularity Self-Attentive Network for Time-Domain Speech Separation,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5759–5763.

[33] Y. Luo and N. Mesgarani, “TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE Press, Apr. 2018, pp. 696–700.

[34] C. Deng, Y. Zhang, S. Ma, Y. Sha, H. Song, and X. Li, “Conv-TasSAN: Separative Adversarial Network Based on Conv-TasNet,” in *Interspeech 2020*. ISCA, Oct. 2020, pp. 2647–2651.

[35] N. Shao, R. Zhou, P. Wang, X. Li, Y. Fang, Y. Yang, and X. Li. CleanMel: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR. [Online]. Available: http://arxiv.org/abs/2502.20040

[36] R. Zhou, X. Li, Y. Fang, and X. Li. Mel-FullSubNet: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR. [Online]. Available: http://arxiv.org/abs/2402.13511