

ABSTRACT

Title of Thesis: **DEMYSTIFYING MONAURAL SPEECH SEGREGATION MODELS**

Rahil Kartik Parikh
Masters of Science, 2022

Thesis Directed by: **Professor Shihab Shamma and Professor Carol Espy-Wilson**
Department of Electrical and Computer Engineering

The ‘cocktail party problem’ is the task of attending to a source of interest, usually speech, in a complex acoustic environment with concurrent sounds. Despite the apparent ease with which humans can group acoustic cues from such an environment and organize them to meaningfully perceive them, the complexity of this problem has inspired generations of neuroscientists, psychologists and engineers to develop multi-disciplinary solutions to this problem, ranging from biologically-inspired frameworks to strictly engineering solutions.

In this dissertation we first explore the biologically plausible ‘Temporal Coherence’ algorithm to perform monaural source segregation based on the timing cues of each speaker. This approach integrates biologically plausible feature extraction and hypotheses of sound object perception with current trends in deep learning. It focuses on speech segregation and de-noising in an unsupervised and online fashion. Our findings suggest that this framework is suitable for de-noising applications but is unreliable for segregating mixtures of speech in its current setting.

We then explore the recent advancements in deep learning which have led to drastic improvements

in speech segregation models. Despite their success and growing applicability, few efforts have been made to analyze the underlying principles that these networks learn to perform segregation. Here we analyze the role of harmonicity on two state-of-the-art Deep Neural Networks (DNN)-based models- Conv-TasNet and DPT-Net. We evaluate their performance with mixtures of natural speech versus slightly manipulated inharmonic speech, where harmonics are slightly frequency jittered. We find that performance deteriorates significantly if one source is even slightly harmonically jittered, e.g., an imperceptible 3% harmonic jitter degrades performance of Conv-TasNet from 15.4 dB to 0.70 dB. Training the model on inharmonic speech does not remedy this sensitivity, instead resulting in worse performance on natural speech mixtures, making inharmonicity a powerful adversarial factor in DNN models. Furthermore, additional analyses reveal that DNN algorithms deviate markedly from the biologically inspired Temporal Coherence algorithm.

Knowing that harmonicity is a critical cue for these networks to group sources we then perform a thorough investigation on ConvTasnet and DPT-Net to analyze how they perform a harmonic analysis of the input mixture. We perform ablation studies where we apply low-pass, high-pass, and band-stop filters of varying pass-bands to empirically analyze the harmonics most critical for segregation. We also investigate how these networks decide which output channel to assign to an estimated source by introducing discontinuities in synthetic mixtures. We find that end-to-end networks are highly unstable, and perform poorly when confronted with deformations which are imperceptible to humans. Replacing the encoder in these networks with a spectrogram leads to lower overall performance, but much higher stability. This work helps us to understand what information these network rely on for speech segregation, and exposes two sources of generalization-errors. It also pinpoints the encoder as the part of the network responsible for these generalization-errors, allowing for a redesign with expert knowledge or transfer learning.

The work in this dissertation helps demystify end-to-end speech segregation networks and takes a step towards solving the cocktail-party-problem.

DEMYSTIFYING MONAURAL SPEECH SEGREGATION MODELS

by

Rahil Kartik Parikh

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2022

Advisory Committee:

Professor Shihab Shamma, Chair/Advisor
Professor Carol Espy-Wilson, Co-Advisor
Professor Ramani Duraiswami

© Copyright by
Rahil Kartik Parikh
2022

Dedication

To my grandfather, Dr. Pravin Parikh for planting the seeds of curiosity and scientific exploration in my mind at an early stage.

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish.

First and foremost I'd like to thank my advisors, Professor Shihab Shamma and Professor Carol Espy-Wilson for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past four years. Their enthusiasm towards this problem and advise has invigorated me to overcome the challenges faced throughout my journey. Without their ideas and insights, this thesis would have been a distant dream. I thank Prof. Ramani Duraiswami for the wonderful discussions and collaborations over the past 2 years and in exposing me to the advances in the deep learning towards speech and music processing. His course Deep Learning for Audio Processing helped me broaden my understanding of the field, making me more captivated by it. I would like to thank Prof. Jonathan Simon, Prof. Behtash Babadi, Prof. Soheil Feizi, Prof. Rama Chellappa, Prof. Philip Resnik, Prof. Jordan Boyd-Graber and Prof. Stéphane Mallat for our enriching discussions over the past 5 years.

My colleagues at Neural Systems Lab and Speech Communication Lab have enriched my life in many ways and deserve a special mention. Working on speech-inversion with Nadee Seneviratne and Ganesh Sivaraman introduced me to performing research in speech processing. Working with Shoutik Mukherjee on temporal coherence based source segregation and Kelsey Dutta and Yashish Maduwantha on the MirrorNet ensured wonderful experiences. Saurabh Sahu,

Neha Joshi and Mohesen Rezaeizadeh made my time at the lab memorable.

I also wish to thank all my collaborators over the years— Ilya Kavalero, Gaspar Rochette, Guilhem Marion, Bowen Zhi and Ophir Gal for our enriching discussions and for sharing your passion towards the field in our work. I also owe my gratitude to my friends at UMD- Julian, Omer, Giulia, Kaan, Ameya, Anton, Mahshid and Sandeep for making my time here enjoyable

I also extend my gratitude to my managers and mentors at Alexa AI— Rahul Gupta, Sun Ming, Christophe Dupuy and Harsha Sundar for giving me wonderful opportunities to explore privacy preserving machine learning and acoustic event and scene classification. Their mentorship and the internships at Amazon Science drastically enhanced my graduate school experience.

I owe my deepest thanks to my family - my mother and father and sister who have always stood by me and guided me through my career, and have pulled me through against impossible odds at times. Words cannot express the gratitude I owe them.

Most importantly, I would like to thank my fiancé Janvi for her insurmountable support and patience. Your love has been a beacon of joy throughout grad-school.

Table of Contents

Dedication	ii
Acknowledgements	iii
Table of Contents	v
List of Figures	vii
List of Abbreviations	ix
Chapter 1: Introduction	1
1.1 The Cocktail Party Problem	1
1.1.1 Speech Segregation Using Temporal Coherence	2
1.1.2 Deep Neural Network (DNN) Based Speech Segregation	2
1.2 Outline of Thesis	4
Chapter 2: Monaural Speech Segregation Using Temporal Coherence	7
2.1 Introduction	7
2.1.1 Motivation and Outline	7
2.2 Methodology	7
2.2.1 Feature Extraction	7
2.2.2 Training Neural Networks	10
2.2.3 Reconstruction	10
2.3 Monaural Segregation	11
2.3.1 Modifications in the Reconstruction of Cortical Features	12
2.3.2 Auto-Encoder (Online Clustering) and Variations	13
2.3.3 Conclusion	18
Chapter 3: Harmonicity plays a Critical role in DNN Based versus in biologically- inspired monaural speech segregation systems	20
3.1 Overview	20
3.2 Inharmonic Sound Sources	21
3.3 Experiments	22
3.3.1 Dataset Construction and Evaluation	22
3.3.2 Results	23
3.3.3 Discussion	28
3.4 Conclusion	30

Chapter 4: Empirical Analysis on the Vulnerabilities of End-to-End Models	32
4.1 Overview	32
4.2 Experiments	34
4.2.1 Datasets and Evaluation Metrics	34
4.2.2 Results	35
4.2.3 Discussion	40
4.3 Conclusion	43
Chapter 5: Summary and Future Work	44
5.1 Summary	44
5.2 Future Work	46
Bibliography	48

List of Figures

1.1	General DNN Speech Segregation System with PIT	3
1.2	End-to-End Speech Segregation Models	4
2.1	Auditory Spectrogram Extraction [1]	8
2.2	Source Segregation Using Temporal Coherence [2]	10
2.3	Baseline Auto-Encoder for n-PCA	12
2.4	Deep Asymmetric Auto-Encoder. Bottleneck layer with softmax activation.	15
2.5	Deep Auto-Encoder with LSTM encoding layer.	16
3.1	Conv-Tasnet Segregates mixtures of harmonic tone complexes (left panels), but fails with inharmonic complexes (right panels).	24
3.2	Mixtures of natural speech with harmonic (left panels) and inharmonic (right panels) tone complexes are segregated only when both are harmonic.	24
3.3	Segregation of overlapped spectra. (Left panels) A mixture of 2 harmonic sources segregate well. (Right panels) Alternating inharmonic spectra become harmonic and incorrectly segregate during overlap.	25
3.4	Performance on mixtures of inharmonic speakers with increasing jitter J for models trained on native WSJ-2-mix	26
3.5	Performance on mixtures of inharmonic speakers with increasing jitter J for models trained on inharmonic WSJ-2-mix with $J = 0.3$	27
3.6	Conv-Tasnet successfully segregates mixtures of harmonic sources (left panels). However, it fails when the fundamentals and some lower harmonics are absent (right panels).	29
3.7	Conv-TasNet segregates two harmonic sources even when synchronous.	30
4.1	ConvTasnet fails to segregate fricatives in natural speech. The red boxes highlight where ‘/s/’ is grouped with Source-1 instead of Source-2.	33
4.2	Performance of ConvTasnet on either LP (blue) or HP (red) filtered speech for different cutoff frequencies, compared to the performance of networks trained on filtered data.	36
4.3	Conv-Tanet fails to segregate mixtures of higher harmonics (left columns) but successfully segregates mixtures of the first three harmonics (centre columns). It fails to segregate mixtures of the first three and fifth harmonics (right columns)	37
4.4	Performance of ConvTasnet on band-stoped data. Each filter is a horizontal line representing the frequencies in the stop-band, and the corresponding SDR.	38

4.5	Discontinuities are inserted into a mixture of alternating tones (upper rows). Errors in segregation can be observed when these discontinuities mute the lower frequency tone (yellow boxes in the middle rows).	39
4.6	A spectrogram-based network can successfully segregate mixtures of the higher (left column), first three (centre column) and first three and fifth (right column) harmonics	40
4.7	A spectrogram-based model can successfully segregate mixtures with inserted silences which are denoted by red and green arrows in the upper rows.	41
4.8	Density over evaluation mixtures of the ratio $\log_2(f_0^{(2)}/f_0^{(1)})$ where $f_0^{(1)}$ and $f_0^{(2)}$ are the average F0s of ConvTasnet's output channels 1 and 2 respectively.	42

List of Abbreviations

DNN	Deep Neural Networks
AE	Auto-Encoder
C-Matrix	Co-incidence Matrix
n-PCA	Non-Linear Principal Component Analysis
E2E	End-to-End
LSTM	Long Short-Term Memory
BLSTM	Bidirectional Long Short-Term Memory
Conv	Convolution Layer
TC	Temporal Coherence
CASA	Computational Auditory Scene Analysis
MFCC	Mel-frequency Cepstral coefficients
IRM	Ideal Ratio Mask
SI-SNR	Scale Invariant Source-to-Noise Ratio
SDR	Signal-to-Distortion Ratio
PIT	Permutation-Invariant Training
HP	High-Pass
LP	Low-Pass

Chapter 1: Introduction

1.1 The Cocktail Party Problem

The ‘cocktail party problem’ is the task of attending to a source of interest, usually speech, in a complex acoustic environment with concurrent sounds. Despite the apparent ease with which humans can group acoustic cues from such an environment and organize them to meaningfully perceive them [3], the complexity of this problem has inspired generations of neuroscientists, psychologists and engineers. Mimicking this ability with computational models is a problem known as computational auditory scene analysis (CASA), the goal of which is to simultaneously segregate objects in complex environments. The multi-disciplinary nature of this problem has motivated CASA algorithms using traditional pitch modulations extractions [4, 5, 6] and biologically-inspired systems which are based on frameworks such as ‘temporal coherence’ (TC) [2, 7]. In contrast to these approaches, there are strictly engineering frameworks such as signal factorization, statistical-speech modelling [8, 9] and recently deep learning [10, 11, 12, 13, 14, 15, 16, 17] which ignore biological plausibility. Advances in deep learning have facilitated drastic improvements in speech segregation [11, 12, 13, 14, 15, 16, 17]. These models can now be used at scale, to separate individual speech from mixtures, enabling downstream audio processing tasks such as Automatic Speech Recognition (ASR), allowing us to use speech technologies in the *real-world*. These model architectures have also shown promising results in music segregation [18] and

sound segregation [19]. In this work we focus on monaural speech segregation, which involves estimating C sources $s_1(t), \dots, s_C(t)$ for a single-channel audio mixture $x_{mix}(t) = \sum_{i=1}^C s_i$. For simplicity, the number of individual sources C are assumed to be 2 in this dissertation.

1.1.1 Speech Segregation Using Temporal Coherence

Following [2], a segregation algorithm based on the Temporal Coherence framework suggests that distinct features belonging to a particular source modulate together in power, differently from the features of another source. This algorithm thus relies on the temporal coherence among the features of a target to identify clusters which modulate together. In [2], this is done by training an Auto-Encoder (AE) to perform non-linear Principal Component Analysis (n-PCA) of speech features for each time-frame. The ‘eigenvectors’ obtained from the n-PCA correspond to each source in the mixture and are consequently used to mask the mixture to obtain the estimated source. This algorithm is thus capable of segregating mixtures in an unsupervised and on-line fashion.

1.1.2 Deep Neural Network (DNN) Based Speech Segregation

Deviating significantly from the Temporal Coherence framework, the majority of the deep-neural network (DNN) based segregation algorithms are trained in a supervised fashion. Supervised DNN models for speech segregation are generally designed as illustrated in Fig 1.1. The input speech mixture is presented to the DNN. The network encodes the input mixture and generates two masks for the mixture, one for each estimated speaker. The element-wise product of the mask with the representation of the mixture produces the output estimated speakers. Since the

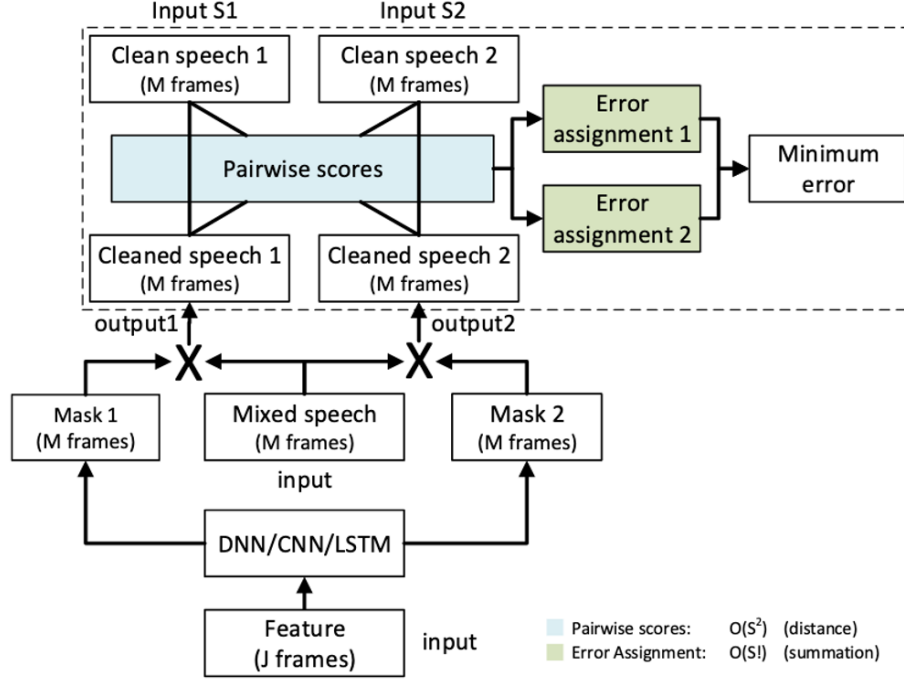


Figure 1.1: A general schematic of a DNN-based speech segregation system using Permutation Invariant Training [17]

assignment of the the estimated speaker to each of the two output channels is unknown, the error between the estimated (cleaned) speech and the ground-truth (clean) speech cannot be computed directly, presenting the *Permutation Problem*. Thus, under the Permutation Invariant Training (PIT) framework [17], the pairwise scores are computed for each permutation of the ground-truth speech and the estimated speech. The permutation which corresponds to the minimum error is then used to back-propagate the error and train the DNN.

The input mixture presented to the DNN can be a time-frequency representation of speech such as a spectrogram. In this case, the Mean Square Error (MSE) loss [20] between the estimated and ground-truth speakers is computed [10, 11, 12, 17, 21]. In end-to-end DNN based speech segregation systems, the input speech mixture is presented to the DNN directly as a waveform

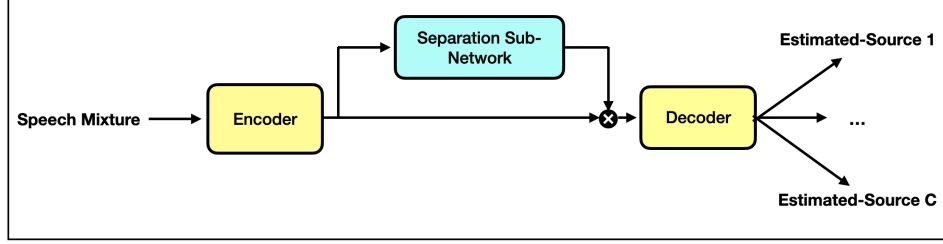


Figure 1.2: The general schematic of a End-to-End Speech Segregation Network

[13, 14, 15, 16]. This is shown in Fig. 1.2. Here, the encoder-decoder network learn an analysis-synthesis transform. These networks typically maximize the Scale Invariant Source-to-Noise Ratio (SI-SNR) or the Signal-to-Distortion Ratio (SDR) [13, 22]. Two end-to-end networks of particular importance in this dissertation are the Convolution-Tasnet (Conv-Tasnet) [14] and Dual Path Transformers Network (DPT-Net) [16]. In Conv-TasNet the segregation network consists of dilated 1-D convolutional networks, whereas DPT-Net consists of a cascade of transformers configured to maximize the temporal receptive field over the encoded mixture.

1.2 Outline of Thesis

In Chapter 2, we extend the work performed in [2] to improve the speech segregation based on Temporal Coherence. We experiment with various variations in the auto-encoder network architecture to perform unsupervised and online-segregation and denoising. More specifically, we investigate the impact of changing the masking strategy, changing the activation-function for the auto-encoder bottle-neck, increasing the depth of the auto-encoder, using recurrent networks and increasing the temporal context at each time-frame on the segregation performance. We conclude that while this framework is in-adequate for monaural speech segregation, it can be

used for unsupervised speech denoising.

In Chapter 3, we perform an empirical analysis on Conv-Tasnet and DPT-Net to demonstrate that harmonicity of speech plays a critical role in end-to-end segregation such as Conv-TasNet and DPT-Net. We evaluate the performance of these networks with mixtures of natural speech versus slightly manipulated inharmonic speech, where harmonics are slightly frequency jittered. We find that performance deteriorates significantly if one source is even slightly harmonically jittered, e.g., an imperceptible 3% harmonic jitter degrades performance of Conv-TasNet from 15.4 dB to 0.70 dB. Training the model on inharmonic speech does not remedy this sensitivity, instead resulting in worse performance on natural speech mixtures, making inharmonicity a powerful adversarial factor in DNN models. Furthermore, we demonstrate that DNN algorithms deviate markedly from the biologically inspired Temporal Coherence framework [2] that rely primarily on timing cues and not harmonicity to segregate speech.

In Chapter 4, we investigate the impact of harmonic and temporal distortions on E2E speech segregation systems. We perform various ablation studies to analyze how the network performs harmonic analysis of the mixture, where we apply low-pass, high-pass, and band-stop filters of varying pass-bands to empirically analyze the harmonics most critical for segregation. We also investigate how these networks decide which output channel to assign to an estimated source by introducing discontinuities in synthetic mixtures. We find that end-to-end networks are highly unstable, and perform poorly when confronted with deformations which are imperceptible to humans. Replacing the encoder in these networks with a spectrogram leads to lower overall performance, but much higher stability. This work helps us to understand what information these network rely on for speech segregation, and exposes two sources of generalization-errors. It also pinpoints the encoder as the part of the network responsible for these generalization-errors,

allowing for a redesign with expert knowledge or transfer learning.

Chapter 5 provides the conclusion to the thesis. It also discusses avenues for future work in the understanding and improvement of DNN based and temporal coherence based speech segregation models.

Chapter 2: Monaural Speech Segregation Using Temporal Coherence

2.1 Introduction

2.1.1 Motivation and Outline

In this chapter we extend the work in [2] to segregate speech using Temporal Coherence. This work focuses on the simultaneous source segregation of monaural mixtures in a biologically plausible fashion. We address both speech segregation (segregating speech from a mixture of two speakers) and speech de-noising (segregating speech from background noise). Here, we investigate variations of the auto-encoder network used to estimate masks for each estimated speaker and explore various reconstruction strategies to obtain the representations of the segregated source from a trained auto-encoder. This work was conducted jointly with Shoutik Mukherjee.

2.2 Methodology

2.2.1 Feature Extraction

In this work we use multi-resolution spectro-temporal features which mimic the representations of sounds in the primary auditory cortex of mammals [1], called the ‘cortical features’. The process of obtaining the cortical features are explained in Section 2.2.1.1.

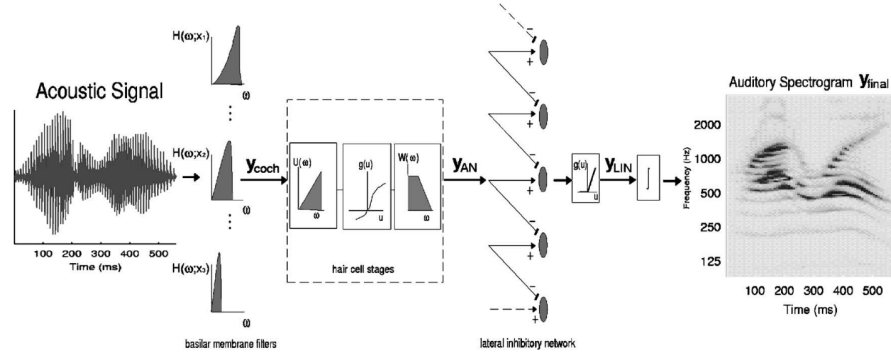


Figure 2.1: Auditory Spectrogram Extraction [1]

2.2.1.1 Cortical Features

The cortical features of the time-domain speech wave-form are a representation of sound that are based on extensive studies of the mammalian primary auditory cortex [1].

Obtaining the cortical representation of a signal, $x(t)$, consists of three steps. The first step is the extraction of the auditory spectrogram. This time-frequency representation is obtained by applying a filter bank mimicking the basilar membrane to the signal to obtain the cochleogram. Non-linearities that mimic the effect of the hair cells are then applied, followed by the lateral inhibitory network. Its outputs are integrated and rectified over small time windows, to produce the auditory spectrogram, $x(t, f)$ as shown in Fig. 2.1.

The cortical representation is obtained by the analysis of the spectral and temporal modulations performed using a 2-D wavelet transform on this auditory spectrogram. A multi-scale wavelet analysis along the spectral axis produces $x_s(t, f; s)$. A multi-rate wavelet analysis is then performed along the time-axis of the modulus of $x_s(t, f; s)$. Low rates correspond to slower temporal dynamics. The result of this analysis is the 4-D cortical representation, $X(t, r; f, s)$, where the

arguments t, r, f, s , respectively denote the time index, rate parameter, frequency channel, and scale parameter.

2.2.1.2 Temporally Coherent Features and the Coincidence Matrix

As mentioned in Chap. 1, Temporal Coherence dictates that the distinct features belonging to a particular source will modulate together in power and distinctively from other features of another source. The algorithm thus relies on identifying clusters of features that are temporally coherent and modulate together. Temporally coherent features are explicitly characterized by computing a coincidence matrix (C-matrix) whose elements reflect coherence of two cortical features.

The C-matrix at time t for rate channel r_i are computed by performing an outer-product of the flattened cortical feature vector. We supplement these cortical features with the pitch information of the two sources that also modulates with the source in time, and thus improves the identification of temporally coherent features. The C-matrix for each modulation rate r_i are stacked to produce the training data for the auto-encoder: a collection of all rate-specific C-matrices $C(t)$, given by

$$C(t) = [C_{r_1}^\top(t), \dots, C_{r_R}^\top(t)]^\top, \quad (2.1)$$

in order to bind the clusters obtained for each rate channel. The rate-channels of $C(t)$ have different temporal dynamics due to the varying time-constants of the corresponding rate-filters. Thus, the features implicitly include a degree of temporal context. This is useful in maintaining continuity in the encoding of sources, essential for streaming and segregation.

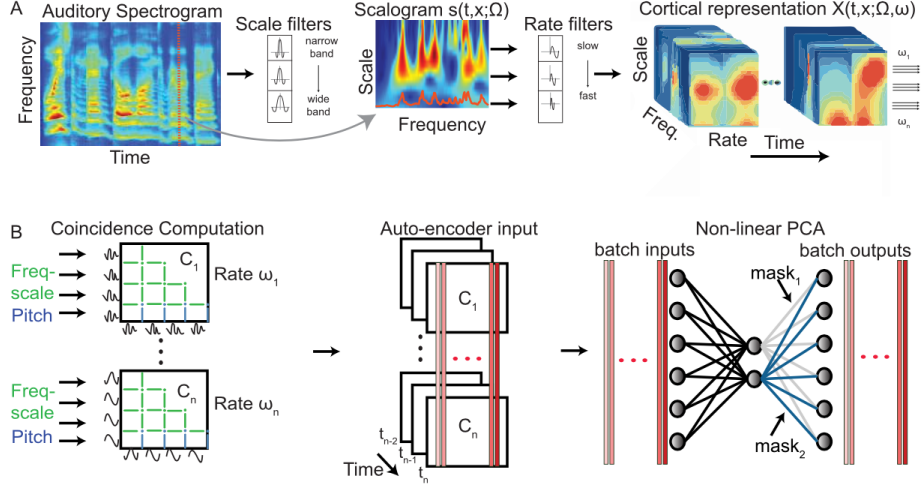


Figure 2.2: Source Segregation Using Temporal Coherence [2]

2.2.2 Training Neural Networks

The neural network architectures we examine are trained as auto-encoders. The objective function used the square loss $\|C(t) - \hat{C}(t)\|_F^2$, where $\hat{C}(t)$ is the network's estimate of $C(t)$. We update the weights at each time-step, using the previous time-step's trained weights as an initialization to enforce continuity in source assignment. Updates to the weight are made using the *Adam* optimization algorithm, which we found had the fastest convergence to the optimal weights. The process of generating the cortical features and the corresponding C-matrices and training the auto-encoder are shown in 2.2.

2.2.3 Reconstruction

After training the neural network at each time-step on the C-matrices, we use the learned weights to segregate the cortical representation of the mixture into its component sources. Each

network architecture we consider has a hidden “bottleneck” layer with two nodes; the motivation is that each of the nodes will correspond to only source. This implies two methods of segregating each source: forward-propagation of the cortical features while keeping only one of the bottleneck layer nodes active; and using the decoder weights directly as a mask on the cortical features. Later in this chapter, we compare the performance of each method of reconstruction.

The time-domain signal of the segregated sources are obtained by the inverse cortical transformation [1]. To improve intelligibility of the resulting wave-forms, we add a small amount of the mixture. Normalizing the wave-forms to zero-mean, unit-variance signals serves to remove biases and amplitude variations, further enhancing the output signals.

2.3 Monaural Segregation

We now discuss our experiments and findings for simultaneous source segregation of monaural mixtures. Our network variations and experiments are motivated primarily by the following principles: to obtain a high-dimension representation for the mixture in which improved source-segregation may be achieved and to increase temporal context to incorporate additional information about the feature trajectories. These discussions are largely qualitative, but provide helpful context for later experiments.

We first look at the auto-encoder utilized in [2] for performing n-PCA and consequently segregation. The network architecture is shown in Figure 2.3. The network consists of a single hidden layer with two nodes, one for each source to be segregated.

Sources were reconstructed using the forward-propagation method in Section 2.2.3. The 2 neurons in the bottle-neck layer are designed to capture the eigenvector of the corresponding

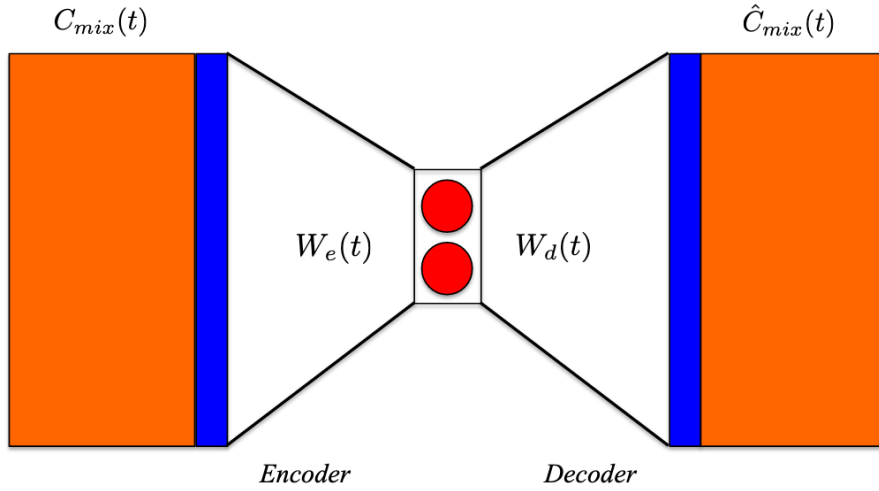


Figure 2.3: Baseline Auto-Encoder for n-PCA

source in the audio-mixture. However, we qualitatively found the mixtures were moderately segregated by the algorithm. We perform the below modifications to this baseline segregation system.

2.3.1 Modifications in the Reconstruction of Cortical Features

In this section we explore variations in reconstructing the cortical features of the segregated source after training the auto-encoder.

2.3.1.1 Reconstruction Using Network Weights

After the auto-encoder in [2] is trained to perform n-PCA, the neurons in the bottleneck layer are used to reconstruct segregated speech i.e. the cortical features of the mixtures are forward-propagated into the auto-encoder and a bottleneck neuron was masked to generate segregated speech. We explore an alternate reconstruction approach, where-in we use the weights of the

auto-encoder for masking. Here, the output weights are simply point-wise multiplied to the cortical features of the mixture. Using the output weights as a mask improves the segregation quality in comparison to the baseline auto-encoder. Additionally, this method is computationally less expensive since the masking operation is a simple element-wise multiplication in contrast to a forward-propagating through a neural network.

2.3.1.2 Cortical Features for Reconstruction

To compute the C-matrix, we perform scale filtering on the spectrogram and then take its absolute value, before performing rate filtering. However, this process is not required to compute the cortical features during reconstruction- where rate filtering can be performed directly after performing scale filtering. This would help maintain information related to phase. Another modification in the reconstruction process is that the auditory spectrogram is computed using a channel of band-pass filters and a low-pass filter, that keeps the DC component of the signal mixture for reconstruction. The inclusion of the DC component is potentially detrimental in training because it adds energy to the lowest frequency channel. This can affect the C-matrix as it becomes more difficult for the network to find truly coincident features. However, the low frequency components can contribute to the quality of reconstructed speech. These changes slightly improve the quality of the reconstruction.

2.3.2 Auto-Encoder (Online Clustering) and Variations

A key limitation to the auto-encoder used in [2] is the bottleneck layer. While it is possible that both eigenvalues can be large in n-PCA, this poses a challenge in clustering the two ‘eigenvectors’.

To enforce that there is dominating ‘eigenvector’ at each time-frame, we use the softmax activation in the bottle-neck layer of the auto-encoder. This alters the n-PCA paradigm in [2] to an online-clustering paradigm.

For the vector x inputted to the AE and the layer weights W , the pre-activation of the i^{th} node is $u_i = [Wx]_i$. The softmax activation of the same node is given by $h_i = \frac{e^{u_i}}{\sum_{j=1}^N e^{u_j}}$. Differences between the activation of nodes are now exaggerated by the exponential function. The normalization ensures that one node of N dominates the layer activation, sparsifying the layer activity and in-turn encouraging clustering. Learning the auto-encoder network weights with the softmax bottleneck layer thus amounts to online clustering of the C-matrix, where the clusters correspond to different components of the mixture.

The quality of segregation improved using the softmax activation in the bottleneck layer relative to the n-PCA in [2]. Due to the sparsification property of the softmax layer, reconstructed mixture components tended to be more distinct than for the n-PCA. These results were corroborated with experiments on arbitrary mixtures of male and female talkers from the TIMIT sentence database. We also applied this modified segregation algorithm to perform de-noising of speech. Additive White Gaussian Noise (AWGN) was added to one sentence at a time. The quality of the de-noised speech was superior to that of the n-PCA auto-encoder algorithm. Additionally, the de-noised speech was generally much more intelligible than the speech obtained from segregating a mixture of talkers.

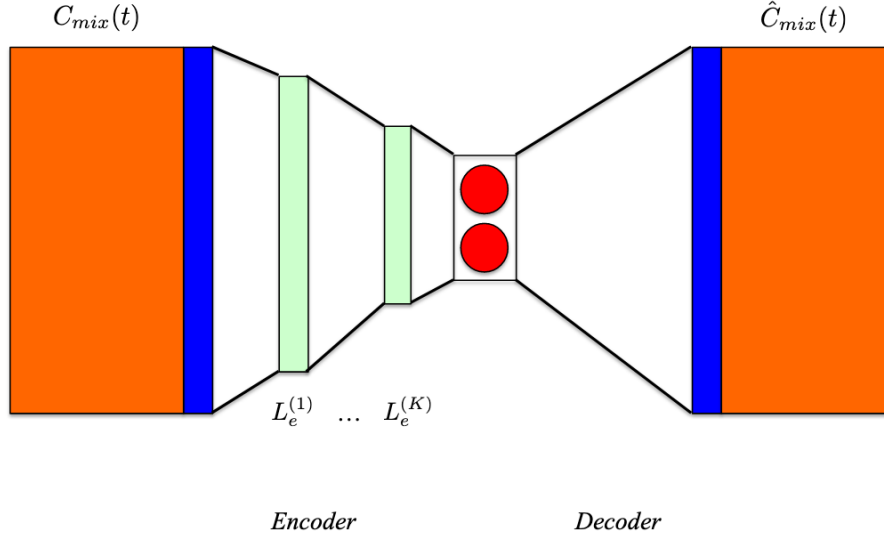


Figure 2.4: Deep Asymmetric Auto-Encoder. Bottleneck layer with softmax activation.

2.3.2.1 Deep Auto-Encoder

The auto-encoder used in [2] consists of a single hidden layer (the bottle-neck layer). We considered the use of a deep encoder to implement the notion of a deep embedded feature space. The deep encoder permits the network to learn a non-linear mapping from the input features to an intermediate set of features. We note that to reconstruct the cortical features using the decoder weights as mask as explained in Section 2.3.1, the decoder must remain shallow, i.e. directly map the bottleneck layer to the C-matrix estimate for that time-frame. These neural networks resemble the schematic in Figure 2.4.

Two fully connected layers with ℓ_1 regularization and dropout of $p = 0.5$ and ReLU activations were added to the encoder. The performance of this deep encoder model is still bounded by the performance of the baseline system. Similar architectures with more fully

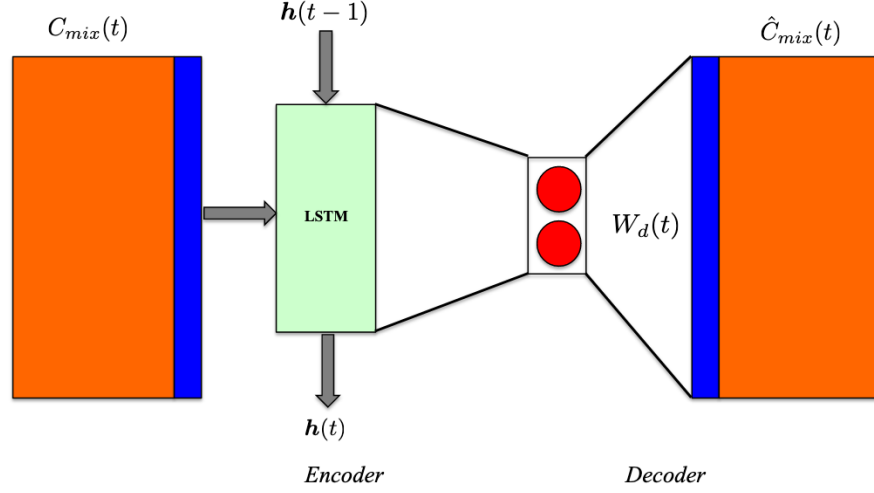


Figure 2.5: Deep Auto-Encoder with LSTM encoding layer.

connected layers were tried with no improvement on the segregation quality. This could be attributed to the fact that the hidden layers in the encoder were very dense and almost always activated, in spite of the L1 regularization, hinting that the model may be over-fitting.

2.3.2.2 LSTM Based Encoders

Continuing with the notion of improving clustering with additional encoding layers, we attempt to incorporate temporal context by using an LSTM. This is similar to the justification of BLSTM layers in the deep clustering approach of [11]. In the online context, the LSTM would have a role analogous to adaptive filtering of the inputs to create an embedding of the C-matrix that includes information about its past realizations. A schematic of the neural network is shown in Figure 2.5.

However, we observed a drop in segregation performance, likely because the LSTM may result in averaging of the temporal information over the different rates, leading to loss of information

over the previous time samples which is crucial for segregation.

Thus, by implementing a separate LSTM for the frequency channels of each rate, we avoid the intermixing and averaging of channels from different rates. This model feeds the frequency channels of each rate into an LSTM that embeds the channel into a more compact representation. The hidden states of the LSTMs are then concatenated and fed into the auto-encoder. This produces better segregation than having a single LSTM. However, the results are still comparable to the shallow soft-max autoencoder model.

2.3.2.3 Concatenating C-Matrices

Next, we look at increasing the temporal context in the features that the network is trained on. To enable the encoder to access longer input sequences, we concatenated the C-matrices across B time samples and used it as an input to the encoder. The encoder thus attempts to de-correlate patterns in the windowed C-matrix and the decoder could learn masks for each window of B samples by optimizing the criterion:

$$L = \sum_{(k-1)B}^{kB} \|C_{mix(t)} - \hat{C}_{mix(t)}\|_F^2 \quad (2.2)$$

Alternatively we can also learn the masks for each time frame using a context of B samples by optimizing the following:

$$L = \sum \|C_{mix(t)} - \hat{C}_{mix(t)}\|_F^2 \quad (2.3)$$

where $\hat{C}_{mix(t)}$ is $\mathcal{N}_{AE}(C_{mix}(t - B + 1), \dots, C_{mix}(t))$ We notice marginal improvements for small

windows of size 2 or 4 but the segregation quality reduces for longer windows. This is possibly because the weights learned for the window become an average of the weights for each time-step.

2.3.3 Conclusion

In comparing the network architectures with optimized source reconstruction, we observe that the shallow encoder based AE demonstrates the best performance. The addition of dense encoding layers did not improve the performance, and the LSTM encoding layer in fact diminished the performance. Rather than a negative result about the incorporation of temporal context or usage of deep features, we interpret this as a consequence of the trade-off between the amount of training data at each time-step and the complexity of the network.

We use the coincidences of the cortical features with the pitchgram at a given time-step as the training data for the network. As we consider a pitchgram with 128 frequency channels, this means the C-matrix has 128 columns. Each column can be viewed as one sample when training the network. Thus, when we increase the number of parameters in the neural network by adding more layers, we are still limited to 128 training samples per time step and, as a consequence, are overfitting the neural network weights.

In other optimization problems, overfitting is often addressed by imposing sparsity constraints on the parameters. We have attempted to incorporate this in our setting by using the ℓ_1 norm of the weights as opposed to the standard ℓ_2 norm as an additional penalty to the cost function. However, we did not see improvements to the quality of segregation, implying that the lack of training data is indeed problematic for more complex networks.

We conclude that in the current form, online and unsupervised source segregation using

Temporal Coherence is more suitable for de-noising than speech segregation.

Chapter 3: Harmonicity plays a Critical role in DNN Based versus in biologically-inspired monaural speech segregation systems

3.1 Overview

In the previous Chapter we demonstrated a biologically inspired speech segregation model. Despite their biological plausibility and well studied underlying principles, these models cannot be used to segregate speech for commercial and real-world applications due to the existing issues with their performance. On the other hand, advances in deep learning have facilitated drastic improvements in speech segregation [11, 12, 13, 14, 15, 16, 17]. Despite their unquestionable success, no work has been done to investigate the underlying principles they learn to perform grouping and consequently segregation. In the following chapters we attempt to bridge the gap between our understanding of traditional CASA systems, biologically inspired algorithms and deep-learning systems by investigating the role of harmonicity on the latter. Since harmonicity of natural speech is linked with pitch extraction, it is a crucial cue for grouping in certain CASA systems [4, 6]. However, Temporal Coherence based algorithms depend solely on the temporal correlation of features from the same source and are capable of performing segregation regardless of harmonicity. While the pitchgram is used to augment the spectrogram in [2] as an additional anchor for segregation, the model can segregate speech even in the absence of this pitchgram.

Recent findings in [23] found that human subjects perform worse in segregating inharmonic acoustic mixtures, performance deficits were relatively modest and often attributable to source assignment confusion rather than true segregation difficulty of the overlapping words in the mixtures. In our work, we begin by exploring the importance of harmonicity on two prominent speech segregation models Conv-TasNet [14] and Dual-Path Transformers (DTP) Net [16]. We investigate how these models perform on mixtures of harmonic speech, inharmonic speech, and a variety of synthetic simple stimuli that exemplify these properties more clearly. We also extend this analysis to the same models when trained on inharmonic speech. We find that-

- DNN based segregation systems trained on natural speech fail to segregate inharmonic speech. The performance of DPT-Net degrades from 20.2 dB to 0.7 dB for imperceptible amounts of harmonic jitter.
- DNNs find it hard to learn to segregate sources when trained on inharmonic speech, instead becoming worse at segregating natural speech.
- They differ from biologically inspired systems in heavily relying on harmonic patterns to segregate speech.

3.2 Inharmonic Sound Sources

Sounds that originate from a periodic process have a harmonic structure in their Fourier representation, consisting usually of a component at the fundamental frequency (F_0) and its integer multiples. Inharmonic sounds instead do not have frequency components at exact integer multiples of F_0 . We use the modified STRAIGHT algorithm [24] to generate inharmonic speech. This algorithm models voiced excitations as a sum of multiple time varying sinusoids, the frequencies

of which are manipulated using ‘jittering’, wherein a random offset is added to each harmonic. This offset is bounded by some fraction- J of F_0 . The n^{th} harmonic is thus modified to:

$$f_n(t) = n f_0(t) + J_n f_0(t); -J \leq J_n \leq J \quad (3.1)$$

Since other characteristics of the speech remain unchanged, synthetically generated inharmonic speech is very intelligible¹. To generate inharmonic tone complexes, we sample from $x_{tone}(t) = \sum_{k=1}^N A_k \sin(2\pi f_n(t)t)$ where $f_n(t)$ is given by (3.1). Sources are harmonic at $J_n = 0$.

3.3 Experiments

To demonstrate the importance of harmonicity as a cue for grouping, we first evaluate performance of Conv-TasNet (trained on native (harmonic) WSJ-2-mix) with mixtures of harmonic or inharmonic complex tones. To extend these results to speech, we evaluate Conv-TasNet and DPT-Net with inharmonic versions of WSJ-2-mix, with varying levels of inharmonicity. We also evaluate these models with mixtures where one speaker is natural and the other is inharmonic. Finally, we train a Conv-TasNet on only inharmonic mixtures.

3.3.1 Dataset Construction and Evaluation

We use the native WSJ-2-mix dataset [11] created using the WSJ0 corpus to train our models on harmonic speech. We create multiple inharmonic versions of the WSJ0-corpus, by varying the harmonic jitter J in (3.1) and create their corresponding inharmonic WSJ-2-mix datasets. We change J from 0.01 – 0.30, corresponding to an average offset in the range of

¹<http://labrosa.ee.columbia.edu/projects/inharmonic/>

$\pm 1.2 - \pm 40$ Hz for male speakers and $\pm 2.1 - \pm 65$ Hz for female speakers. A higher value of J corresponds to a higher deviation from the native source. We also generate versions of the WSJ-2-mix datasets wherein one source is harmonic and the other is inharmonic.

We evaluate source-segregation performance using the metric signal-to-distortion-ratio improvement (SDRi) [25]. The results are baselined against the model’s performance on the harmonic WSJ-2-mix dataset with $J = 0$. Since each of the speech datasets contain differently jittered versions of the same samples from WSJ-2-mix, we can fairly compare the evaluation performance across these datasets.

3.3.2 Results

3.3.2.1 Speech Segregation Models Trained on Natural Speech

We begin by demonstrating that Conv-TasNet trained on the native WSJ-2-mix relies primarily on harmonicity to perform separation. The test stimuli are mixtures of alternating tone complexes (with no overlap) as illustrated in Fig 3.1. The model performs well when both complexes are harmonic ($J = 0$ and $F_0 = 110, 210$ Hz) as illustrated in the left panels of Fig. 3.1. It, however, fails completely when the complexes are harmonically-jittered (right panels) with $J = 0.20$.

This phenomenon applies equally when the pair of stimuli are very different as in Fig.3.2, where one is a tone complex and the other is speech. The segregation is perfectly effective when both are harmonic (left panels) while it fails when one of the sources (the tone complex) is inharmonic (right panels) in spite of the speech being harmonic.

Harmonicity may also be a transient phenomenon during speech mixtures due to the overlap of harmonic components. To illustrate the immediate sensitivity of the DNN models to the

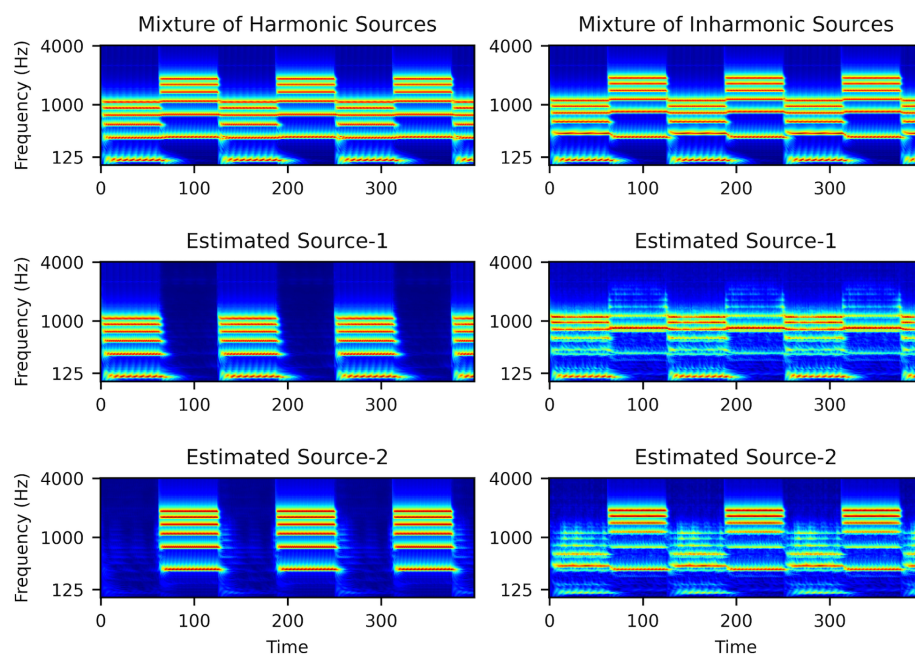


Figure 3.1: Conv-Tasnet Segregates mixtures of harmonic tone complexes (left panels), but fails with inharmonic complexes (right panels).

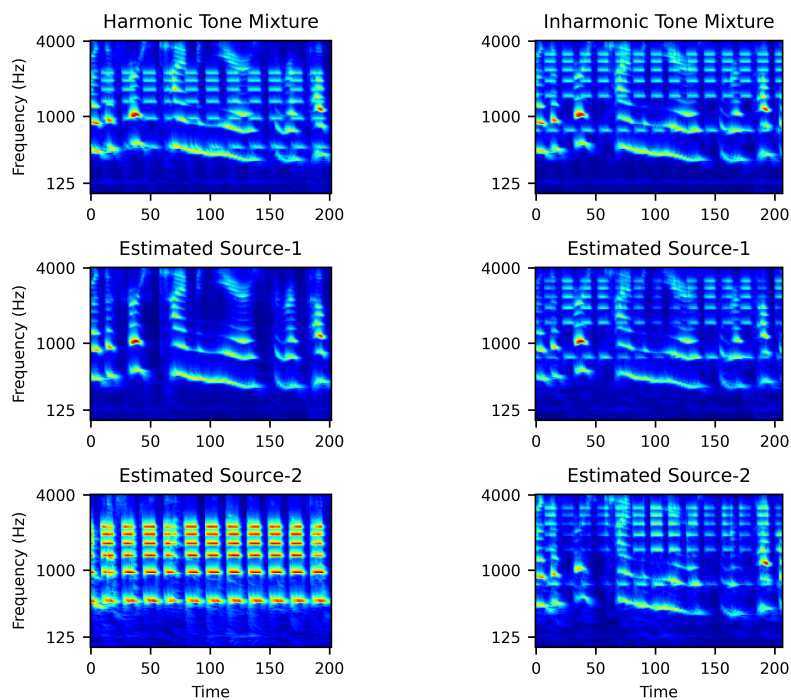


Figure 3.2: Mixtures of natural speech with harmonic (left panels) and inharmonic (right panels) tone complexes are segregated only when both are harmonic.

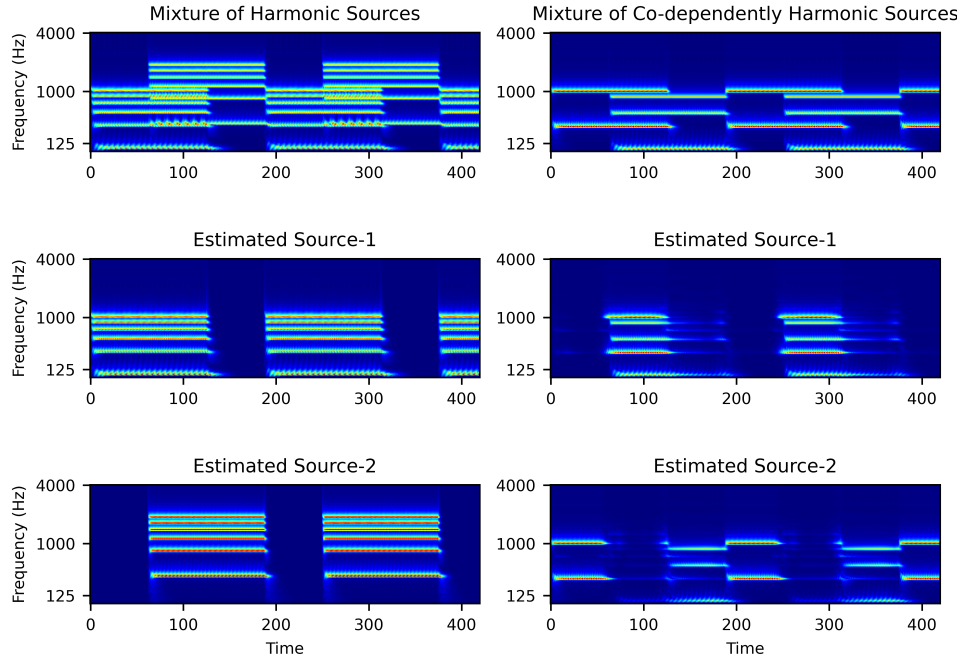


Figure 3.3: Segregation of overlapped spectra. (Left panels) A mixture of 2 harmonic sources segregate well. (Right panels) Alternating inharmonic spectra become harmonic and incorrectly segregate during overlap.

emergence (or absence) of harmonicity, we illustrate in Fig.3.3 two alternating mixtures with 50% overlap. In the left panels, both tone complexes are completely harmonic ($F_0=100$ and 190). On the right, the complexes are inharmonic with frequencies $200, 600$ Hz and $100, 300, 500$ Hz. During the intervals of overlap, the effective instantaneous spectrum becomes the sum of the two complexes. On the left, the harmonic patterns still exist in the mixture, and hence are extracted out and segregated perfectly. But in the right panels, the tone mixture strongly resemble the harmonics of $F_0=100$ Hz. The DNN model incorrectly groups this harmonic series of tones as one source, while the two individually inharmonic series become the other source. This demonstrates the strong dependence of the model on the harmonicity of the spectrum.

These tonal stimuli illustrate the vulnerabilities that affect models when harmonicity is disturbed. To extend our analysis to speech mixtures we evaluate the models using mixtures of: two inharmonic

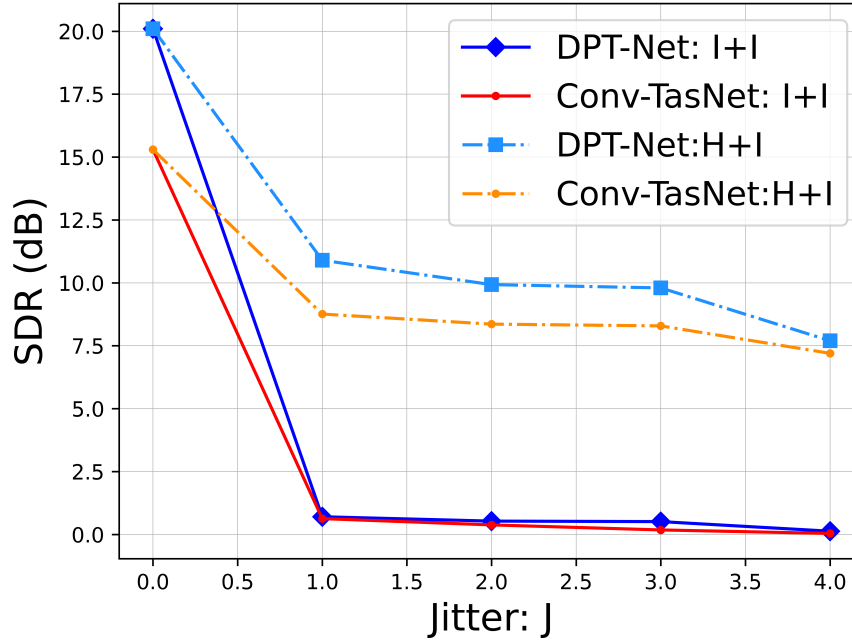


Figure 3.4: Performance on mixtures of inharmonic speakers with increasing jitter J for models trained on native WSJ-2-mix

speakers (I+I), and one harmonic and one inharmonic speaker (H+I), using datasets discussed in Sec. 3.3.1. The results in Fig.3.4 illustrate that the models’ performance degrades severely, especially on mixtures of two inharmonic speakers– 15 dB on natural speech to 0.7 dB for ConvTasnet and 20 dB to 0.7 dB for DPT-Net, for $J = 0.03$, which reveals a remarkable vulnerability.

3.3.2.2 Training DNN-Models with Inharmonic Speech

An important question is whether these DNN models can be made more robust to inharmonicity by training them on inharmonic speech. Here, we trained the Conv-TasNet on the WSJ-2-mix datasets with both inharmonic speakers. We note that the model finds it more challenging to learn to segregate with increasing inharmonicity. Our results from a model trained only on inharmonic

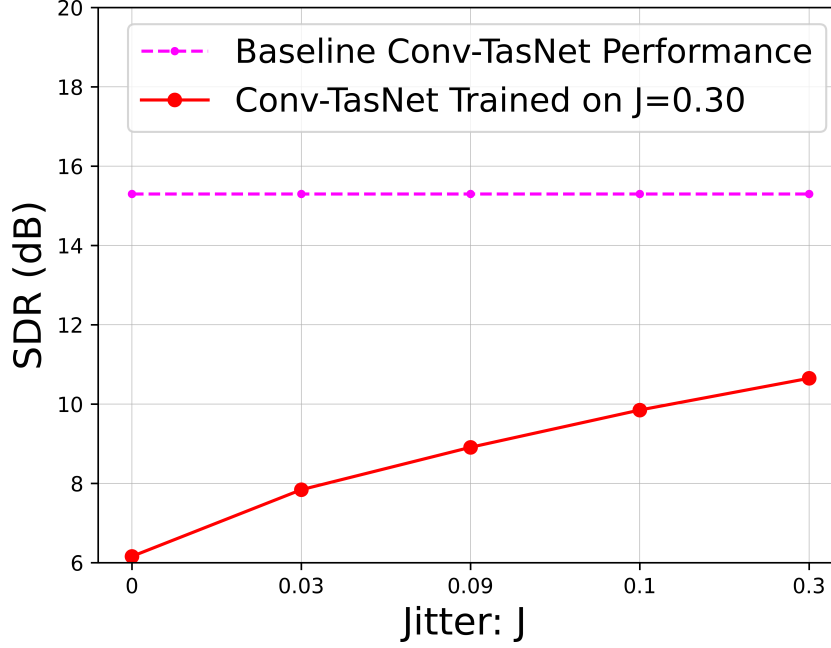


Figure 3.5: Performance on mixtures of inharmonic speakers with increasing jitter J for models trained on inharmonic WSJ-2-mix with $J = 0.3$

data with $J = 30\%$ are illustrated in Fig. 3.5. In spite of being trained on data from the same distribution, the performance of the network drops from 15.3 dB to 10.6 dB. Interestingly, this network now performs poorly on harmonic and mildly inharmonic speech. Furthermore, when the model is trained on a combination of natural and inharmonic mixtures, it still finds it hard to segregate inharmonic mixtures but learns to segregate natural speech. We believe that learning to segregate inharmonic mixtures is a harder problem since the space of inharmonic patterns is very large, while that of harmonic patterns is sparse. This makes it difficult to cover all possibilities during training. These results underscore that the networks rely strongly on the constrained harmonic structure for segregation.

3.3.3 Discussion

Our experiments indicate that end-to-end models trained on natural speech are incapable of segregating an inharmonic source. Furthermore, we have shown that the models cannot even be trained to segregate inharmonic mixtures.

These results provide clear evidence that DNN-based speech segregation is heavily reliant on learning the harmonic patterns that are available in natural speech, but that are absent or rarely available in inharmonic training. Their sensitivity to harmonicity implies the ability to track the F0, or pitch. But they seem to do so from the full spectral set of harmonic patterns and not just from a few harmonics as humans are readily able to do. For instance, in Fig.3.3, the network could not treat the sparser pattern of tones as part of a larger harmonic pattern, expecting instead to see the full patterns. To further test this, we generate a mixture of alternating harmonic tone complexes with and without the fundamental component and the second harmonic, as illustrated in Fig. 3.6. Since the model fails to segregate the tones, it is likely it has learned to recognize only the full harmonic patterns. They then proceed to *track* the trajectory of the pitches implied by these harmonics the varied mixtures of speakers and tones, and despite their often proximate values, patterns, and even inter-crossings and gaps and other well-known difficulties. These experiments provide evidence of the important cues involved in the remarkable performance of the DNNs, but they also demonstrate their vulnerabilities.

In additional experiments not described here for the sake of brevity, we also observed that the models have difficulties in segregating fricatives and unvoiced segments of speech. Fricatives are plentiful in the training set but are not harmonic in structure, indicating that these

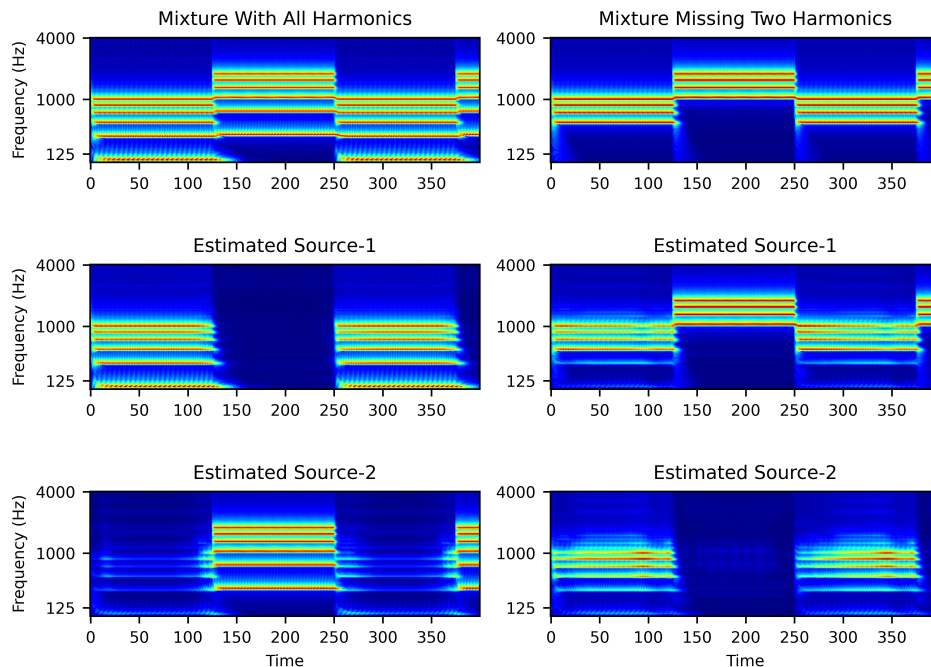


Figure 3.6: Conv-Tasnet successfully segregates mixtures of harmonic sources (left panels). However, it fails when the fundamentals and some lower harmonics are absent (right panels).

vulnerabilities are not a result of testing on ‘out-of-distribution’ data, but instead stem from the underlying mechanisms used by the DNNs to perform the segregation. Finally, we have also confirmed that these same vulnerabilities are not limited to time domain based models trained in an end-to-end fashion but also extend to spectrogram based segregation models.

3.3.3.1 Divergence of DNN Models from Temporal Coherence

Our results also indicate another significant divergence of the DNN-based segregation models from biologically plausible and perceptual models that perform speech segregation using TC. While TC partially leverages pitch information and consequently harmonicity, they primarily rely on the coincidence of the onsets and timing information to group the features emanating from a single source. This divergence between DNNs and such algorithms is evident in Fig. 3.7,

which illustrates that Conv-TasNet can generally segregate two harmonic complexes that have the same exact onset and offset time. Humans and TC algorithms fail to do so because they would group all these components as coming from a single source. Conversely, unlike Conv-TasNet, TC algorithms would have been able to separate the tones in Fig. 3.3.

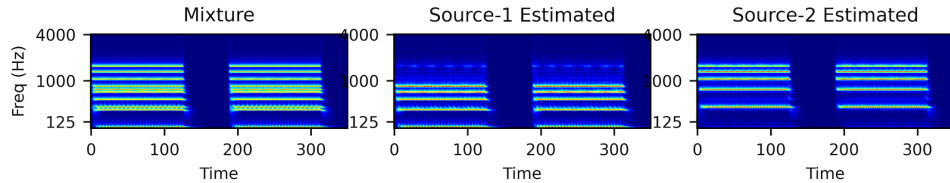


Figure 3.7: Conv-TasNet segregates two harmonic sources even when synchronous.

3.3.3.2 Inharmonic Sources as Adversarial Inputs

We have demonstrated that speech segregation networks perform poorly if an inharmonic speech or tone source is present in the mixture. An adversary can thus generate inharmonic speech which sounds similar to natural speech as an adversarial input to hinder source-segregation and its downstream-tasks. Since the presence of an inharmonic source is independent of the model architecture, the adversarial nature of such sources is not limited to Conv-TasNet and DPT-Net.

3.4 Conclusion

We have shown that DNN-based speech segregation models such as Conv-TasNet and DPT-Net rely on the harmonicity of the sources in the mixture to separate them. They, therefore, fail to segregate sources when their harmonic structure is disturbed, and also find it harder to learn to segregate mixtures with such inharmonic sources. This study demonstrates that these networks implicitly learn to *estimate* pitch by learning the harmonic patterns abundant in the

training corpus. They then proceed to track these harmonic complexes to segregate speech. This effort to understand the underlying principles that these networks learn helps us narrow the gap between our understanding of auditory perception and end-to-end learning-based segregation models.

Chapter 4: Empirical Analysis on the Vulnerabilities of End-to-End Models

4.1 Overview

The sensitivity of E2E speech-segregation models to mixtures of inharmonic speech is demonstrated in the previous Chapter, indicating that these networks are heavily reliant on harmonicity to group and segregate sources. This is corroborated by our observation that despite the abundance of fricatives in the training data, ConvTasnet often fails to segregate fricatives in natural speech as demonstrated in Fig 4.1, although, this error is usually not heard due to the robustness of our speech perception. Appreciating the importance of harmonicity for segregation, we empirically investigate the factors responsible for these models to produce a harmonic analysis of the mixture and subsequently track these harmonic patterns to eventually perform segregation.

The goal of our paper is to analyze the shortcomings of these E2E models to better understand the principles responsible for their success. We believe that this would help us - 1) understand the current limitations of these networks, 2) interrelate traditional Computational Auditory Scene Analysis (CASA) algorithms which rely on engineering [8, 9], biological [2, 7], or pitch tracking [4, 5, 6] based solutions with E2E networks, and 3) improve these networks by training them to be invariant to these shortcomings. We observe that-

- Speech Segregation is more reliant on cues from the lower harmonics of speech than the higher harmonics.

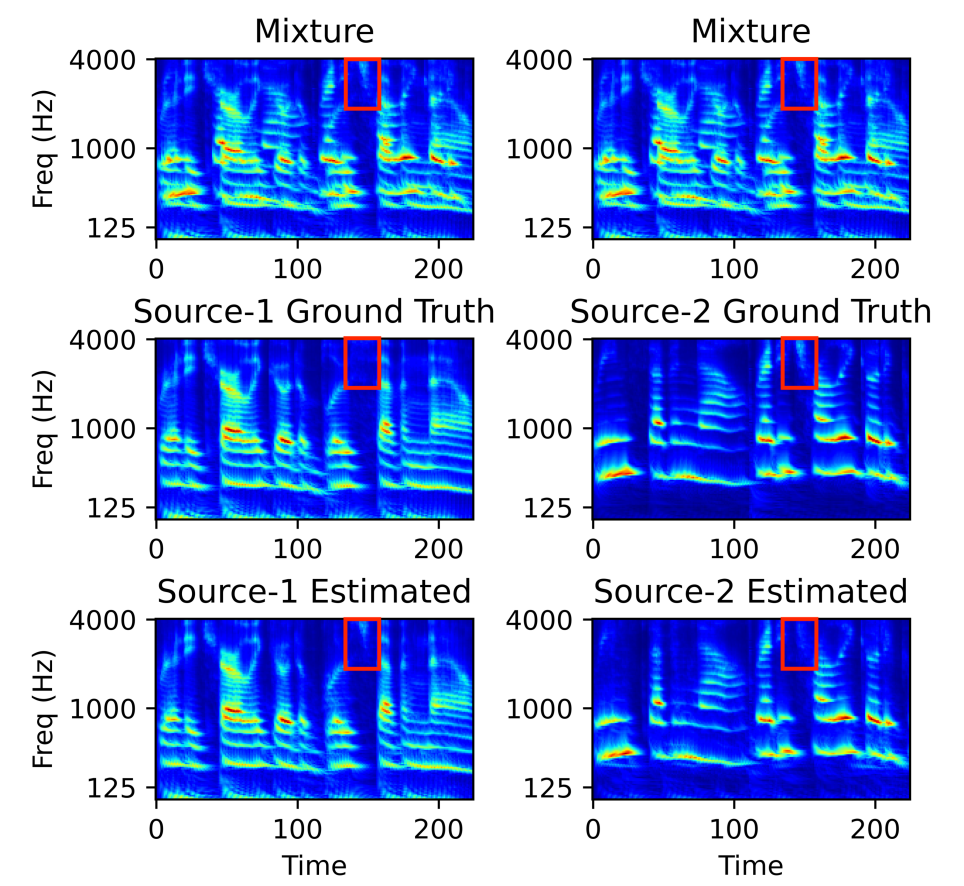


Figure 4.1: ConvTasnet fails to segregate fricatives in natural speech. The red boxes highlight where ‘/s/’ is grouped with Source-1 instead of Source-2.

- E2E source segregation models are very sensitive to a discontinuous harmonic space. These networks fail when any intermediate harmonic is absent.
- Short discontinuities of 20ms in the audio cause a Causal ConvTasnet to perform incorrect assignment near the point of discontinuity
- Spectrogram-based models trained to generate masks directly on the mixture spectrogram are more robust to harmonic deformations and temporal discontinuities.

This work was done in collaboration with Gaspar Rochette.

4.2 Experiments

We perform our analysis on ConvTasnet and DPT-Net trained to segregate 2 speakers. First, we investigate whether the models attend to the higher harmonics which are more prone to overlapping or to the lower harmonics which generally contain more energy. We then analyze the importance of continuity in the harmonic structure for grouping the channels of each estimated speaker. Next, we investigate the importance of temporal continuity to analyze how the models approach the assignment problem of deciding which estimated speaker should be assigned to which channel. Lastly, we contrast these results with speech-segregation networks trained directly on spectrograms [1].

4.2.1 Datasets and Evaluation Metrics

We perform our analysis on ConvTasnet and DPT-Net trained on the WSJ-2-Mix data [11]. We report experiments on both, mixtures of speech from the WSJ-2-Mix test-set and mixtures of non-overlapping alternating tones which are more controllable.

We demonstrate the network’s behaviour visually by generating mixtures of two alternating, non-overlapping harmonic tones, where each tone has N harmonics and is given by $x_i(t) = \sum_{k=1}^N a_k \sin(2\pi k f_0^{(i)} t)$, $i \in \{1, 2\}$. A mixture of these tones containing all the N harmonics can be successfully segregated by the networks. It has been shown in previous work that ConvTasnet relies very strongly on the harmonic structure [26]. Input stimuli of such nature allow us to focus on this harmonic structure and investigate how the network analyses it, by removing certain harmonics such as in Sec. 4.2.2.1 and Sec 4.2.2.2 and precisely inserting discontinuities in Sec 4.2.2.3. To perform our analysis on speech, we filter the validation set of the WSJ-2-Mix dataset

using low-pass, high-pass or band-stop-filters to analyse the network’s performance when missing the lowest, highest or intermediate harmonics. We evaluate the ConvTasnet model using the SDR.

4.2.2 Results

4.2.2.1 Model Performance on Low-Passed and High-Passed Filtered Speech

We investigate whether a network trained on natural speech attends more to the lower or higher harmonics for segregation. We low-pass (LP) filter the WSJ-2-mix dataset with cut-off frequencies at 300Hz, 700Hz, and 1200Hz and evaluate a model with this filtered data to analyze if the network can detect the harmonic structure necessary for segregation [26] in the absence of the higher harmonics. Similarly, we also high-pass (HP) filter the dataset, at cut-off frequencies 180Hz, 300Hz, 400Hz and 700Hz and evaluate the model on this filtered data. We baseline these results against networks trained and evaluated on the corresponding filtered dataset. An alternate baseline is to segregate natural speech using the model trained on natural speech, and filtering the estimated sources. In both cases, we observe performances of around 15db and thus only report the former.

These results are illustrated in Fig 4.2, where it is evident that the network is very sensitive to speech missing its lowest harmonics (e.g. telephonic speech). The network is more robust to the LP filtered data, indicating that the model relies on cues in the lower harmonics more than in the higher harmonics to characterize the incoming mixture.

These results are corroborated by our observations in Fig. 4.3 on mixtures of two alternating tones with F0 as 117Hz and 201Hz. The left column indicates the model’s inability to segregate

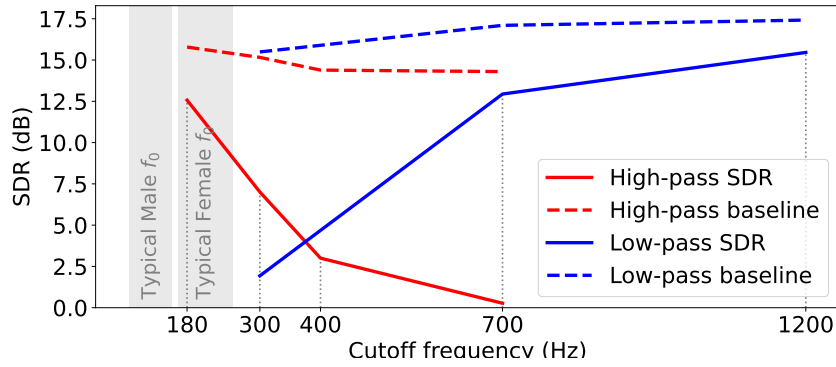


Figure 4.2: Performance of ConvTasnet on either LP (blue) or HP (red) filtered speech for different cutoff frequencies, compared to the performance of networks trained on filtered data.

a mixture of tones with the harmonics: $4, \dots, 8$ and the middle column indicates the model’s ability to segregate a mixture with the lowest three harmonics.

4.2.2.2 Model Performance on Band-Stopped Filtered Speech

We investigate the network’s sensitivity to discontinuities in the harmonic structure. We generate speech missing a few intermediate harmonics by applying a band-stop filter to the mixture before doing the segregation. We perform this evaluation with a set of 8 band-stop filters, with stop-band frequencies ranging from 200Hz-800Hz, to 350Hz-400Hz. We report the results in Fig. 4.4 in which each horizontal line represents a band-stop filter by its low and high cutoff frequencies, and reports the corresponding SDR. As in 4.2.2.1, multiple baselines can be considered which all resulted in performances above 15dB. For clarity, we only report the ConvTasnet native performance of 15.8dB in Fig. 4.4.

We repeat this study on the mixture of tones in Fig 4.3. While the network can segregate a mixture of only the first 3 harmonics (2nd column), it fails when we introduce the 5th harmonic (3rd column). While doing so brings extra information, it is perceived as a missing 4th harmonic which

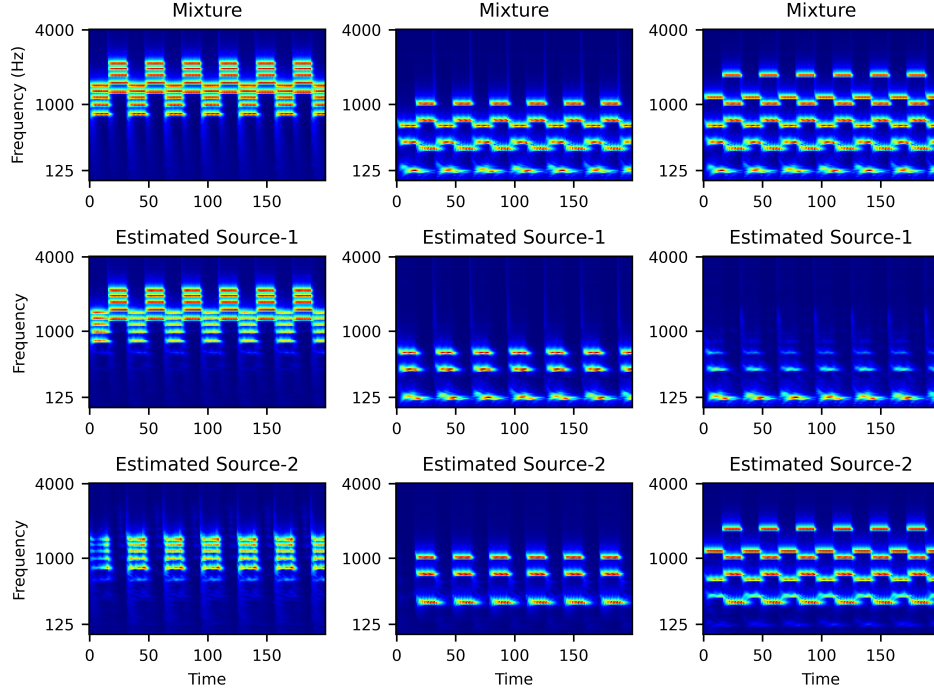


Figure 4.3: Conv-Tanet fails to segregate mixtures of higher harmonics (left columns) but successfully segregates mixtures of the first three harmonics (centre columns). It fails to segregate mixtures of the first three and fifth harmonics (right columns)

challenges the network’s expectations. Unsure of the harmonic structure’s position, the network fails and segregates both tones into the same source, indicating its dependence to a continuous harmonic structure.

4.2.2.3 Performance on Mixtures with Temporal Discontinuities

We study the temporal consistency of the network’s output by introducing short silences in the data. To visualize the quality of segregation at the point of discontinuity, we use a causal ConvTasnet for these experiments. We create mixtures of two alternating and insert silence ranging from 15ms to 100ms since such silence is usually present in speech, on which the network is trained. Fig 4.5 shows our results when tone period is 62ms (i.e. 31ms of energy) and the inserted silence (discontinuities) is 31ms in duration. This silence is inserted to the mixture

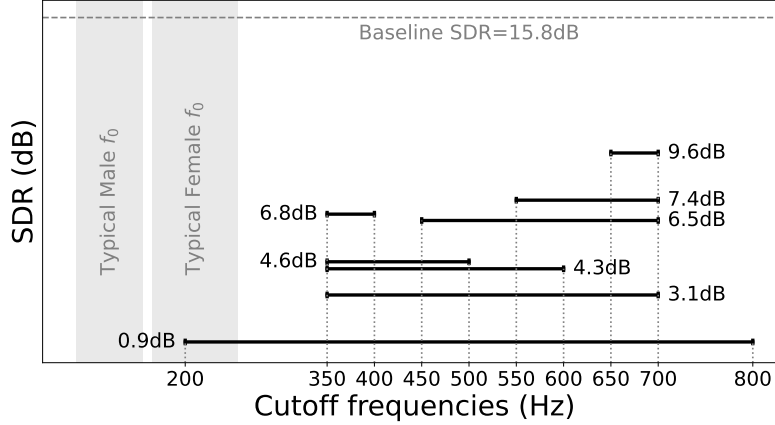


Figure 4.4: Performance of ConvTasnet on band-stopped data. Each filter is a horizontal line representing the frequencies in the stop-band, and the corresponding SDR.

at the points of the arrows in the upper row. The red (respectively green) arrows indicate that the silence has muted the low (respectively high) frequency tone. The corresponding red and green arrows in the mixture spectrograms (upper-rows) indicate the missing tone. We observe that the inserted discontinuity leads to an error in segregation, denoted by the yellow boxes in the estimated spectrograms. These errors are localized to the neighborhood of discontinuity, after which the network begins to segregate the sources correctly. More interestingly, these errors only arise when the low frequency tone is muted (red boxes in the mixture spectrogram). Discontinuities that arise from muting the higher frequency tones do not seem to cause an error in segregation. This observation is consistent across tones of periods ranging from 30ms to 125ms and the duration of the discontinuities as transient as 10ms, repeated using 10 random seeds. This behavior is discussed in Sec 4.2.3.

4.2.2.4 Comparison of E2E Models with Spectrogram Based Models

We analyze the behavior of spectrogram-based speech segregation models on tones with the same harmonic and temporal deformations. We train the separation sub-network of ConvTasnet

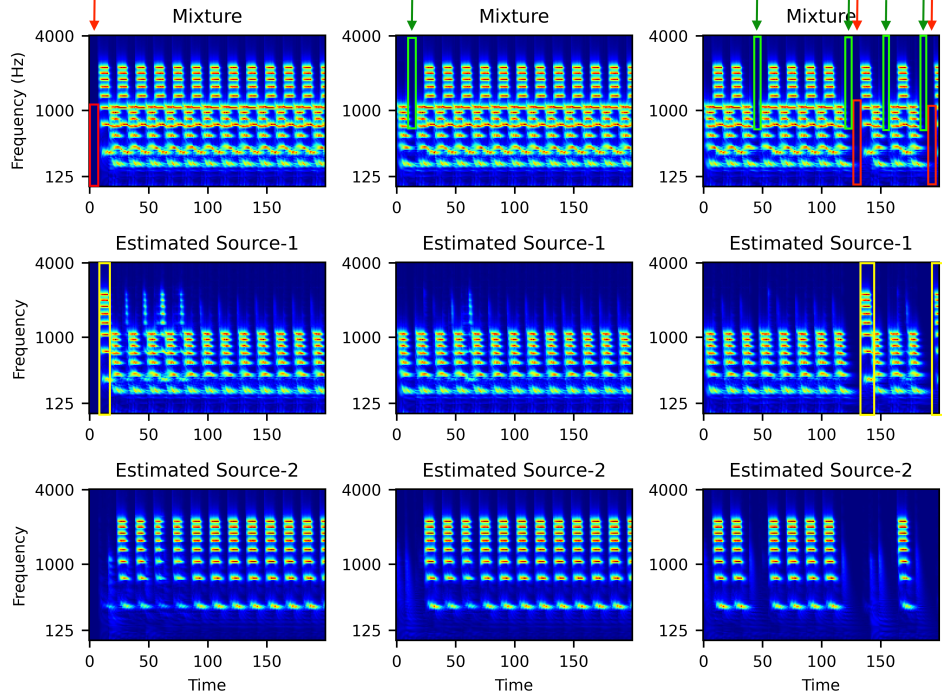


Figure 4.5: Discontinuities are inserted into a mixture of alternating tones (upper rows). Errors in segregation can be observed when these discontinuities mute the lower frequency tone (yellow boxes in the middle rows).

to directly generate masks on the spectrogram. This network is trained to optimize the L2 norm between the estimated and true spectrograms using PIT and is also trained on the WSJ-2-mix dataset. It performs equivalently to a state-of-art spectrogram-based model [17] on natural speech. As illustrated in Fig 4.6, when presented with the same stimuli as that in Fig 4.3, we observe that the spectrogram based model is not sensitive to missing harmonics tests. It can successfully segregate mixtures of just the higher harmonics (left column), lower harmonics (centre column), and mixtures with missing harmonics (right column). We also analyze the sensitivity of spectrogram-based models to temporal discontinuities and present the model with the same stimuli discussed in Sec 4.2.2.3. Fig 4.7 illustrates that the spectrogram based-model is robust to these deformations and can successfully segregate mixtures with periods of silence.

Our results indicate that spectrogram-based models are significantly more robust to deformations

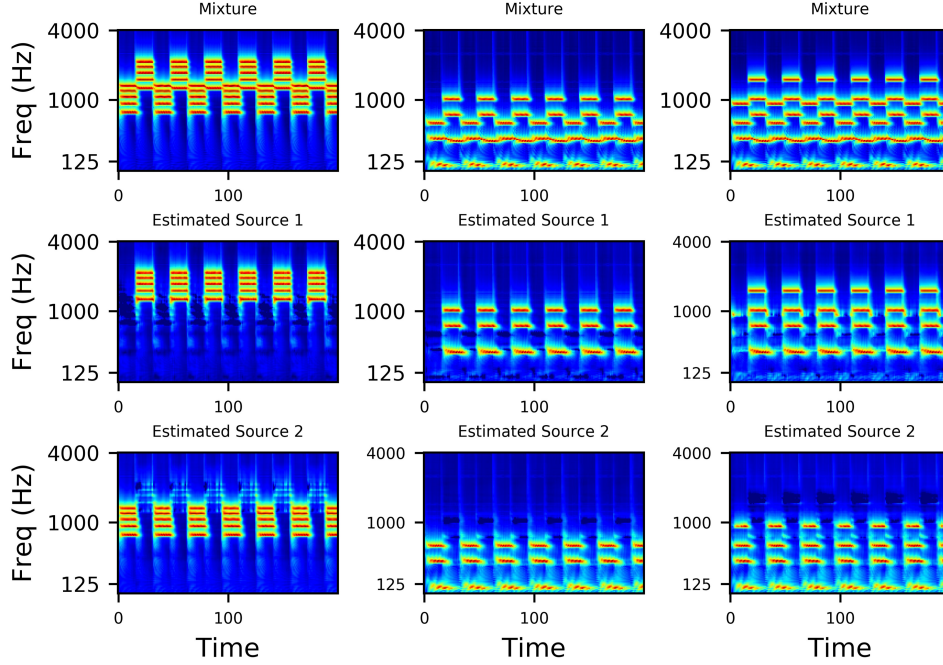


Figure 4.6: A spectrogram-based network can successfully segregate mixtures of the higher (left column), first three (centre column) and first three and fifth (right column) harmonics

in the evaluation data, in-spite of being trained on the same dataset.

4.2.3 Discussion

We have demonstrated that ConvTasnet is very sensitive to transformations on the harmonic structure. Removal of the lower harmonics (HP filtering), higher (LP filtering) or intermediate harmonics (band-stop filtering) often occur in the *wild*. These transformations result in an imperceptible deformation to humans but have a catastrophic impact on the network. We repeat our experiments on DPT-Net and observe similar results.

4.2.3.1 Sensitivity of Networks to Deformations in Harmonics

Section 4.2.2.1 indicates that E2E segregation networks cue on to the lower harmonics to characterize the harmonic structure. A possible explanation for this is that the lower harmonics

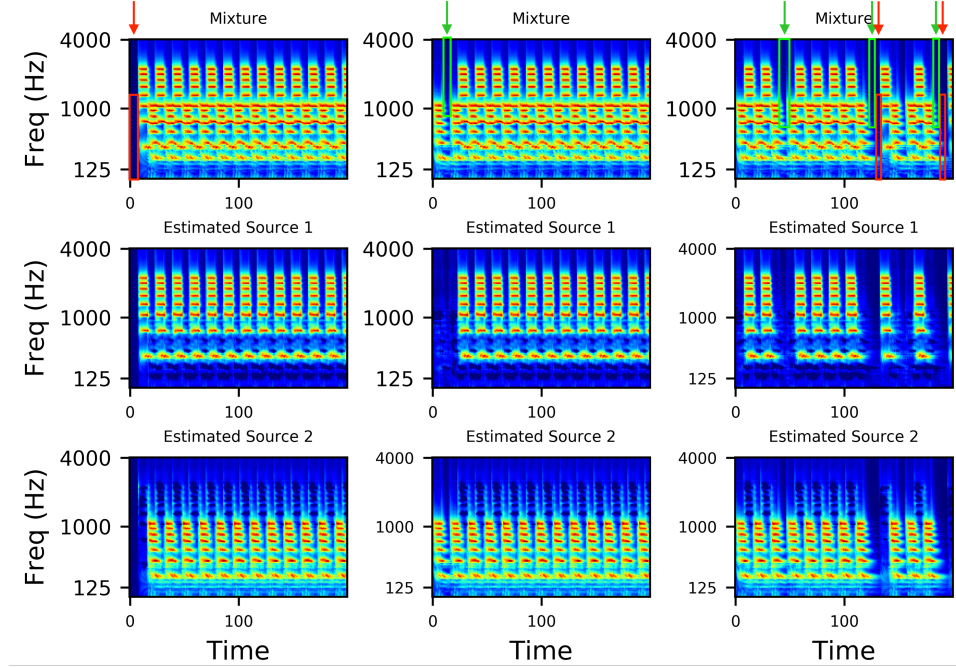


Figure 4.7: A spectrogram-based model can successfully segregate mixtures with inserted silences which are denoted by red and green arrows in the upper rows.

have more energy and are often well separated, even in a mixture. Computing the F0 from the higher harmonics is less trivial. For example, energy at 250Hz could either be the F0 of a female speaker, or the second harmonic of a male speaker whose F0 is 125Hz. However, energy at 900Hz could be the 4th or 5th harmonic of a female speaking at respectively 225 or 180Hz, or the 6th, 7th, ..., or 10th harmonic of a male speaking at respectively 150Hz, 128Hz, ..., or 90Hz. However, the collection of higher harmonics are perceptually important since humans can segregate speech composed of just the higher harmonics and compute the F0 [27]. Our experiments in Sec 4.2.2.2 show that even removing a small frequency band from 350 – 400Hz or 650 – 700Hz considerably degrades the network’s performance, from above 15dB to 6.8dB and 9.6dB respectively. This effect is stronger when removing lower frequencies than higher frequencies, and when increasing the range of the stop-band.

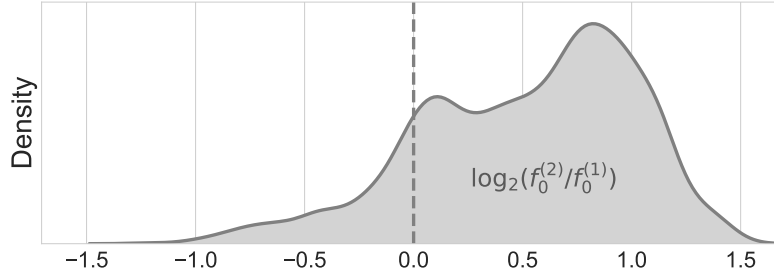


Figure 4.8: Density over evaluation mixtures of the ratio $\log_2(f_0^{(2)}/f_0^{(1)})$ where $f_0^{(1)}$ and $f_0^{(2)}$ are the average F0s of ConvTasnet’s output channels 1 and 2 respectively.

4.2.3.2 Sensitivity of Networks to Temporal Discontinuities

Our experiments in Sec 4.2.2.3 demonstrate that E2E models quickly lose track of the speaker identities after short silences. This may be due to the difficulty of the assignment problem: even after performing speaker grouping, the model needs to decide at every time frame which speaker it should assign to a given output channel. This has been addressed in the past as the label permutation problem - networks are now trained using utterance-level permutation invariant training (U-PIT) [17, 28], which makes the loss invariant to the order of speakers in the output. However, although the chosen order does not affect the loss function, the network still needs to assign each speaker to a channel in a way that is stable over time. In feed-forward networks, this is decided without any information about past assignments. An option is to compute a statistic that can reliably order the speakers in a temporally consistent way. We show in Fig. 4.8 the distribution of $\log_2(f_0^{(2)}/f_0^{(1)})$ where $f_0^{(i)}$ is the average F0 estimated from the output channel i . This distribution is computed using the WSJ-2-mix test-set. It is strongly biased towards positive values, with one peak just above 0 and one near 1, corresponding to mixtures of speakers with the same or different genders respectively. We observe that for 86% of the mixtures, the lower frequency speaker is assigned to channel 1. The network seems to solve the

assignment issue by ordering the speakers’ fundamental frequencies. Interestingly, this learned trait is in alignment to pre-defined conventions used in early segregation models [29], without it being explicitly constrained during training. When unsure, the networks seem to assign almost all the energy to the channel 1, explaining the behavior at points of temporal discontinuity. Thus, they rely on the self-consistency of the speakers to perform segregation.

4.3 Conclusion

Our work investigates the performance of E2E speech segregation networks to data with harmonic deformations and temporal discontinuities. Given their poor performance for speech missing the lower and intermediate harmonics, we believe that these models strongly rely on cues from the lower harmonics to detect a continuous harmonic structure. The network’s sensitivity to silences in the mixtures indicates that it enforces consistency of its output over a time window much shorter than its perceptive field. The network assigns speakers to different channels by comparing short-term features (*e.g.* pitch) and is only consistent over time because of the data’s consistency. We also demonstrate that spectrogram based models are more robust to temporal and harmonic deformations, indicating that ConvTasnet’s sensitivity to these deformations is not a result of a dataset that prevents *out-of-distribution* generalization but rather a consequence of the learned time-frequency representation. Given that the performance of E2E models have surpassed human performance in ideal conditions, their next challenge is to generalize to speech with such deformations. We believe that using transfer learning [30] or expert-knowledge in designing representations that are more invariant to common deformations, and data-augmentation during training may help bridge this gap in robustness.

Chapter 5: Summary and Future Work

5.1 Summary

In this dissertation we first review a biologically plausible speech segregation algorithm that is based on the principle of Temporal Coherence. This algorithm mimics the human-brain and its hypothesised approach towards solving the cocktail-party problem, in that it uses timing cues in an unsupervised and online fashion. However (like humans), this algorithm has shortcomings and is unable to perform ideal speech segregation when presented with a single-channel speech mixture. We find that such an algorithm is instead more suitable for a speech de-noising task in its current state.

We then investigate DNN based speech segregation algorithms that are found to out-perform humans and the IRM. Despite their remarkable performance, the underlying principles used by these networks to perform segregation is largely unexplored. Performing our analysis on Conv-Tasnet and DPT-Net, we demonstrate that these networks largely depend on harmonicity in speech to segregate the sources, and that they are especially vulnerable to inharmonic sources. We also show that these networks find it difficult to learn segregation cues when trained on inharmonic speech, underscoring the importance of harmonicity for DNN based segregation networks. Additionally, we also demonstrate that these networks diverge from humans and the discussed Temporal Coherence algorithm and are do not use timing information as a primary cue

for segregation. We believe that these networks implicitly learn to track the harmonic structure of speech which is abundantly present in the training data-set. It thus learns to perform a non-trivial task that was foundational to CASA algorithms e.g. pitch-tracking algorithms. We believe that the ability of the DNN to perform this task is responsible for its superior performance on natural speech but also vulnerable to any perturbations to the natural harmonic structure of speech, e.g. in fricatives.

Lastly, we investigate other vulnerabilities in end-to-end speech segregation networks to data with harmonic and temporal discontinuities. We demonstrate that these models are very sensitive to discontinuities in the harmonic structure of speech and completely fail to segregate speech with missing harmonics, e.g. band-stopped filtered speech. We also demonstrate these networks are more reliant on the energies in the lower harmonics to perform tracking than from the higher harmonics. Interestingly, we demonstrate that these networks are locally vulnerable to very short silences and discontinuities, indicating that the network enforces consistency of its output over a time window much shorter than its perceptive field. We also observe that the network solves the ‘assignment problem’ by comparing short-term features such as *pitch*. We also demonstrate that spectrogram based models are more robust to temporal and harmonic deformations, indicating that ConvTasnet’s sensitivity to these deformations is not a result of a dataset that prevents *out-of-distribution* generalization but rather a consequence of the learned time-frequency representation.

5.2 Future Work

5.2.0.1 Binaural Source Segregation Using Temporal Coherence

Given the challenges associated with data-limitations in monaural source segregation using Temporal Coherence, a promising next-step is to investigate binaural source segregation, wherein the spatial location of the sources can be used by comparing the relative delays between the two channels. This would allow the location to be used as an additional anchor along with the pitch. Furthermore, the use of two microphones would immediately double the number of features usable to train the network, potentially addressing the problems associated with lack of training data

5.2.0.2 Improving DNN Based Monaural Speech Segregation

Given that the performance of E2E models have surpassed human performance in ideal conditions, their next challenge is to generalize to speech with harmonic deformations. We are confident that using transfer learning [30], pre-training or expert-knowledge in designing representations that are more invariant to common deformations will enhance segregation performance. While data-augmentation techniques like SpecAugment [31] have shown promising results in other speech and audio processing tasks [32], these methods have not been investigated in speech segregation. Training the models on mixtures with spectral and temporal discontinuities such as in SpecAugment will make it more robust.

5.2.0.3 Training DNN Models Without Permutation Invariant Training

We demonstrate that DNN models solve the assignment problem using local features such as pitch. We believe that knowledge of how the network assigns an estimated speaker to a channel will help us circumvent the permutation problem. Not having to compute all possible permutations of speaker assignment can drastically reduce the computational demands of training these models.

Bibliography

- [1] Taishih Chi, Powen Ru, and Shihab A Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.
- [2] Lakshmi Krishnan, Mounya Elhilali, and Shihab Shamma. Segregating complex sound sources through temporal coherence. *PLoS computational biology*, 10(12):e1003985, 2014.
- [3] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.
- [4] Srikanth Vishnubhotla and Carol Y Espy-Wilson. An algorithm for speech segregation of co-channel speech. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, page 109. IEEE, 2009.
- [5] Michael Stark, Michael Wohlmayr, and Franz Pernkopf. Source–filter-based single-channel speech separation using pitch information. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):242–255, 2010.
- [6] DeLiang L Wang and Guy J Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE transactions on neural networks*, 10(3):684–697, 1999.
- [7] Mounya Elhilali and Shihab A Shamma. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6):3751–3771, 2008.
- [8] Ron J Weiss and Daniel PW Ellis. Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech & Language*, 24(1):16–29, 2010.
- [9] Martin Cooke, John R Hershey, and Steven J Rennie. Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1–15, 2010.
- [10] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015.

- [11] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.
- [12] Xiao-Lei Zhang and DeLiang Wang. A deep ensemble learning method for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(5):967–977, 2016.
- [13] Yi Luo and Nima 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)*, pages 696–700. IEEE, 2018.
- [14] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.
- [15] Yi Luo, Zhuo Chen, and Takuya 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)*, pages 46–50. IEEE, 2020.
- [16] Jingjing Chen, Qirong Mao, and Dong Liu. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. *arXiv preprint arXiv:2007.13975*, 2020.
- [17] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017.
- [18] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.
- [19] Ilya Kavalero, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 175–179. IEEE, 2019.
- [20] H Vincent Poor. *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.
- [21] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey. Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*, 2016.
- [22] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

- [23] Sara Popham, Dana Boebinger, Dan PW Ellis, Hideki Kawahara, and Josh H McDermott. Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature communications*, 9(1):1–13, 2018.
- [24] Daniel PW Ellis, Josh H McDermott, and Hideki Kawahara. Inharmonic speech: a tool for the study of speech perception and separation. 2012.
- [25] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.
- [26] Rahil Parikh, Ilya Kavalero, Carol Espy-Wilson, and Shihab Shamma. Harmonicity plays a critical role in dnn based versus in biologically-inspired monaural speech segregation systems. *arXiv preprint arXiv:2203.04420*, 2022.
- [27] Shihab Shamma and David Klein. The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *The Journal of the Acoustical Society of America*, 107(5):2631–2644, 2000.
- [28] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.
- [29] Chao Weng, Dong Yu, Michael L Seltzer, and Jasha Droppo. Deep neural networks for single-channel multi-talker speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(10):1670–1679, 2015.
- [30] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- [31] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.
- [32] Anthea Cheung, Qingming Tang, Chieh-chi Kao, Ming Sun, and Chao Wang. Improved student model training for acoustic event detection models.