EXTRACTING SINGLE TALKER SEGMENTS FROM AUDIO MIXTURES IN REVERBERANT ENVIRONMENTS

BY

AVINASH SUBRAMANIAM

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Adviser:

Professor Romit Roy Choudhury

Abstract

Classifying the number of simultaneous speakers in a reverberant environment remains a difficult problem to solve. The less complicated but no less important problem of detecting whether a single speaker or multiple speakers are present also poses a challenge, as multipath often distorts or alters the features commonly used in speaker number classification. When this classification system is used as a front-end to a learning-based system, then it becomes imperative that the classifier be accurate.

This thesis presents *SinguDetect*, an end-to-end system which uses the inter-aural phase differences between a pair of in-ear microphones to classify each 64 ms time window of audio as belonging to one speaker or not. Even in heavily reverberant environments (RT60 = 2.0 s), *SinguDetect* is still able to maintain a precision of around 0.8 where the number of simultaneous sources is not higher than three, whereas other state-of-the-art algorithms tend to suffer decreases in precision and f-score in this range.

To my family and friends.

Acknowledgments

This work has been done under the guidance of Professor Romit Roy Choudhury. I would like to thank him for all the valuable feedback and discussions throughout the course of my studies. Thanks to the University of Illinois Electrical and Computer Engineering Department for providing me with a teaching assistantship, allowing me to financially sustain myself during my graduate studies. And finally, thanks to my parents and numerous friends who endured this long process with me, always offering support.

Table of contents

Chapter 1

Introduction

Machine learning and artificial intelligence (AI) are fast encroaching onto all fields of research. Using a model-free approach, machine learning is capable of solving many tasks such as classification and regression with just data needed during the learning process — sometimes without the data being explicitly labeled. In light of such advances, there still remains a need to provide formatted data to such algorithms using model-based methods. This thesis's motivating example fits this criteria. In [1], A. Xu and R. R. Choudhury use spatial clustering as a front-end to a neural network, identifying which snippets of speech are suitable to learn from and which are not. Specifically, it selects snippets where only one talker is present, which we will refer to as singular frames, and the property of which we will refer to as the singularness of talkers in speech. The totality of these singular frames is what we refer to as single talker segments. However, the clustering used in [1] degrades heavily in performance in reverberant environments. Therefore, herein we address that shortcoming. *SinguDetect* has been specifically tuned to function in reverberant environments, and performs better than the state-of-the-art algorithms in such conditions. Any system that requires knowledge of the singularness of talkers in a segment of audio can rely on our algorithm in environments of varying reverberation and signal-to-noise (SNR) ratios.

The main contribution of this thesis is a method to detect the singularness of talkers in speech. This detection is on the order of tens of milliseconds and can operate in extremely reverberant environments. With knowledge of these singular frames, this work allows for models that rely on single talker data to be implemented in conversation settings.

Our proposed approach operates on binaural audio, just as its inspiration did. Binaural audio is the most natural format for human audition, as with monaural audio, almost all spatial features are lost, and thus, monaural audio will sound flat or incomplete to most listeners [2]. For classifiers, monaural audio is similarly deficient, as many classifiers exploit the spatial features present in recordings by two or more microphones. *SinguDetect* is one such algorithm, as we will cover in detail in the system design section.

The rest of this thesis is divided into five more chapters. Chapter 2 introduces the background information relevant to understand the subsequent sections, particularly the system design and experiment and results chapters. Specifically, we will go over the signal model for source to sensor channels, an explanation of the feature primarily used for our classifier, the GCC-PHAT algorithm and its relevance to *SinguDetect*, and,

finally, an overview of several competing methods and their limitations compared to our proposed approach. Chapter 3 will cover the design of *SinguDetect*. In particular, we will develop the theory behind each step of the algorithm with special attention to inputs and outputs. Chapter 4 will cover the design of the experiments, which are used to elucidate the performance of the competing methods and our proposed method, and the corresponding results. Chapter 5 discusses the limitations of the current work, as well as possible future work in the same direction as this thesis. Chapter 6 concludes the thesis and examines the implications of this work.

Chapter 2

Background

In this chapter, we introduce the models and formulae necessary for understanding the design of the proposed method. We briefly present several state-of-the-art algorithms and highlight their differences to the presented algorithm and the choices that resulted in these differences.

2.1 The Signal Model

We try to follow the signal model from the work of L. Wang et al [3]. Indeed, the basis for *SinguDetect* is the algorithm presented in [3]. The differences between this work and that paper will be expanded upon in Section 2.3.

Consider $M = 2$ microphones with an inter-microphone distance of d. At any given time instant, we represent the audio received by a single microphone as a mixture of the following description,

$$x_m(n) = \sum_i^N (\mathbf{a}_{m,i} \circledast \mathbf{s}_i)(n) + v_m(n) \tag{2.1}$$

where $\mathbf{x}_m$ corresponds to the signal received at microphone m, n to the time index of the signal, $\mathbf{a}_{m,i}$ to the room impulse response from the m'th microphone to the i'th source, $\mathbf{s}_i$ to the i'th speech source, $\mathbf{v}_m$ to the noise present at the m'th microphone, and $\circledast$ to the convolutional operator. Note that the room impulse response length L_{rir} is typically much shorter than the speech source length, and noise is uncorrelated between the two microphones.

If we analyze the signal received by the microphones using the short-time Fourier transform (STFT) to enter the time-frequency domain, we obtain the following equation.

$$X_m(k,l) = \sum_i^N A_{m,i}(k) S_i(k,l) + V_m(k,l) \tag{2.2}$$

where all upper-case letters correspond to the same meaning as their lower-case counterparts, k refers to the k'th frequency bin of the fast Fourier transform (FFT), and l refers to the l'th time frame. For the purposes of this thesis, the total number of frequency bins will be referred to as D and the total number of frames as L. Equation (2.2) represents a simplified signal model that is mathematically easier to work with, and therefore, it will be the one we will work with going forward. Note that we have made the narrowband transfer function assumption. However, instead of directly dealing with the measured signal, We use the inter-aural phase

difference (IPD) of the microphones. This refers to the phase difference between the signal captured by the first microphone and the second, induced by the specific acoustic transfer functions (ATFs) between the microphones and the sources producing the signal. It is estimated by the following equation, assuming a static (in location) source with respect to a single time frame,

$$\psi(k,l) = \angle \frac{X_2(k,l)}{X_1(k,l)} \tag{2.3}$$

where then angle is derived by taking the inverse tangent of the imaginary part of the signal over the real part [4]. However, there is a major problem with deriving the IPDs in this way. If there are multiple speech sources active in the same time-frequency bin, or the noise is uncorrelated and present at all frequencies, then the IPD will potentially be heavily corrupted and unreliable for further use. These possible speech sources include echoes or reverberations, since these arrive at a different angle of arrival (AoA) than the direct path from the source to the microphone. Nevertheless, as we will see in the results, if there is only one source that is active for the duration of the time frame, we can usually derive the correct AoA for that source. Although the resulting signal will be noisy, we will be able to fit linear curves to it, resulting in a fairly accurate angle estimate.

In an anechoic scenario, our IPD simplifies further, as illustrated in Figure 2.1. Our phase difference becomes a function of the time difference of arrival (TDoA) τ between the same signal entering the first and second microphones and the specific frequency of sound, as shown below [3].

$$\psi_s(k,l) = 2\pi f_k \tau_s + 2\pi p_k \tag{2.4}$$

where f_k refers to the frequency in Hz corresponding to the k'th frequency bin of the FFT, τ_s to the TDoA for source s, p_k to the phase wrapping term for the k'th frequency bin, and ψ_s to the IPD for source s. Here, we clearly obtain a linear relation between frequency and the IPD — the only confounding variable is the phase wrapping term p_k. This wrapping term is also linearly dependent on frequency, and corresponds to the following: $|2\pi f_k \tau_s|_{2\pi}$ [3]. We have the same terms from before, only $|\cdot|_{2\pi}$ refers to the integer $2\pi f_k \tau_s \bmod 2\pi$. To mimic human audition, binaural microphones are often placed within a distance typically around 22 cm of each other [5], so in these cases the τ_s factor is not too large. However, at higher frequencies — especially at larger AoAs — phase wrapping frequently occurs. This, along with the problems of noise, reverberation, and multiple simultaneous sources, will pose a problem for us later on.

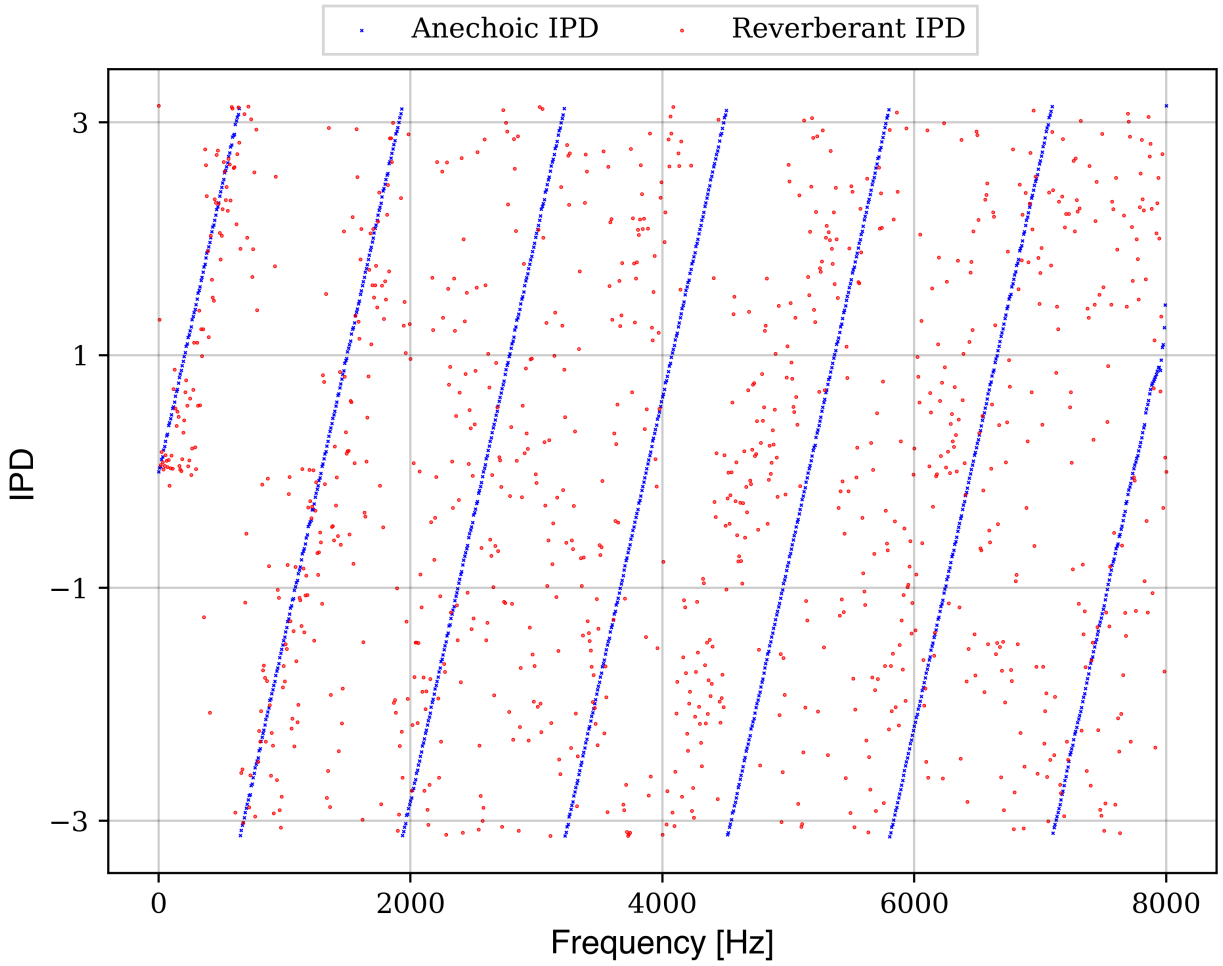

Figure 2.1: Illustration of an anechoic IPD vs. a reverberant IPD

2.2 Generalized Cross-Correlation Phase Transform

With the IPDs being noisy due to a number of factors listed in Section 2.1, it is more efficient to have a hypothesis of what the IPD line should be. That is, rather than attempting to discern structure from the IPD itself, we can formulate some idea of what the IPD should look like and check how close this hypothesis is for classification. Such a hypothesis will be more akin to the blue trend line of Figure 2.1 rather than the red. In [3], L. Wang et al. make the argument that the GCC-PHAT algorithm [6] can be modified to find each τ_s from each source in a multi-source environment. They prescribe the function,

$$R_w(\tau) = |\sum_{k,l} W_{TF}(k,l) e^{j\psi(k,l)} e^{-j2\pi f_k \tau}| \tag{2.5}$$

where W_{TF} refers to a time-frequency weighing factor dependent on an SNR weight, a coherence weight, and a cancellation weight, all of which will be explained in Chapter 3. The peaks of the GCC correspond to the most likely AoAs, although discerning the direct path AoAs from the multipath ones is quite challenging.

However, assuming there is only one source, and that the near path signal is the loudest, the largest GCC peak should correspond to the direct path source AoA [4], a concept depicted in Figure 2.2. Therefore, we use the τ derived from the largest GCC peak as the hypothesis to fit an IPD line. We will expand on this concept in Chapter 3 when we go over the central algorithm of the thesis.

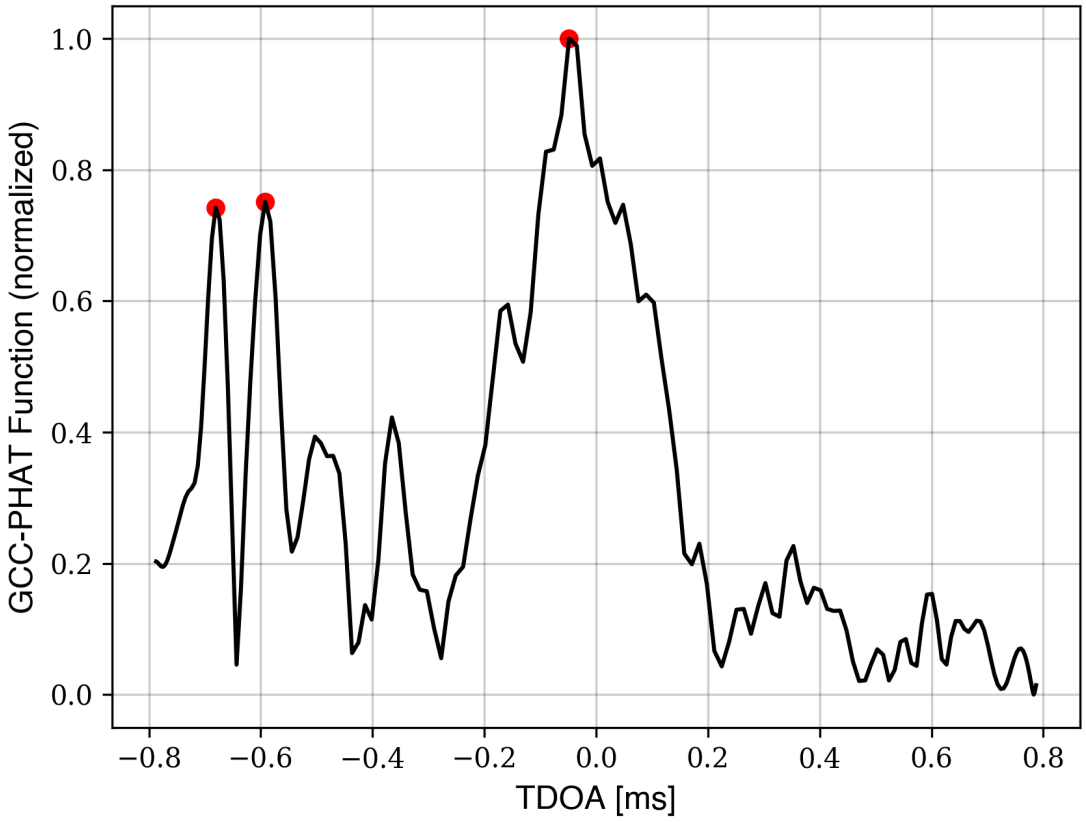

Figure 2.2: GCC-PHAT function of a single speaker in a heavily reverberant environment. The red dots indicate likely source TDoAs, with the rightmost peak being the correct one. This function was calculated over 10 frames of the STFT.

2.3 Competing Methods for Singularness Detection

We now begin briefly covering competing methods for detecting the singularness of talkers in speech. The first method is "Source Counting and Separation Based on Simplex Analysis" [7] by B. Laufer-Goldshtein et al. The second method is "Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking of Fundamental Frequency" [8] by A. O. T. Hogg et al. And the final method we will cover in this section is the inspiration behind the algorithm for *SinguDetect*, "An Iterative Approach to Source Counting and Localization Using Two Distant Microphones" by L. Wang et al [3].

2.3.1 Source Counting and Separation Based on Simplex Analysis

B. Laufer-Goldshtein et al. implement a statistical method to estimate the number of sources using data from many different microphones. The crux of their paper is that the correlation between two instantaneous relative transfer functions (RTFs) $\tilde{\mathbf{W}}(l)$ and $\tilde{\mathbf{W}}(n)$, where each RTF is a D dimensional vector corresponding to the frequency bins of a single time frame of index either l or n, is

$$E\left\{\frac{1}{D}\tilde{\mathbf{W}}^T(l)\tilde{\mathbf{W}}(n)\right\} = \begin{cases} \sum_{i=1}^{N} p_i(l)p_i(n) & \text{if } l \neq n \\ 1 & \text{if } l = n \end{cases} \tag{2.6}$$

where $p_i(l)$ represents the probability of a source i being represented in a single time-frequency bin in the observation at frame l.

The relative transfer function is defined as the ratio of acoustic transfer functions (ATFs) between two sensors [9], which in our case would be $\frac{\mathbf{A}_{2,i}}{\mathbf{A}_{1,i}}$. Acquiring ATFs is difficult in practice, so [7] estimates the instantaneous RTF through the following equation

$$\tilde{W}(k,l) \equiv \frac{\hat{\Phi}_{y^2y^1}(k,l)}{\hat{\Phi}_{y^1y^1}(k,l)} \equiv \frac{\sum_{n=l-T/2}^{l+T/2} X_2(k,n)X_1^*(k,n)}{\sum_{n=l-T/2}^{l+T/2} X_1(k,n)X_1^*(k,n)} \tag{2.7}$$

where $\hat{\Phi}_{y^2y^1}$ refers to the estimate of the cross power spectral density of the received signal, $\hat{\Phi}_{y^1y^1}$ to the estimate of the power spectral density of the first microphone, and T to the size of the window to compute the RTF.

Across all sources in a single time frame, probabilities sum to 1: $\sum_i^N p_i(l) = 1$. Each RTF $\tilde{\mathbf{W}}^T(l)$ can be thought of as the result of D trials, where the outcome of each trial is the time-frequency bin of one of the unknown RTFs $\frac{\mathbf{A}_{2,i}}{\mathbf{A}_{1,i}}$. Therefore, [7] assumes the W-disjoint orthogonality of speech, where each time-frequency bin is only composed of a single source. That is, there is no mixture of sources within each time-frequency bin.

From this formulation, B. Laufer-Goldshtein et al. make the argument that the correlation matrix, of shape $L \times L$, where L is the number of time frames, can be decomposed using the eigendecomposition, and that the resulting N eigenvectors corresponding to the N largest eigenvalues form a simplex that corresponds to the probability of activity of each source along the observation index $1 \leq l \leq L$ [7]. They estimate the number of sources N by finding the number of eigenvalues larger than a certain scalar multiple of the largest eigenvalue. This scalar's value can be varied to elicit better performance in certain scenarios and is an important hyperparameter. Through mathematical process, they produce an estimate of the matrix $\mathbf{Q}$ such that the inverse of this matrix multiplied by the eigenvectors recovers the original source probabilities, as shown below.

$$\hat{\mathbf{p}}(l) = \hat{\mathbf{Q}}^{-1}\boldsymbol{\nu}(l) \tag{2.8}$$

Here, $\hat{\mathbf{p}}(l)$ is the estimate of the source probabilities, $\hat{\mathbf{Q}}^{-1}$ is the aforementioned $\mathbf{Q}$ matrix, and $\boldsymbol{\nu}(l)$ is the matrix of N eigenvectors of the correlation matrix corresponding to the N largest eigenvalues. With these estimated probabilities, we can theoretically fulfill the goals of *SinguDetect*, as we can go frame by frame and determine whether the frame is probabilistically dominated by any source. The authors of [7] decide this by setting a parameter β, and declaring a frame to be a single speaker frame if any value in the vector $\hat{\mathbf{p}}(l)$ is greater than β. The value of β in the paper is 0.9, which is a logical choice as that would indicate the frame

heavily dominated by a single speaker. In this thesis's implementation of the paper's algorithm, β was set to 0.6 instead of 0.9 because it was found this value worked better for the considered experimental conditions.

Why did we choose the IPD based approach over this correlation matrix based approach? Firstly, the signal model has some weaknesses. It assumes that the probability is fixed per frame — every time-frequency bin across the same time frame has an independent probability of belonging to each source. However, this is not necessarily true, as different talkers are likely to have different pitched voices. Thus, frequencies around the pitch and harmonics of a talker are more likely to belong to that talker. This weakens the convergence of Equation (2.6), as there would not be a single probability for each talker for every time frame — more probabilities are required for separate frequency bins as well. Secondly, the W-disjoint orthogonality assumption is sometimes violated in practice [10]. Again, this weakens the signal model as the RTF's time-frequency bins $\tilde{W}(k,l)$ cannot be represented merely by a single source RTF's corresponding time-frequency bin . Finally, the most important reason is that the correlation matrix is difficult to reckon with. Although B. Laufer-Goldshtein et al. estimate the probability vector, and use that as the feature for classification, any improvements in the classification algorithm have to tackle the estimation of the $\mathbf{Q}$ matrix. Thus, any tweaks or improvements would be non-trivial in nature. In contrast, IPDs are easy to digest visually, and, thus, any classifier built on top of them can be improved upon more easily and transparently — i.e. we can tell why a proposed improvement actually effects an increase in accuracy.

2.3.2 Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking of Fundamental Frequency

As opposed to using the correlation matrix as the central feature in their algorithm, A. O. T. Hogg et al. use pitch. To detect pitches from multi-speaker data, the paper uses PEFAC [11]. However, PEFAC produces a number of possible pitches and harmonics without delineation between them at each time step, meaning we do not know which values correspond to the pitches and which to the harmonics of one of these pitches. Therefore, in order to narrow down the number of candidate pitches and harmonics, the paper produces a set of non-overlapping pitches and their respective harmonics for every frame of data [8]. These pitches are determined as the lowest possible values — above and below a certain threshold F_{min} and F_{max} — that have at least one other value in the set close to an integer multiple of itself. These integer multiples (within a tolerance) of the proposed pitch are the hypothesized harmonics belonging to that pitch. With the set of possible pitches and their associated harmonics, a simple iterative pruning process is conducted per time step. First, the longest observation, that is the pitch with the highest number of associated harmonics, is added to the final set, and any other pitches with harmonics that intersect with this observation are removed from consideration [8]. Then, this step is repeated until no candidates are left, and the final set containing non-overlapping pitches and harmonics is produced.

Then, the paper moves onto using Kalman filters and multiple hypothesis tracking. For each time instant, usually 10 ms in time, pitches and their harmonics are tracked. Depending on past values, overlapping observations, etc. tracks are either terminated or propagated [8]. The end result is an estimate of the number of speakers, given by the number of 'alive' tracks, at each time instant. Therefore, unlike part of [7], this work is not truly a detector of whether there is one speaker or multiple, but rather of the number of speakers total in each time instant. Nevertheless, we can clearly and easily change it into such a detector.

This pitch based approach was not taken for two reasons. Firstly, PEFAC has empirically produced pitches and harmonics that either do not correspond to any speaker or have to be mapped to a speaker already responsible for another set of pitches and harmonics. This derails the hypothesis tracking process, as

the cardinality of tracks is taken as the number of speakers per time instant. Secondly, and related to the first point, is that the performance of the algorithm in heavily reverberant environments is rather unreliable, as we will see in the results section. It seems that pitch detection is tricky in these environments. For example, in single speaker scenarios, multiple simultaneous sets of pitches and harmonics have been produced where there should only be a single set. Therefore, it is impossible for any downstream algorithm to rectify this. There is no way to differentiate the 'real' set from the false ones, if such a delineation even exists, and correctly classify the frame as singular as a result.

2.3.3 An Iterative Approach to Source Counting and Localization Using Two Distant Microphones

Finally, we move onto the paper that provides the basis for this thesis. L. Wang et al. use the IPD as their primary feature for classification. They use the GCC-PHAT to produce a trend line for the IPD using the derived τ corresponding to the largest GCC peak similar to what was alluded to before. Here, the major difference is that for each iteration of the algorithm, the trend line is potentially shifted by an offset. This offset is decided according to a loss criterion, defined below.

$$\rho'(k,l,\tau_q) = |\angle e^{j(\psi(k,l)-(2\pi f_k \tau_q + \delta'_q))}| \tag{2.9}$$

where τ_q represents the TDoA from the q'th iteration's GCC-PHAT peak, and δ'_q refers to the shift in the trend line $2\pi f_k \tau_q$. With this loss function, the optimal shift is defined as the following

$$\delta_q = \arg \min_{\delta} \sum_{k,l} |\angle e^{j(\psi(k,l)-(2\pi f_k \tau_q + \delta))}| \tag{2.10}$$

The paper mentions that an exhaustive search is conducted for the value δ_q in the range $[-\pi/3, \pi/3]$. Searching for this offset is important, because as Figure 2.1 and Figure 2.3 illustrate, the IPD often deviates away from the hypothesized trend line.

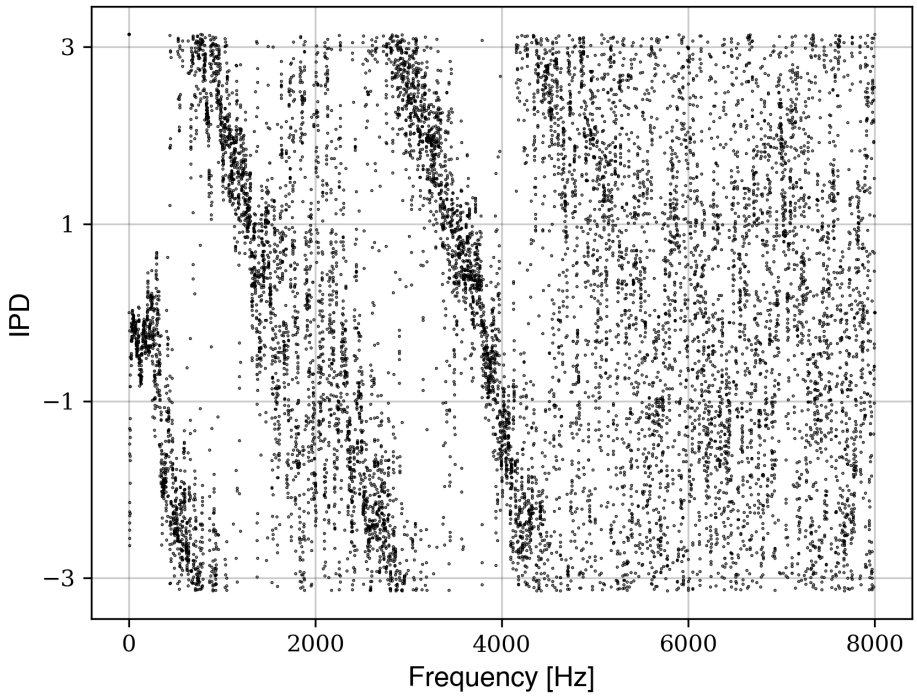

Figure 2.3: IPD of a single speaker in a heavily reverberant environment. This IPD is a composite of 5 consecutive frames of the same speaker.

Upon discerning the optimal offset, the algorithm then alters the weight of W_{TF} from Equation (2.5) to have zeros where the time-frequency bins of the IPD are sufficiently within a threshold ρ_{TH} of the potentially shifted trend line $2\pi f_k \tau_q + \delta_q$, and moves onto the next iteration. Every time frame is considered for the GCC calculation (see Equation (2.5)), and the time-frequency bins are cancelled across all time frames as well.

The stop criterion for this process is three sided. First, if the number of time-frequency bins that are not zeroed out is 1% of the total number of time-frequency bins, then the iteration stopped. Alternatively, if the kurtosis value of the GCC-PHAT function is smaller than the threshold K_{TH}, then the process is also halted. Finally, if the number of iterations is equal to ten, the process stops. Every unique TDoA that the GCC-PHAT produces during these iterations is recorded as belonging to a source. If two hypothesized sources have TDoAs that are relatively close to each other, according to the following equation, they will be merged.

$$|\tau_p - \tau_q| < \frac{dsin(A_{min})}{c} \tag{2.11}$$

where A_{min} is a minimum separation angle, defined as 10° by the paper, τ_p and τ_q are the TDoAs produced by GCC-PHAT for two separate iterations and theoretical sources p and q respectively, d is the separation distance between microphones (typically around 22 cm for binaural microphones), and c is the speed of sound, which will be 343 m/s in this thesis, corresponding to its speed in air at 20°C [12]. If this criterion is met,

then the TDoAs will be merged according to which GCC-PHAT peak was larger, that is

$$\tau_m = \begin{cases} \tau_p & \tilde{R}_o(\tau_p) > \tilde{R}_o(\tau_q) \\ \tau_q & \text{otherwise} \end{cases} \tag{2.12}$$

where $\tilde{R}_o(\tau_p)$ refers to the GCC-PHAT function's value at τ_p during the p'th iteration. See Equation (2.5) for the details of the implementation of GCC-PHAT here. The cardinality of the final set of TDoAs is taken as the number of sources in the particular audio file presented to the algorithm.

This algorithm suffers from a few weaknesses with respect to singularness detection, the solution to which is the main contribution of this thesis. Firstly, the algorithm uses every time frame in the STFT of the observation for its operation. This severely reduces the granularity of the source counting mechanism — it cannot classify whether single frames are dominated by a single source. If the number of time frames is restricted to provide a sense of granularity, then further issues occur. In particular, with 10 frames of data for example, the kurtosis of the GCC-PHAT function as part of the stop criteria of the algorithm was empirically found to be unreliable given heavily reverberant data. Oftentimes, either the algorithm terminated too early, when there were a greater number of sources than the number of iterations conducted, or the algorithm hit the iteration limit because the kurtosis did not decrease below the threshold. With the algorithm stopping prematurely, the correct number of sources may not be detected. For the purposes of singularness detection, if only one source is detected when there are in fact many, an error will occur.

Additionally, merely shifting the IPD trend line may not be sufficient to capture the perturbations reverberation produces in the IPD, as the paper posits [3]. We can see this potentially occurring in Figure 2.1. If we cancel out enough time frequency bins according to one trendline, we may not be able to recover the same TDoA from the GCC-PHAT peaks. This would prevent us from discovering the trend that starts at 4000 Hz for example in the figure. Thus, we will effectively miss some time-frequency bins that belong to the same trend line derived from a certain TDoA that will affect the accuracy of our algorithm in further iterations. To summarize, *SinguDetect* was built to address the shortcomings of the current state-of-the-art algorithms used in either source counting or singularness detection. Although *SinguDetect*'s algorithm is based upon the work done in [3], it has several important improvements that are substantial in making progress towards solving the overall problem of detecting the singularness of talkers in speech.

Chapter 3

System Design

Now, we delve into the system design of *SinguDetect*. First we will cover the calculation of the IPDs, followed by the GCC-PHAT equation. Next, we will cover the notion of structure in our IPD, how we quantify that structure, and we will explain why measures of structure are useful in classifying singularness. Then, we will formulate an improved method of establishing a trend line based on the given GCC-PHAT peak, which has a higher probability of more closely following the measured IPDs in singular frames. Finally, we describe how *SinguDetect* assimilates this information to predict the singularness of talkers in a frame.

To provide an overview of the system, *SinguDetect* begins by iterating through every frame of data in the STFT of the microphone observations. Then, we calculate the GCC-PHAT function using the STFT around that frame, including several frames of data in the calculation. Using the GCC-PHAT function, we construct a trend line based off the most likely TDoA. We segment the trend line into regions between phase wrappings, indexed $i \in \{1, ..., I\}$. For each segment i, we establish which time-frequency bins of the current frame's IPD are within a threshold ρ_{TH} of the segment. If the number of bins within this threshold are sufficiently large relative to the length of the segment, then we posit that the IPD has enough structure that the frame at least has one source present in it. Otherwise, if no segments meet this criteria, then we declare that the frame has no sources dominating it as its IPD is too unstructured, hence it is not singular. In the former case, we repeat the same process after cancelling out the time-frequency bins within the threshold ρ_{TH} of the IPD of the current frame, as well as the corresponding bins in the current window of the GCC-PHAT function. If we again find sufficient structure in the IPD, then we posit that the frame has another source present in it, so we mark the frame as non-singular. If we do not meet this condition, then we finally mark the frame as singular.

3.1 IPD Calculation

In order to produce the IPD, L. Wang et al. use Equation (2.3). While this method generally produces suitable results, in the noisy and reverberant environments tested in this thesis, it was found to be unsatisfactory. *SinguDetect* uses a more robust means of IPD calculation. Rather than take the angle between the STFT of the first microphone and the second, the algorithm instead uses the angle of the estimated instantaneous relative transfer function. We use the method of estimating the RTF from [7]. Once we have estimated the

RTF, then estimating the IPD is straightforward. We simply take the angle of the RTF as shown below.

$$\hat{\psi}(k,l) = \angle(\tilde{W}(k,l)) = \angle(\frac{\sum_{n=l-T/2}^{l+T/2} \hat{\Phi}_{y^2y^1}(k,n)}{\sum_{n=l-T/2}^{l+T/2} \hat{\Phi}_{y^1y^1}(k,n)}) \tag{3.1}$$

Using the estimated RTF rather than taking the angle of the ratio of the microphone signals directly has empirically given better results, as it averages out errors in the IPD due to noise or reverberation over the prescribed window. With too short a window, our IPDs will be too noisy to use for classification, but with too long a window, our IPDs will not adequately represent the frame they are designated for.

3.2 GCC-PHAT

Although we have examined the GCC-PHAT equation, we did not explain the W_{TF} term. This term in fact consists of three separate components. One is an SNR term, one is based on coherence, and the final is used to cancel out time-frequency bins [3]. These terms are simply multiplied by each other to produce the final weight

$$W_{TF}(k,l) = W_{SNR}(k,l)W_{coh}(k,l)W_{cancel}(k,l) \tag{3.2}$$

and the final term serves as a mean of eliminating the impact of time-frequency bins that have too low of an SNR, too low coherence, or are to be canceled out during the algorithm's iteration. The W_{SNR} term is calculated by detecting whether the local SNR $\lambda(k,l)$ exceeds a threshold λ_{TH}, and setting the term to zero if it does not meet this criteria. $\lambda(k,l)$ is calculated by the following equation

$$\lambda(k,l) = \min\,(\frac{P_{x_1}(k,l)}{P_{v_1}(k,l)} - 1, \frac{P_{x_2}(k,l)}{P_{v_2}(k,l)} - 1) \tag{3.3}$$

where $P_{x_m}(k,l) = |X_m(k,l)|^2$ is the power of the m'th microphone signal, while $P_{v_m}(k,l)$ is the power of the noise signal [3]. We can estimate $P_{v_m}(k,l)$ with the following equation, assuming our noise is stationary and that we have frames of data L_v where we know only noise is present

$$P_{v_m}(k) = \frac{1}{L_v}\sum_{l=1}^{L_v} |X_m(k,l)|^2 \tag{3.4}$$

which is a simple estimation of the expectation of the power of the noise at every frequency. Once we have obtained this value, we can then proceed to set W_{SNR} appropriately in the following equation

$$W_{SNR}(k,l) = \begin{cases} 1 & \lambda(k,l) > \lambda_{TH} \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

Next, we settle the issue of the W_{coh} weight. Coherence is a metric that indicates how likely it is that a single time-frequency bin is dominated by a singular source [3]. Therefore, we set W_{coh} to 1 if our coherence at a time-frequency bin, $r(k,l)$ is higher than a threshold r_{TH}, and 0 otherwise, similar to the W_{SNR} term. This coherence $r(k,l)$ is defined as the following [3]

$$r(k,l) = \left| \frac{E(X_1(k,l)X_2^*(k,l))}{\sqrt{E(X_1(k,l)X_1^*(k,l))}\sqrt{E(X_2(k,l)X_2^*(k,l))}} \right| \tag{3.6}$$

where $E(\cdot)$ refers to the expectation. We estimate the expectation by averaging the value enclosed by the operation over $2C+1$ consecutive time frames. Finally, we address the W_{cancel} weight. We simply set certain time-frequency bins of W_{cancel} to 0 if the previous iteration's IPD trend line approached the IPD within a threshold ρ_{TH}. To reiterate, to produce the final weighting W_{TF} we multiply all three weights together: $W_{SNR}W_{coh}W_{cancel}$.

Finally, we resolve the issue of narrowing the GCC-PHAT calculation find the TDoA of a single frame. We can simply sum across a window of time frames W to localize the GCC-PHAT

$$\tilde{R}_w(\tau) = \left| \sum_{l'=l-\frac{W}{2}}^{l+\frac{W}{2}} \sum_k W_{TF}(k,l') e^{j\hat{\psi}(k,l)} e^{-j2\pi f_k \tau} \right| \tag{3.7}$$

where l is the time frame in the current iteration. However, we need to be careful that the window W is not too large such that the derived τ no longer represents an AoA found in the current frame, but not too small that the function becomes unreliable. We can then center and normalize $R_w(\tau)$ by subtracting its minimum value from it, and then dividing it by its maximum value. We do this in order to ensure that our peak finding remains consistent.

3.3 Measuring Structure in the IPD

Once we have the localized GCC-PHAT function $\tilde{R}_w(\tau)$, we find the most probable source TDoA by taking the τ value corresponding to the highest peak. We can generate an estimate an estimate of the single source IPD trend line using the following equation to calculate the slope of the line

$$\mu = 2\pi f_{inc}\tau \tag{3.8}$$

where f_{inc} refers to the increment in frequencies across each STFT frequency bin. For example, an FFT size of 1024 and a sampling frequency of 16,000 Hz would yield an f_{inc} of 15.625 Hz. Then we merely multiply μ by the vector of frequencies $\mathbf{f} = \{0, \ldots, f_{max}\}$ where f_{max} corresponds to the integer $\frac{fs}{2}$, where fs refers to the sampling frequency, and the number of entries in $\mathbf{f}$ to $\left\lfloor \frac{K}{2} \right\rfloor + 1$ to get the trend line γ. In order to conform to the phase wrapping in the IPD, we wrap the trend line by using the following equation

$$\bar{\gamma} = (\gamma + \pi)_{2\pi} - \pi \tag{3.9}$$

Now we have a hypothesis of what the observed IPD should be, assuming the TDoA we have captured corresponds to the direct path from a source to the microphone. However, if our frame contains multiple sources within it, we will find that the IPD can differ greatly from this hypothesis, especially in reverberant and noisy scenarios. Uncorrelated noise will perturb the IPD, because it will randomly (across frequencies) bias the measured IPD positively or negatively [13]. Reverberation will similarly disrupt the structure of the IPD, as the image of the source from angles other than the direct path can produce a different IPD pattern than the direct path scenario. This pattern can clash with the direct path IPD as well as other multipath IPDs across frequency bins, again reducing the structure of the measured IPD. This means that the

presence of two or more speakers may severely clobber the structure of the measured IPD, and consequently the GCC-PHAT, which relies on the IPD to calculate the most probable direct path TDoA. This situation would violate the key assumption in [3] — that the time-frequency bins closest to the estimated IPD trend line belong to the source that corresponds to the TDoA that generated the trend line. We can see this illustrated in Figure 3.1. The singular frame on the left clearly corresponds to a source with an AoA of around 0° at the microphone, whereas the two source frame on the right is relatively unstructured and it is not obvious which AoAs the two sources produce, or whether there are two sources present at all. Therefore, the iterative algorithm in [3] is unlikely to produce the correct direct path AoAs, and successfully cancel out the time-frequency bins related to the two sources.

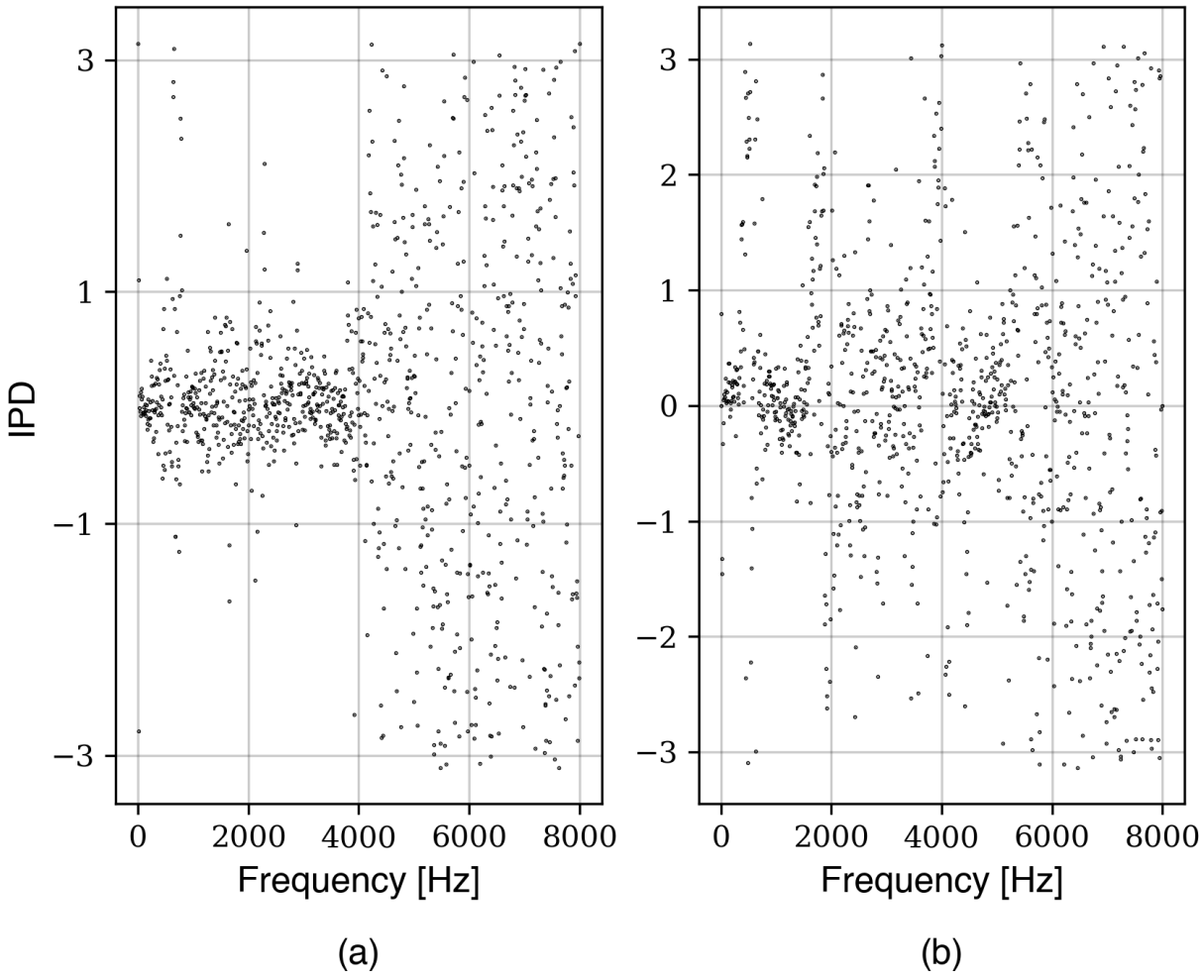

Figure 3.1: Illustration of an IPD in a singular frame vs. two source frame. The source producing the singular frame is present in the two source frame. (a) Singular frame, (b) Two source frame.

Thus, we must devise a method to account for the degradation of the structure of the IPD in multi-source frames, rather than relying on the ability to count the number of unique AoAs across iterations. An efficient method to detect the structure of the IPD is to check how closely the trend line $\bar{\gamma}$ matches the measured

IPD. If there are a sufficient number of frequency bins where the IPD is significantly close to the trend line, then we claim that the frame is highly likely to contain at least one source in it. We reason that if these conditions are met, then the calculated anechoic IPD derived from the highest peak of the GCC-PHAT must accurately represent the measured IPD, thus at least one source is present. In this case, just as in [3], we cancel the time-frequency bins associated with the first TDoA from both the GCC-PHAT and the IPD, and repeat the process. If we find that the new trend line derived from the new TDoA fulfills the aforementioned condition, then we have at least a two source frame, and thus we mark the frame as non-singular. In the opposite case, we can finally mark the frame as singular.

However, we must carefully consider how to count the number of frequency bins where the trend line successfully approximates the measured IPD. The frequency bins in which perturbations caused by noise and reverberation reside are difficult to predict, and therefore can disrupt the structure of the IPD unevenly. For example, in Figure 2.3, the IPD has significant deviations occurring around 0 Hz and around 2000 Hz. In order to mitigate this, we propose that we should count these frequency bins between each phase unwrapping. Although these perturbations do not necessarily occur between phase wrappings, we posit that it is simple, efficient, and effective to account for these disturbances by counting between each phase wrapping. Generally, the presence of a single region between phase wrappings with enough frequency bins matching the trend line was enough to successfully identify the frame as at least containing one source.

To summarize, we begin our algorithm by finding where the phase wraps occurs on the trend line. To do this, we check where the difference between the consecutive values of the trend line is greater than 5 radians. Then, for each segment of the trend line between phase wraps, $\bar{\gamma}_i$, where i refers to the index of the segment, we count the number of time-frequency bins where the IPD is within a threshold ρ_{TH} of $\bar{\gamma}_i$. We use the absolute error to determine the distance of the IPD from $\bar{\gamma}_i$, which is shown in Equation (3.11 below. We use this distance to find C_i, the number of frequency bins in the trendline $\bar{\gamma}_i$ sufficiently close to the IPD $\hat{\psi}(k,l)$ as shown in Equation (3.10) below.

$$C_i = \sum_{k \in \Omega_i} \mathbb{1}_i(k) \tag{3.10}$$

where the indicator function $\mathbb{1}_i(k)$ is defined as the following

$$\mathbb{1}_i(k) = \begin{cases} 1 & \text{if } \left|\hat{\psi}(k,l) - \tilde{\gamma}_i(k)\right| < \rho_{TH} \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

where l refers to the current frame. If C_i is greater than another threshold ζ_i, then we predict that our trend line $\tilde{\gamma}$ correctly represents the TDoA of a source present in that frame, as it contains sufficient structure. Otherwise, we predict that the frame is non-singular, and move onto the next frame. The reason why the threshold is indexed by i is that we set it to be a multiple of the length of the line segment. Since line segments vary in length — substantially across time frames — it is logical to examine the ratio of well predicted IPD values to the number of time-frequency bins within the segment. The higher this ratio is, the higher the probability the trend line represents a real TDoA.

Assuming that there are enough time-frequency bins to meet ζ_i, we now proceed to check if there is another TDoA present. We use the current TDoA to establish trend lines in all frames in the current window, and then we cancel out all time-frequency bins from all STFT frames in the window W, and all bins from the current IPD where the trend lines are within ρ_{TH} of the corresponding IPDs. Let us refer to the set Ω_{cancel} as the union of the sets of time-frequency bins that are to be canceled. This union is composed of the sets Ω_l

where the l refers to the frame index, as before.. We cancel the bins in the GCC-PHAT function window accordingly

$$W_{cancel}(k,l) = \begin{cases} 0 & \text{if } k \in \Omega_l \\ 1 & \text{otherwise} \end{cases} \tag{3.12}$$

And then we cancel the bins in the IPD as follows

$$\hat{\psi}(k,l) = \kappa_{max} \text{ if } k \in \Omega_l \tag{3.13}$$

where κ_{max} refers to an arbitrarily high value. We 'cancel' the IPD by setting it to κ_{max} in order to ensure that the subsequent trend lines have no chance of claiming these time-frequency bins again.

Then, we repeat the process — calculate the GCC-PHAT function, fit trend lines, and check for the two thresholds ρ_{TH} and each ζ_i. If we find the GCC-PHAT has a peak in the same location as before within a tolerance A_{min}, we find the TDoA corresponding to the next largest value of the function that guarantees it to be outside of this tolerance. This time, if we find that any threshold ζ_i is met, we know that there are at least two TDoAs present in the time frame — thus, we have a non-singular frame. On the other hand, if the threshold is not met, we ultimately predict a singular frame. We reason that with the contributions from the previous TDoA canceled out, the new TDoA corresponding to the peak or next largest viable value of the re-calculated GCC-PHAT function should belong to another source. And, with the contributions from the previous TDoA canceled out in the IPD, our trend line will not consider bins from the previous TDoA. Thus, if the trend line exceeds the two thresholds, it has a high possibility of representing an actual TDoA.

3.4 Offsetting the Trend Line

Although we have settled the issue of the lack of structure in non-singular IPDs thwarting the algorithm in [3], we have one final problem to address. The offset calculation from Equation (2.10) is a global optimum, i.e. it involves calculating the loss over every time frame and frequency bin. Even if we were to restrict the loss to a single time frame to fulfill our purposes, the loss function still presents a problem. As stated previously, the IPD suffers difficult to predict perturbations at certain frequencies, and in environments where strong early reflections are present, the IPD can have non-linear characteristics [3]. This can be observed in Figure 2.3, where the IPD clearly varies non-linearly with respect to frequency at around 0 Hz. Thus, we may encounter a situation where the optimal offset leads the trend line to be misplaced, such that the IPD appears unstructured, and therefore non-singular according to our previous metrics for structure. We can observe this phenomena in Figure 3.2. The original offset calculation method, prescribed in [3], fails to capture the locally linear trends in the IPD because of the globality of the loss function.

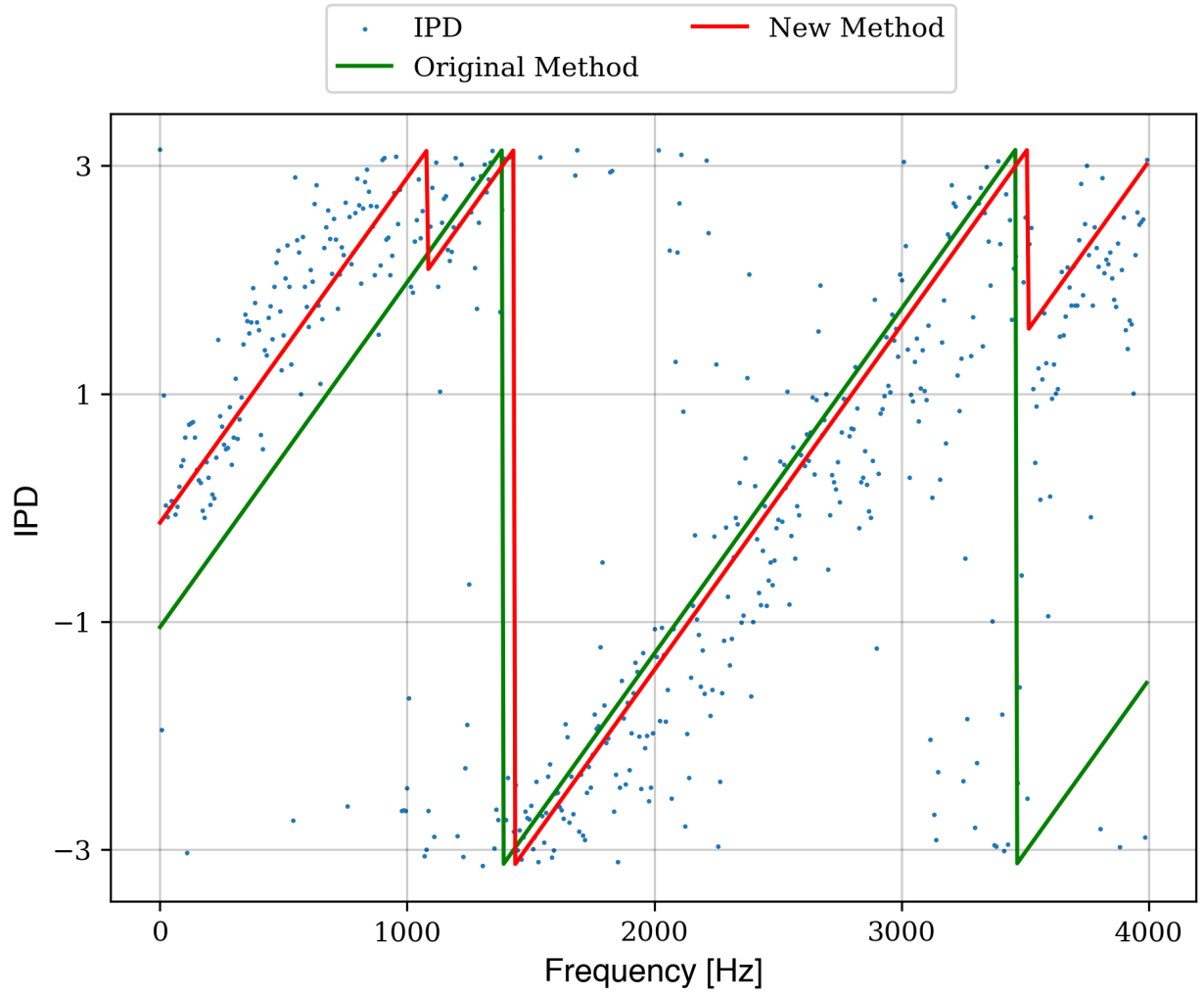

Figure 3.2: Illustration of the results of the original method in calculating the offset of the trend line vs. the new method of calculating the trend line compared to the measured IPD

In line with our previous thinking, we must therefore devise a more local means of calculating the offset, such that we can approximate the IPD more correctly, assuming the TDoA derived from the GCC-PHAT matches the direct path TDoA, and assuming the frame is singular. Therefore, we prescribe that the optimal offset be calculated for each segment of the trend line between phase wrappings, producing a piece-wise linear function similar to the one depicted in Figure 3.2. Through this method, we can increase the probability that the trend line matches the measured IPD.

Thus, at each phase wrap, we produce the best offset such that the absolute error between $\bar{\gamma}_i(k+\phi)$ and the IPD is minimized, that is

$$\hat{\phi}_i = argmin_\phi \sum_{k \in \Omega_i} |\bar{\gamma}_i(k+\phi) - \psi(k,l)| \tag{3.14}$$

where Ω_i refers to the frequency bins in the current segment. Starting from the first segment, we calculate its optimal offset $\hat{\phi}_1$ and produce the first segment $\tilde{\gamma}_1$, of our trend line, with $\tilde{\gamma}_i$ referring to the i'th optimally

shifted line segment in our final trend line and $\tilde{\gamma}$ to the line itself. To produce the rest of the trend line, we repeat the process, beginning with the end of the previous segment. Of course, these offset segments may need to be phase wrapped themselves, as seen in Figure 3.2, so we do so and begin the next segment on the second phase wrap of these segments.

3.5 Summary

This entire sequence of steps for *SinguDetect* can be summarized in algorithm 1, found below.

Algorithm 1 SinguDetect

Input: $\mathbf{X}_1$, $\mathbf{X}_2$ — the microphone observations and $\mathbf{L}_v$, the noise frames

Output: Singularness of each frame

1: $\rho_{TH} \leftarrow 0.5$
2: Calculate the IPDs using Equation (3.1)
3: **for** l in L **do**
4: Calculate the GCC-PHAT using Equation (3.7)
5: Calculate the slope μ of the initial trend line using Equation (3.8)
6: Calculate the trend line $\tilde{\gamma}$ using Equations (3.9) and (3.14)
7: $fit \leftarrow False$
8: **for** Segment seg in $\tilde{\gamma}$ **do**
9: $\zeta_{seg} \leftarrow len(seg) \cdot 0.5$
10: calculate C_{seg} using Equation (3.10)
11: **if** $C_{seg} > \zeta_{seg}$ **then**
12: $fit \leftarrow True$
13: **end if**
14: **end for**
15: **if** fit is $True$ **then**
16: cancel all time-frequency bins in the GCC-PHAT according to Equation (3.12)
17: cancel all time-frequency bins in the current IPD according to Equation (3.13)
18: repeat steps 4–13
19: **if** fit is $True$ **then**
20: Declare current frame l as non-singular
21: **else**
22: Declare current frame l as singular
23: **end if**
24: **else**
25: Declare current frame l as non-singular
26: **end if**
27: **end for**

Chapter 4

Experiments and Results

In the previous section, we introduced the algorithms and procedures used in *SinguDetect*. In this section, we present experimental results that test the ability of the proposed algorithm to detect singularness. In particular, we compare our proposed approach against against the algorithms in Section 2.3 on three different scenarios. Since the performance of *SinguDetect* is dependent on various parameters, we first discuss the specific choices made for our experiments.

4.1 Parameters and Details

We list the parameters associated with the algorithm in Table 4.1 below. To produce the STFT, we use an FFT size of 2048 and an overlap of 1024 or a hop length of 1024. Our experiment consists of 100 trials using three different sets of binaural room impulse responses (BRIRs), and under five different signal to noise ratios (SNRs). Thus, the total number of trials run was 1500. These 100 trials consisted of five different numbers of speakers, ranging from 1 to 5, with each speaker number being assigned to twenty trials each. Each trial comprised no more than 11 seconds of audio and no less than 1 second, and each speaker began and ended at a random time. The candidate pool for the speakers was fixed to 10 different speakers chosen from the CSTR VCTK Corpus from the University of Edinburgh [14], with equal representation from men and women. This corpus was chosen because it comprised anechoic and noiseless mono audio. In order to make the resulting audio reverberant, we convolved the BRiRs with the corpus samples, producing reverberant binaural audio suitable to replace actual microphone signals. The list of SNRs used was 25 dB, 20 dB, 15 dB, 10 dB, and 5 dB. At 5 dB, the results degraded to a point that further decreasing the SNR would no longer produce any insightful information. For the experiment, we added uncorrelated zero mean Gaussian noise to both microphone signals by sampling from a Gaussian distribution with a set standard deviation. These standard deviations were chosen such that they would produce the corresponding desired SNR.

Table 4.1: Parameters used by SinguDetect

Parameter	Section	Value
λ_{TH}	3.2	5
C	3.2	5
r_{TH}	3.2	0.9
ρ_{TH}	3.3	0.5
T	3.1	6
W	3.2	6
ζ_i	3.3	length$(\bar{\gamma}_i)/2$
c	2.3.3	343.0 m/s
d	2.3.3	0.27 m
α_{TH}	4.4	0.8
A_{min}	3.3	12°

4.2 Algorithms for Comparison

We compare the proposed *SinguDetect* with two of the algorithms previously outlined, the simplex analysis of B. Laufer-Goldshtein et al. [7] and the pitch detection of A. O. T. Hogg et al [8]. In the case of the former, the threshold β was set to 0.6, because this empirically produced the best results. Otherwise, all other parameters for the simplex analysis algorithm remained the same. As for the pitch detection algorithm, only a small change was made. Instead of using a minimum amplitude of 10^5 for PEFAC, we used an amplitude of $4 \cdot 10^5$ which was found to be more performant on the data for this experiment.

4.3 Explanation of Scenarios

4.3.1 Scenario (a)

As previously explained, we produced our data using BRiRs from three different datasets. The first — scenario (a) — was derived from the dataset *360° Binaural Room Impulse Response (BRIR) Database for 6DOF spatial perception research* [15]. Here, a portion of the dataset has the Neumann KU100 binaural head rotated in 3.6° increments to provide a full 360° set of measurements in a reverberant concert hall. In particular, we use the C2m set of BRiRs for our trials, and we use every second angle in order to avoid providing sources that are too angularly similar. Thus, the true angular increment for our purposes is 7.2°. Finally, the average RT60 of the concert hall where the dataset was recorded is 2.1 s [15]. Finally, the speaker is located at a position sufficient for far-field audio recording, a distance of 2 meters.

4.3.2 Scenario (b)

The second scenario was derived from the dataset *A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios* [16]. As before, part of the dataset was captured using the Neumann KU100 binaural head rotated in 1° increments to provide a full 360° set of measurements in a large broadcast studio. In particular, we use the LBS_KU_MICS_PAC set of BRiRs for our trials, and we use every tenth angle in order to avoid providing sources that are too angularly similar. Thus, the true angular increment for our

purposes is 10°. Finally, the average RT60 of the large broadcast studio is around 1–2 s for frequencies below 8000 Hz [16], and the speaker is located at a position sufficient for far-field audio recording.

4.3.3 Scenario (c)

The third scenario was derived from the dataset *Binaural room impulse responses of a 5.0 surround setup for different listening positions.* [17]. Unlike before, the dataset was collected using a KEMAR head rotated in n 1° increments to provide 180° set of measurements in listening room Calypso at Technische Universität Berlin. This 180° coverage refers to the arc in front of the KEMAR head, such that no recordings with the speaker behind the microphones are recorded. In particular, we use the 6_KEMAR_Calypso_Surround_X0.0m_Y0.0m set of BRiRs for our trials, and we use every second angle in order to increase the level of difficulty for *SinguDetect*. Thus, the true angular increment is 2°. Finally, the average RT60 of the listening room is 0.17 s at 1000 Hz [17], and the speaker is located at a position sufficient for far-field audio recording, a distance of 2 meters as we select speaker 2 for our BRiRs.

4.4 Evaluation Measure

We shape our evaluation measure according to the original use case. Because [1] was interested in capturing data where single talkers were present, precision, recall and F-score are natural tools for evaluation, just as in [3]. In this case, precision refers to the ratio of correct identifications of single talker frames to the total number of such identifications for the algorithm in question. Meanwhile, recall refers to the ratio of correct identifications of single talker frames to the total number of actual single talker frames. Therefore, the F-score can be thought of as a balance between these two scores, with the lower of the two scores more heavily influencing the F-score. Furthermore, we can use a factor of β_R to define how important the recall score is relative to the precision [18]. Let us define the number of correct identifications as S_{true}, the total number of identifications as S_{total}, the total number of actual single talker frames as N_{single}, and finally precision, recall, and f-score as P, R, and F_{score} respectively. Equations for the three metrics can be found below.

$$P = \frac{S_{true}}{S_{total}},\ R = \frac{S_{true}}{N_{single}},\ F_{score} = (1+\beta_R^2)\frac{P \cdot R}{(P \cdot \beta_R^2) + R} \tag{4.1}$$

Clearly, we wish for any prospective singularness detector to have high precision and recall. If it were to have a low precision, then the data that it produces cannot be trusted for use in downstream applications. However, if it were to have a low recall, then it would require a lot of samples in order to produce enough data for downstream applications which is also undesirable. Nevertheless, below a certain precision score, the recall rate becomes irrelevant. For example, a precision score of 0.5 means we can only trust half the samples reported to be singular, which would badly pollute any data for upstream applications. Therefore, we set the β_R parameter accordingly to 0.5, because precision is twice as important for our application than recall is. Overall, the f-score provides us with a good measurement for the success of any algorithm for singularness detection.

Our definition of a single talker frame is as follows. If the ratio of the sum of the spectral density of the noisy mix to the sum of the spectral density of any of the clean sources is greater than a threshold α_{TH}, then we declare that frame to be singular. That is,

$$G(l) = \begin{cases} 1 & \text{if } \exists i \text{ such that } \frac{\sum_{k=1}^{K} |S_i(k,l)|^2}{\sum_{k=1}^{K} |X_m(k,l)|^2} > \alpha_{TH} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where $G(l)$ refers to whether or not the $l'th$ frame is singular. Note that α_{TH} should be a fairly high value, as the noisy mix rarely equals the sum of the power spectra of the noise and the individual clean sources due to destructive and constructive interference. Therefore, we do not wish to have a singular frame where more than one source is 'dominant' by virtue of the threshold.

4.5 Effects of finding the Optimal Offset

As explained in Section 3.4, we break the trend line into segments, and solve for the optimal offset for each segment. This is in contrast to [3], which solves for the optimal offset among all frequency bins overall. In order to elucidate the advantage of examining each segment separately, we therefore compare the results of the original *SinguDetect* algorithm with one that is modified to implement a line fitting algorithm more closely aligned to [3]. We refer to this modified algorithm as the iterative algorithm. To elaborate, on the first iteration it produces the best offset for the phase wrapped line with the slope derived from Equation (3.8). Then, it iteratively cancels the time-frequency bins, just as in the original *SinguDetect* algorithm, such that the GCC-PHAT function produces a different highest peak, indicating a new AoA. Every time the same AoA is produced, the optimal offset is recalculated, and time-frequency bins are accordingly canceled. If this process repeats 5 times, then it is halted, and the next largest peak is used to produce a new phase wrapped line. Therefore, the main difference in result between this iterative algorithm and *SinguDetect* occurs when the loss function in Equation (2.9) leads to an offset of the entire trend line such that the trend line substantially differs from the linearly varying portions of the IPD in a singular frame. Alternatively, a difference would emerge if the original *SinguDetect* algorithm erroneously identifies a frame as singular, as a result of producing a piece-wise linear trend line that sufficiently matched the observed IPD.

We can see the slight advantage of the non-iterative algorithm in Figure 4.1. In scenario (a), the two algorithms have almost indistinguishable F-scores, as they match F-scores at every SNR and number of sources. However, in scenario (b), we can see that *SinguDetect* boasts a slight advantage in terms of F-score, which is present at a lower number of sources, and diminishes at a higher number of sources. This advantage can be as high as 0.05 — somewhat marginal, but still important as we will see late in distinguishing *SinguDetect* from other algorithms. It seems that in scenario (b), as opposed to scenario (a), there exists instances where the iterative process does produce an offset such that some singular frames are incorrectly identified as too unstructured. These instances reduce in frequency at higher source numbers likely because both algorithms are overestimating the number of singular frames, which the segment examination method does not prevent. Note that for Figure 4.1, the graphs are cutoff below an F-score of 0.4 because any data below that number are practically irrelevant, and showing such numbers would make the important data less readable.

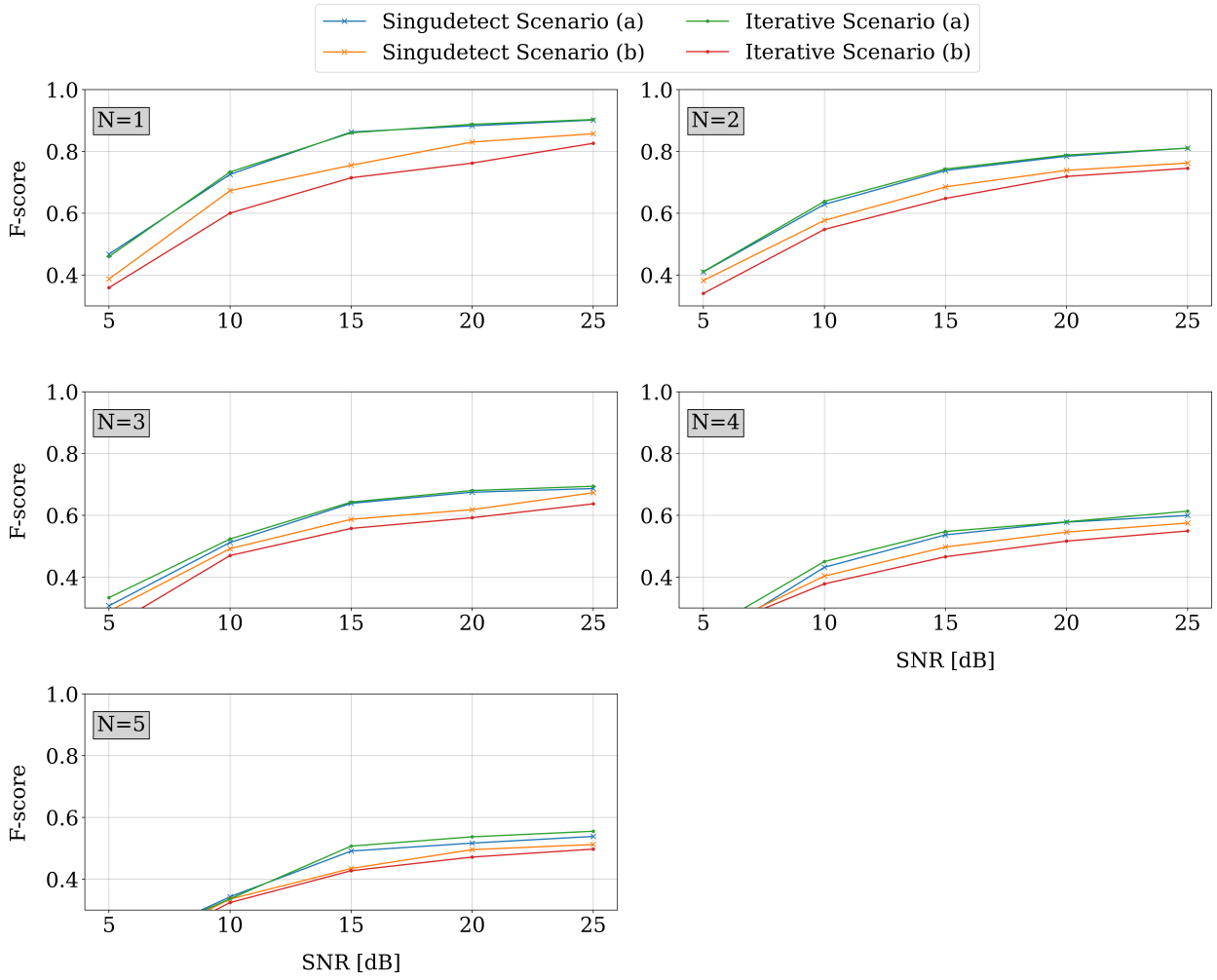

Figure 4.1: F-score vs. SNR across the possible number of sources for *SinguDetect* using the iterative method vs. the non-iterative method

4.6 Results of BRiR Scenarios

Now, we begin comparing the algorithms by presenting the same type of data as in Section 4.5 with the F-scores averaged across all three scenarios. Examining Figure 4.2, we can see that F-scores intuitively decrease with a lower SNR. Particularly at the higher source numbers of 3 and above, we can see that no algorithm performs well at SNRs below 15 dB. This is why the range of SNRs was set from between 5 dB to 25 dB, as further information resulting from SNRs below the 5 dB mark would be uninformative. Although it is clear why a low SNR degrades the F-scores of the algorithms so severely, it is not clear why having a higher number of sources has such a negative impact. With our proposed approach, it is possible that the unpredictable behavior of the sources interfering with each other often does produce a clear IPD trend line that *SinguDetect* perceives as a singular source. For the simplex method, one would assume that the estimated probabilities for each individual source would remain low in the presence of more sources, resulting in accurate predictions as to the lack of singularness in the audio clips. Similarly, the pitch detection method

should be able to distinguish the greater number of pitches, and predict a lack of singularness. Thus, the degradation of the algorithms' performance given a greater number of sources is puzzling. Nevertheless, we can clearly see that our given method outperforms all other algorithms with respect to the F-score, except at a source number of 1, where it matches the pitch method. At lower source numbers, particularly at 2 and 3, we can see that *SinguDetect* boasts an increase of as much as 0.1 in the F-score at higher SNRs compared to the other algorithms.

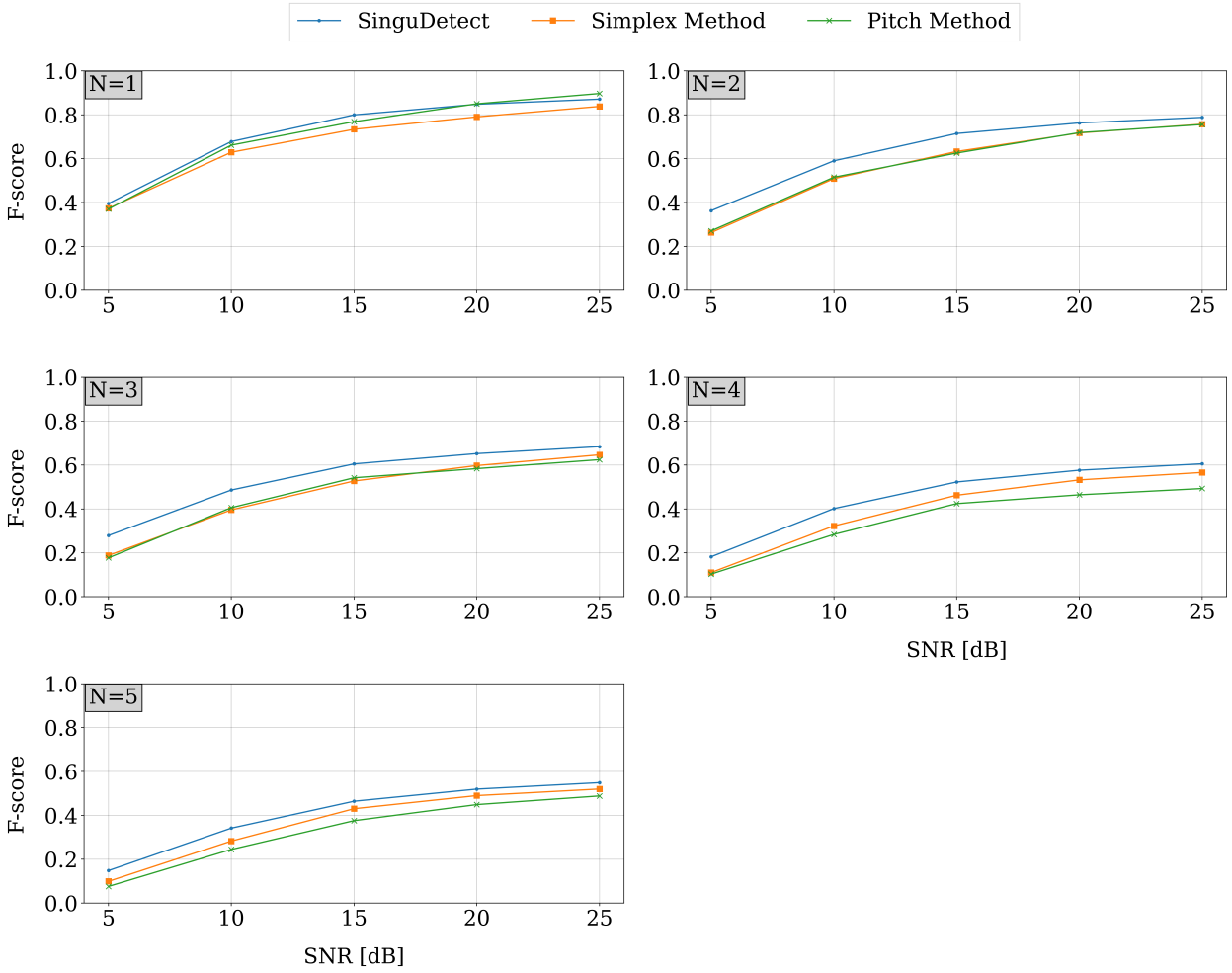

Figure 4.2: F-score vs. SNR across the possible number of sources for the different singularness detector algorithms. Data is averaged across all three scenarios.

We dive into the precision, recall, and f-scores in scenario (a) for each algorithm. From Figure 4.3, we can piece together why the F-score decreases rapidly in lower SNR scenarios. Here, we only have three speakers present in each audio clip. Clearly, the recall score does not degrade too much for the simplex method and *SinguDetect*. Therefore, the issue with the lower F-score lies singularly with precision for these two algorithms, whereas for the pitch method, both precision and recall decrease with SNR. The pitch method suffers in its recall score at lower SNRs because it labels almost every frame as non-singular, and in general, the method labels fewer frames as singular compared to the others. Additionally, with a lower SNR, these singularness

detection algorithms overestimate the number of singular frames, resulting in a lower precision. Notably, *SinguDetect* has a superior F-score to every other algorithm at every SNR, and matches the pitch method in terms of precision, but has a lower recall compared to the simplex method. Nonetheless, we believe this disadvantage in recall does not offset the advantage in precision, thus our setting of β_R to 0.5 is justified.

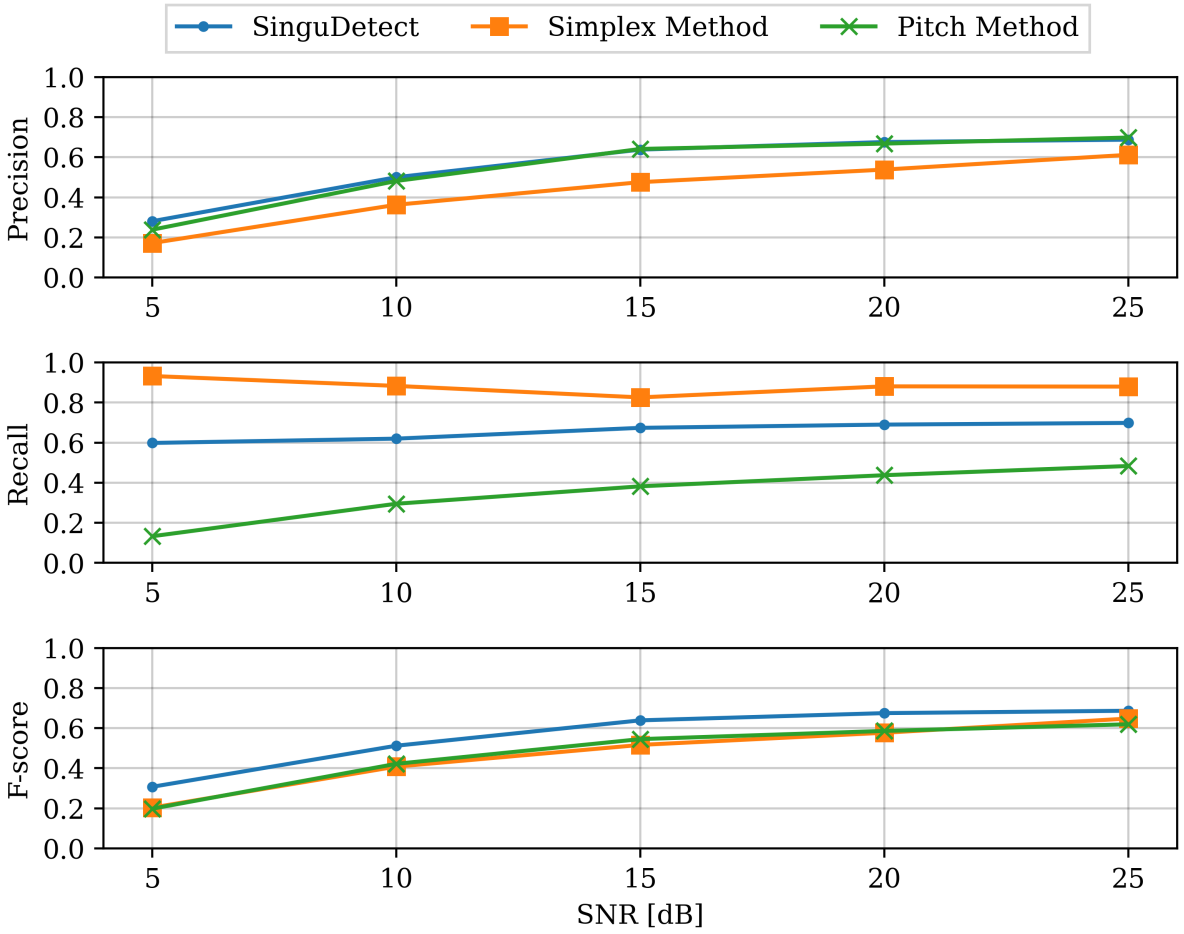

Figure 4.3: Precision, recall, and F-score vs. SNR for the different singularness detector algorithms. The scenario providing the data is scenario (a), and the number of speakers is set to 3.

Finally, we break the results into each scenario, citing the precision, recall, and f-score across each number of sources at 15 dB SNR. In Figure 4.4 we can see that the featured algorithms have quite different results in each scenario. In particular, precision scores are lower for *SinguDetect* and the simplex method for scenario (c), and the f-scores of the three algorithms are close in scenario (b). With scenario (c), we have a lower overall reverberation time, but a smaller room. Therefore, it is likely that there is a strong early reverberation component in the audio, which confounds the simplex method and *SinguDetect*. In the case of the former, these early reverberations may present themselves as different sources, which would reduce its accuracy since there would be a separate probability for different echoes presenting themselves as sources. In the case of the latter, the early reverberations may smear the IPD in unpredictable ways, resulting in a particularly low recall rate at higher source numbers and a lower precision rate overall. Additionally, the chances of there being two sources located relatively close together due to the small angular increment in scenario (c) would

hurt the precision of *SinguDetect* more so than the other algorithms. In the case of scenario (b), it is not clear why the recall rate of *SinguDetect* suffers as compared to scenario (a) — the culprit of its lower F-score. Scenario (a) and (b) are quite similar — taking place in large, reverberant rooms. Regardless, perhaps there is some variation among IPDs of different rooms such that *SinguDetect* becomes more hesitant to declare frames singular — decreasing its recall score but not its precision. Nevertheless, at 15 dB, it is clear that *SinguDetect* is superior in terms of F-score in every possible situation except in scenario (b) with only a single source present.

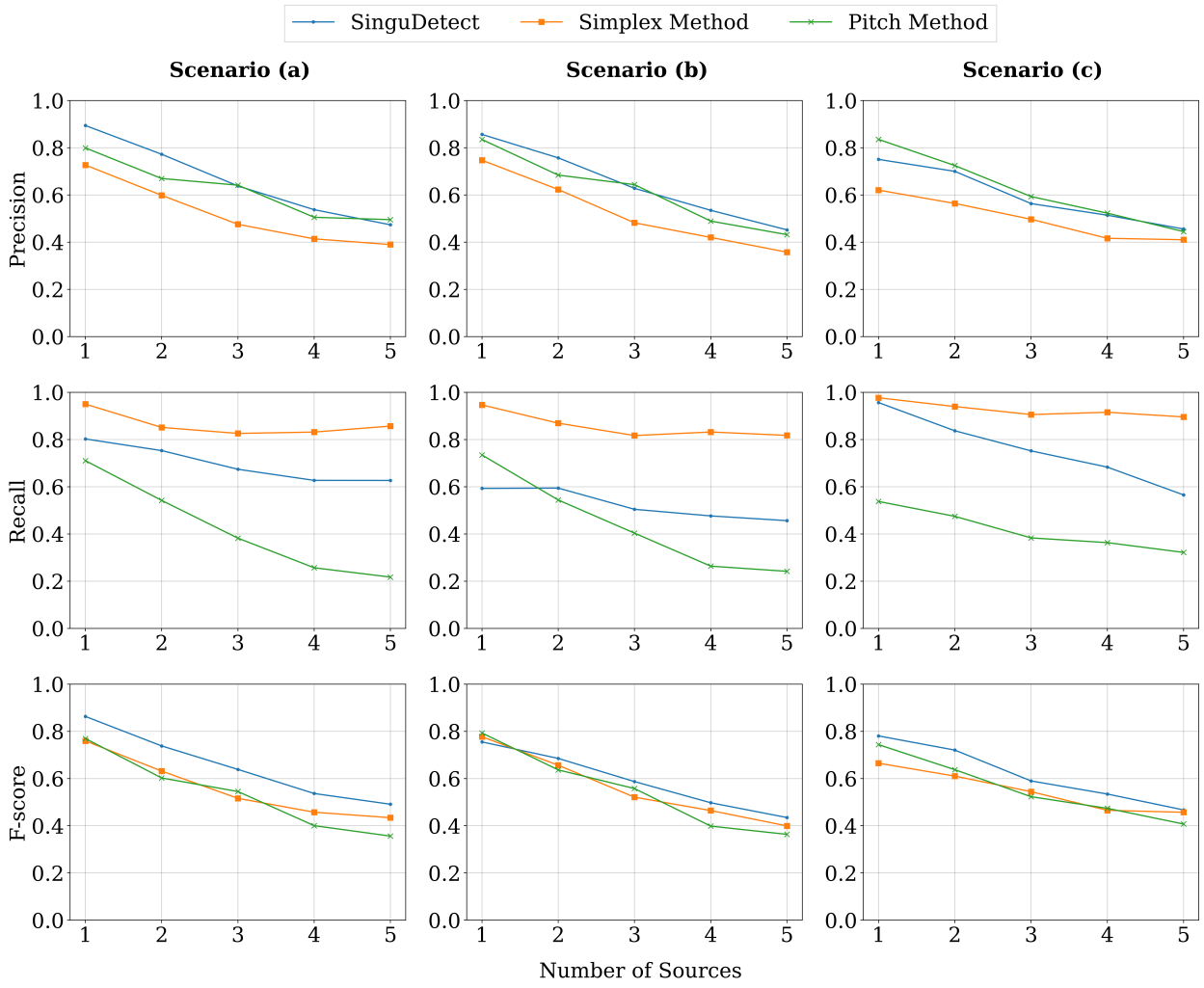

Figure 4.4: Precision, recall, and F-score vs. the number of sources in the trial across the three scenarios for the singularness detector algorithms. SNR is set to 15 dB.

Chapter 5

Limitations and Future Work

Given the original purpose of *SinguDetect*, it remains obvious that there is a lot of work still to be done. Although the algorithm boasts an increased performance as compared to its competitors, its performance at lower SNRs and a higher number of sources is quite weak. Some of this is due to the inherent nature of IPDs — at low SNRs the IPD of a binaural recording is likely to be obliterated if the noise is uncorrelated. However, the lack of success at at a higher number of sources across all three methods remains puzzling. Perhaps the method of measuring the ground truth singularness of talkers is flawed, and a superior method exists to establish which time frames of the STFT are singular. The beamforming metrics of [7] were considered, where the singular frames were used to build beamforming weights for each individual talker, but ultimately were decided as too difficult to interpret within the context of the original problem.

Overall, we can attribute the weaknesses of *SinguDetect* to the limitations of a classification algorithm purely using IPDs. Often times when the algorithm had a low precision, it was because it correctly identified a trend line in a frame that was not dominated by any singular source. Therefore, any future algorithms must consider a breadth of features to improve results to the point that the problem is satisfyingly solved. Features that were considered for *SinguDetect* include inter-aural level differences (ILDs) and Mel-frequency cepstral coefficients (MFCCs), and could be helpful in augmenting classification performance. Intuitively, as humans, we do not rely on any particular one feature, whether it is pitch or the IPD. Even if two people were to have the same voice and spoke from the same direction, as long as they were saying different words, a human would be able to distinguish the lack of singularness, albeit at a coarse temporal level. This indicates a more holistic algorithm would have more success.

Of course, the field of machine learning, particularly with unsupervised or semi-supervised methods, may also advance to the point that it is able to solve the problem. This would obviate the need for classical methods.

Chapter 6

Conclusion

In this thesis, we have proposed a singularness detection algorithm that performs better than the state-of-the-art algorithms. Although there are clear shortcomings with *SinguDetect*, it performs remarkably well in heavily reverberant and noisy environments. Generally, it maintains a precision greater than 0.5 above an SNR of 10 dB at lower source counts, with a granularity of 64 ms for its singularness detection capabilities. Additionally, it is capable of providing a measure of certainty about its predictions and accommodates moving sources since it judges singularness on a frame-by-frame basis. Nonetheless, much work remains to be done to solve the difficult problem of detecting the singularness of talkers in every feasible acoustic environment, whether that be through classical methods or newer data-driven methods.

References

[1] A. Xu and R. R. Choudhury, "Learning to Separate Voices by Spatial Regions," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 24 539–24 549. [Online]. Available: https://proceedings.mlr.press/v162/xu22b.html.

[2] R. Garg, R. Gao, and K. Grauman, *Geometry-Aware Multi-Task Learning for Binaural Audio Generation from Video*, 2021. arXiv: 2111.10882 [cs.CV].

[3] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An Iterative Approach to Source Counting and Localization Using Two Distant Microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1079–1093, 2016. DOI: 10.1109/TASLP.2016.2533859.

[4] M. I. Mandel, S. Araki, and T. Nakatani, "Multichannel Clustering and Classification Approaches," in *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, Ltd, 2018, ch. 12, pp. 235–261, ISBN: 9781119279860. DOI: https://doi.org/10.1002/9781119279860.ch12. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119279860.ch12. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119279860.ch12.

[5] M. Akcakaya, C. Muravchik, and A. Nehorai, "Biologically inspired antenna array design using Ormia modeling," in *Biomimetic Technologies*, ser. Woodhead Publishing Series in Electronic and Optical Materials, T. D. Ngo, Ed., Woodhead Publishing, 2015, pp. 335–364, ISBN: 978-0-08-100249-0. DOI: https://doi.org/10.1016/B978-0-08-100249-0.00016-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780081002490000161.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976. DOI: 10.1109/TASSP.1976.1162830.

[7] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source Counting and Separation Based on Simplex Analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6458–6473, 2018. DOI: 10.1109/TSP.2018.2876349.

[8] A. O. T. Hogg, C. Evers, A. H. Moore, and P. A. Naylor, "Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking of Fundamental Frequency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1479–1490, 2021. DOI: 10.1109/TASLP.2021.3067161.

[9] X. Li, R. Horaud, L. Girin, and S. Gannot, "Local relative transfer function for sound source localization," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 399–403. DOI: 10.1109/EUSIPCO.2015.7362413.

[10] J. Yang, Y. Guo, Z. Yang, L. Yang, and S. Xie, "Estimating Number of Speakers via Density-Based Clustering and Classification Decision," *IEEE Access*, vol. 7, pp. 176 541–176 551, 2019. DOI: 10.1109/ACCESS.2019.2956772.

[11] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014. DOI: 10.1109/TASLP.2013.2295918.

[12] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, "Fundamentals of Acoustics, 4th Edition," 1999.

[13] J. C. R. Licklider, "The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise," *The Journal of the Acoustical Society of America*, vol. 20, no. 2, pp. 150–159, Mar. 1948, ISSN: 0001-4966. DOI: 10.1121/1.1906358. [Online]. Available: https://pubs.aip.org/asa/jasa/article/20/2/150-159/718948 (visited on 06/28/2023).

[14] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)*, 2019. DOI: 10.7488/ds/2645.

[15] B. I. Bacila and H. Lee, "360° Binaural Room Impulse Response (BRIR) Database for 6DOF Spatial Perception Research.," in *Audio Engineering Society Convention 146*, Mar. 2019. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20371.

[16] Stade, Bernschütz, and Rühl, *A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios*, Zenodo, May 2020. DOI: 10.5281/zenodo.3930833. [Online]. Available: https://doi.org/10.5281/zenodo.3930833.

[17] H. Wierstorf, *Binaural room impulse responses of a 5.0 surround setup for different listening positions*, Zenodo, Apr. 2016. DOI: 10.5281/zenodo.49691. [Online]. Available: https://doi.org/10.5281/zenodo.49691.

[18] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," in *Advances in Information Retrieval*, D. E. Losada and J. M. Fernández-Luna, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 345–359, ISBN: 978-3-540-31865-1.