Anonymising Speech in Surveillance using Speech Masking and Background Separation

Carl Örnberg

Master's thesis
2024:E52

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Anonymising Speech in Surveillance using Speech Masking and Background Separation

Carl Örnberg

Lund, 5th of June 2024

Master's Thesis in Mathematical Statistics

Supervisor LTH: Andreas Jakobsson
Supervisors Axis Communications AB: Rebecca Svensson, Alfred Widell
Faculty of Engineering, LTH
Centre for Mathematical Sciences

Preface

I express my deep gratitude towards my supervisors at Axis Communications AB, Rebecca Svensson and Alfred Widell, for giving me the privilege of discovering the world of digital audio signal processing and its application in surveillance technology. Our weekly meetings, your full commitment when answering my numerous questions and your steadfast encouragement spurred me to do my best - everyday.

I also want to earnestly acknowledge my supervisor at The Division for Mathematical Statistics at Lund University - Andreas Jakobsson - who helped me navigate through the different stages of this Thesis, and for introducing me to rigorous spectral analysis methods alongside providing me with initial steps in digital signal processing. Indeed, all of you supervisors have been the compass I needed to finish this Master's Thesis.

A special thanks is given to fellow students Katharina Papst and Erik Karlsson Strandh for your invaluable collaboration in early lab experiments which laid the groundwork for my further undertakings in audio processing. I thank Associate Professor Anders J. Johansson at The Department of Electrical and Information Technology for granting access to the anechoic chamber lab itself.

Also I would like to express my gratitude towards my fellow student Simon Söderlund for providing valuable insights when proof-reading this document.

To my friends and family, I thank you kindly for always supporting me, even though, perhaps, you could not always fully grasp what I was doing. Over the years, you have given me a sense of purpose, which today is manifested in this document.

- Carl Örnberg, Lund, June 2024

Abstract

The modern society is associated with widespread sound recording in public environments as well as in the workplace and at home, which motivates an increased use of speech anonymisation techniques in recorded audio. A demonstrative example is masking recorded speech in hospitals' waiting rooms for eavesdropping listeners inside a control room, thus ensuring privacy of information. This thesis evaluates Short-Term-Objective Intelligibility of offline speech separation and its cancellation from other background audio, maintaining the background sound. The intactness of background sound(s) is measured with a metric based on cross correlation. In addition, comparing the resulting masking effect due to different additive speech masking signals is performed, to evaluate their effect on intelligibility and perceptual classifiability.

Populärvetenskaplig sammanfattning

Anonymiserat tal i övervakning genom talmaskering och bakgrundsseparation

Det moderna övervakningssamhället är förknippat med omfattande ljudupptagning i offentliga rum, på arbetsplatsen och även i hemmet, vilket motiverar ökad användning av s.k. talanonymiseringstekniker. Att värna rätten till privatliv samtidigt som övervakningskapaciteten säkerställs är balansgången detta examensarbete tar itu med i två delar. Möjligheten att blint extrahera hela talsegment ur ljudupptagningar utan att fördärva bakgrundsljudbilden studeras först. Sedan presenteras en röstförvrängare liksom ett förslag på talmaskerare för att maskera vem som talar respektive vad som sägs. Båda med realtidsfunktion. Resultat utvärderas med s.k. objektiva mått för talbegriplighet samt ett mått baserat på korskorrelation.

Genom att använda oberoende komponentanalys på observationer från flertalet mikrofoner för att separera tal från bakgrundsljud demonstreras den stegrande svårighetsgrad när ljudkällor förflyttar sig i rummet över tid med varierande ljudnivå relativt till bakgrundsljudets. Examensarbetet finner emellertid att bevarandet av bakgrundsljud är enklare att uppnå än att utvinna själva rösten från ljudupptagningen. Detta antas härröra ur icke uppfyllda egenskaper hos modellen som oberoende komponentanalys antar, vilken hittar en ortogonal bas i observationsrummet som maximerar avståndet till Gaussiska fördelningar. Examensarbetet visar också genom sina resultat att talbegriplighetsmåtten måste anses som begränsade i sin överförbarhet till att mäta annat än begripligheten för brusreducerade talsignaler.

Contents

1 Background

1.1 Introduction - The Aim of Speech Anonymisation

This thesis presents the claim that there is some ambiguity to the problem description of anonymising speech. Some studies impose so-called *masking sound* on audio files and play them back in subjective speech intelligibility experiments, as in [1]–[3]. The aim being to reduce overall speech intelligibility, understanding what is being said. Others aim to *obfuscate speech* by employing real-time pitch-shifts to recorded audio, the aim being to anonymise the speaker identity but retaining speech intelligibility, as performed in [4]. Technology to transform voices of different genders to a generic gender substitute voice is a recent application presented by [5], utilising pitch-shifting. The trade-off of this dichotomy in problem description is that masking either intelligibility or identity implies use of various methods attaining different (possibly) incomparable results. Indeed, removing recorded speech is the ultimate anonymisation, attaining both of these goals. Possibly, a masking sound loud enough to mask both identity and intelligibility of speech could be attainable, but might be undesirable due to factors to be mentioned later. And implementing such a solution has an easy solution - simply turning off playback of audio during speech. This is the crux of this Master Thesis:

Ensuring that audio recordings are not compromising speech privacy of individuals, and not compromising the surveillance capability itself.

Three scenarios are presented and a simple scenario can be seen in Figure 1. In an airport security control room, surveillance personnel monitor the audio feed from various locations within the airport, such as terminals, baggage claim areas, and security checkpoints. The soundscapes contain public announcements, rolling of luggage, footsteps, crowd murmur and more identifiable speech of speakers closer to the surveillance device. Besides the need for ensuring speech privacy of these individuals, the need to identify other audio events such as a sudden loud noise can prompt security personnel to take immediate action.

Figure 1: The waiting room scenario for anonymising speech in surveillance.
Left: Two people are speaking in a surveilled waiting room.
Right: Two people in a control room eavesdropping on the recorded conversation, the speech waveform depicted in blue.

In a hospital emergency department, audio monitoring is used to ensure the safety and well-being of patients in waiting areas, treatment rooms and hallways. The recorded conversation and its specific speech contents between affiliates to patients and or hospital staff need not be conveyed to the listeners in a control room. However, hearing background sounds such as medical equipment alarms, patient distress calls and the general hustle or mood of the emergency room is vital. These sounds help medical staff quickly respond to emergencies, locate necessary medical personnel, and ensure that patient care is not compromised by focusing solely on individual conversations.

In a public transportation control centre, operators monitor the audio feed from buses, trains and stations to maintain safety and service efficiency. Again, anonymising speech of people's colloquial conversations must happen simultaneously as ensuring that operators can monitor other audio events. Background sounds such as engine noises, door operations, station announcements, and the general murmur of passengers provide essential information. For instance, detecting unusual mechanical sounds attributed to equipment issues needing immediate attention, like an accident, can help ensure smooth operations and rapid responses to incidents like accidents or emergencies.

Hence, one still needs to hear what happens besides speech activity, but since some classification itself can happen earlier in the processing pipeline, the resulting audio output should be classifiable to a human listener. This demands background audio to be as intact as possible after a speech anonymisation process, meaning audio cannot be temporarily silent during speech. The methods of attaining speech anonymisation is to:

- Perform speech separation using Independent Component Analysis on audio signals with both speech and background noise stimuli as well as other audio events.
- Impose masking sounds on resulting audio artefacts due to faulty speech separation.
- Mixing obfuscated speech audio back into the background sounds.

These methods will be evaluated through objective metrics of

- Speech Intelligibility
- Similarity of two audio files using normalised cross-correlation

and results are presented and discussed in corresponding sections. This will constitute the first part of this Master's Thesis. With successful speech separation, the speech signal can be discarded from the output audio for playback in a control room. However, this means no sign of speech from the audio itself can notify the human listener in the control room, listening back to the recording. The role of the *speech masker* becomes one of three phenomena.

- Anonymise speaker identity (preserving speech content)
- Render speech content unintelligible (possibly exposing speaker identity)
- Both of the above.

Communicating that speech is ongoing could already have been done earlier before speech separation, and by other means than manipulating audio, thus the audio manipulation to make speech content unavailable need not rely on the speech signal itself. Therefore the second part of this Thesis will

- Measure speech (un)intelligibility of manipulated speech signals

- Measure speech intelligibility of anonymised voice audio

The former in order to potentially demonstrate the restricted use of recognised objective speech intelligibility measures, and the latter in order to investigate how these measures of speech intelligibility perform on anonymised speech.

1.2 Human Speech, Hearing and Psychoacoustics

In order to understand the principles of speech processing, a brief overview of human speech and speech perception is merited. Voice is produced when airflow originating from the lungs passes through the vocal tract; the larynx, which houses the vocal cords and further through the pharynx, oral and nasal cavities. The passing of air induces, with longitudinal waves converted to energy, resonance in the vocal folds. These folds can be tightened or loosened to produce sound of different frequencies, and the vocal tract above the larynx, such as the jaw, tongue, soft palate and lips, formulates different sounds that indicate consonants and vowels. Voice production forms words, described by either syllables or in more detail by *phonemes*, smallest units of speech sounds. The voice itself contains a multitude of periodicities; a fundamental frequency F_0 and other pitches which are located at whole number multiples of F_0 with decreasing amplitudes called harmonics. This is the lowest periodicity of the sound estimated to range from roughly 90 to 130Hz typically for males or 180 to 220Hz typically for females. [6]–[8]

In the work in [9], human speech is introduced in order to introduce speech perception. It is explained that at rapid speech rate the number of phonemes per second can reach 30, a mean phoneme duration of roughly 33ms. Though, speech seldom exhibits a constant rate and is considered to be a more complex signal. The author further stresses that a given speech sound does not have a fixed acoustic pattern in the speech waveform. Instead, its acoustic pattern varies in a complex manner depending on the preceding and following sounds. Ongoing speech is not only dependant on cues present in the audio waveform; when a part of speech is replaced with an extraneous sound (e.g a coughing sound), listeners still recall the missing sound, studies found. This introduces the topic of human speech perception. [9]

When sound enters the ear and upon reaching the cochlea in the inner ear, certain electrochemical events performs filtering and processing of the sound, such that the mechanical longitudinal waveform is translated into electrical waveform. Frequency, intensity and time information is obtained during this mechanical to electrical conversion, and the intricate design yields a human hearing frequency range of about 20 to 20000Hz. The audibility of a sound depends also on its duration in time and intensity, where e.g. minimum duration for audibility is around 16 ms for 14dB for a 250Hz sine wave. [7]

Psychoacoustics is the empirically-based science dealing with the psychological correlates of the physical parameters of acoustics. A relevant prelude to speech masking is the

phenomenon of *auditory masking*, relying on the sub-phenomena called *tonal* and *temporal masking*, covered in both [7] and [9]. The initial setting and resulting effects are similar and explain the correlation between an observed need for higher intensity threshold for a target signal to be audible, and, the introduction of a so-called *masking signal*. The former by configuring the masker signal, or tone, to possess neighbouring frequency peaks to the target signal (like a rectangle window around target signal's dominating frequencies, then referred to as *noise masking*), and the latter by placing the masking signal in time just before onset or after offset of the target signal (called *forward* and *backward* masking). Forward masking has been found to be effective at a $75 - 100$ms difference between the two audio stimuli, whereas backward masking is effective only up to a 50ms difference. Presence of beats and combination tones affects the precision of these quantities. The human hearing using both ears affects the zone of masking. What masking a complex target signal such as speech entails is to be elaborated upon later in this thesis, due to its association with speech intelligibility; introduced below.

1.3 Speech Intelligibility as an Objective Measure

Speech Intelligibility is here defined as a measure of the comprehensibility of speech, the primary method of communication between humans, transmitted through some channel (like its direct path in air to the listener, or through a wire, digital conversion, etc). The transmission channel influences the speech intelligibility by possibly introducing reverberation, echoes, background noise, altered speech levels, distortion or other auditory phenomena chalked up to psychoacoustical events. Traditionally, speech intelligibility has been measured subjectively using perceptual judgements, in which listeners scored the clarity and understandability of speech samples according to some scoring format. However, such subjective ratings are time-consuming and susceptible to biases caused by individual listener variability and environmental factors, stated in [10]. To mitigate these limitations, objective measures of speech intelligibility have garnered significant attention within the scientific community. Objective Intelligibility Measures (*OIMs*) aim to quantify speech intelligibility using computational algorithms and signal processing techniques, thus allowing standardised and reproducible assessments independent of listener biases. One such example is the international standard Speech Intelligibility Index (IEC 60268-16:2020, *STI*) [11], where a numeric representation of communication channel characteristics ranging from 0 to 1 denoting qualities *bad, poor, fair, good, excellent* is used. It is calculated based on factors including modulation transfer function, speech-to-noise ratio, and reverberation time.

In 2010, earlier experimental findings on STI and other OIMs were examined and concluded to be suitable for several types of degradation, yet less suitable for processing noisy speech by a time-frequency weighting, like speech separation. A new OIM was presented in [12], which displayed high correlation with subjective speech intelligibility data, called *Short-Term Objective Intelligibility* (STOI). It relies on access to a reference ground truth signal. The method is designed for a sampling rate at 10kHz to cover the human speech spectrum, and signals with other sampling rates must be resampled. In addition the two digitally-converted signals to be compared has to be aligned in time. The reference recording called *clean speech* has to be available for comparison to the *denoised speech*.

Later, an Extended STOI (ESTOI) measure was introduced in [13], providing a better intelligibility measure for denoised speech signals contaminated with temporally-modulated noise maskers; the noise is fluctuating over time and causing the aforementioned masking phenomena. The authors state that noise can be another speaker, referring to the non-linear, highly temporal modulating characteristics of a speech signal. It is thus believed to be a better measure for separated speech in soundscapes with both speech and background sounds. Its calculation is analogous to STOI and further details are presented in Section 2.3.

1.4 Speech masking

1.4.1 Overview

In a notable early paper, office staff in two separate offices were subjected to masking sounds, consisting of synthesised static noise at sound levels $45 - 51$ dBSPL (decibel relative to sound pressure level) with frequency range $100 - 10000$ Hz, throughout their workday during several weeks. Simplified, a listener could hear a speech signal alongside an imposed external noise, which impaired the ability to understand the content of the speech signal. The paper found primarily relevant that introducing static noise in the workplace caused three effects:

- It annoyed several subjects when noticeable (especially when forewarned).
- Greater sound levels induced lower speech intelligibility scores.
- A frequency range around 2000 Hz induced lower speech intelligibility scores.

The scoring was done subjectively for about a dozen volunteers. [14]

More recent previous research on speech masking, including [1]–[3], [7], [15], has developed and evaluated several novel masking sounds. They are generally divided into two classes - *energetic* and *informational* masking. The former occurs when the imposed background noise has (for the listener) greater excitation or neural response in a given frequency spectrum than the original speech. The latter is achieved when the imposed sound prevents the listener's ability to distinguish specific spectro-temporal areas of sound to understand the context of the original speech. Spectro-temporal areas include both phonemes and speech cues. The underlying psychoacoustical events are thought to be the spectro-temporal masking mentioned in Section 1.2. Commonly studied sound maskers include both stationary background noise, such as white or pink noise as well as HVAC sounds (*Heating, Ventilation, Air Conditioning*), and fluctuating noise, such as so-called imposed *babble* noise as in [16], artificial speech sounds, or other *speech-like sounds*. Sometimes, reverberation is added to the latter type to decrease perceived irritation or disturbance for the listener (referred to as *annoyance*). Regardless of its class and in accordance with [14] findings, speech maskers can be formalised to primarily have the effects of reducing the intelligibility of speech and secondarily inducing the emotional response of irritation or disturbance to the listener. This is the reason previous publications on the subject consider that an *optimal speech masker* in priority minimises

1. Speech intelligibility
2. Induced irritation

However, measures of perceived distraction are sometimes introduced, especially for speech-like maskers that can exhibit discontinuities in the masking sound. Also, the so-called T*arget-to-Masker Ratio* - the original speech's sound level related to the masker's sound level in decibels, has been reportedly shown to be on the one hand proportional to better masking but on the other hand to increase the level of irritation. In contrast to the first minimisation objective, induced irritation is not a well defined concept, wherefore considering the *A-weighting* of a reference and the masking sound is a possibility. A-weighting is a frequency weighting filter that adjusts the sensitivity of a sound level meter, matching human ear's sensitivity to different frequencies. It mimics the response of the human ear to sound, particularly at moderate to high levels and is outlined in [17]. Irritation is not solely dependant on frequency and intensity, and the unmeasured temporal structure is an unaccounted for factor. [16]

1.4.2 Speech-like maskers

Time-reversed speech masking is a type of informational masking where each speech segment is divided into smaller segments (*frame(s)*), for example in the range 160 − 500ms. Subsequently, the order is flipped for each smaller sound segment and is imposed again on the original audio. So the process occurs in segmentation, time reversal, and imposition on the original speech segment. This method has proven to be more effective at reducing intelligibility than other speech-like sounds such as default babble noise, probably because the masking by definition possesses similar spectral components as the target speech and nullifies the ability to distinguish phonetic features and understand overall context. Time reversal functions less well on palindrome words, since such words exhibit symmetry in inversion. One way to counteract this more rare case is to randomise the imposition of audio segments. Resulting irritation has been judged to be somewhat less in estimation with more compact dispersion scores than previously mentioned maskers, but a statistically significant difference in irritation levels between time reversal and, for example, pink noise or babble noise has not been achieved. A challenge with time reversal has been adapting the method to real-time format. Here, the size of the pieces in the segmentation (buffer for the next speech segment) plays a role as well as the inability to completely randomise the imposition while the masker records speech segments. A proposed method (called *'OLaW'*) in [18] has been to detect and mark pitches (*pitch detection*) through calculation of short-term energy spectrum of incoming segments and place with overlap the reversed segment in phase with the next 'frame' (and possibly add echo), which has been both feasible and effective as a real-time speech masking solution, demonstrated in [19].

2 Theory

2.1 Speech Removal in the context of Blind Source Separation

The objective of isolating several simultaneous speech sources within a mixed signal might contain background or additive noise or music is known as *Speech Separation*. It falls under the broader category of *Source Separation* where general mixed signals are isolated using no, some or sufficient *a priori* signal information. In the first case it is referred to as *Blind Source Separation* (BSS), in the other case *Semi-blind Source Separation* and in the latter case one might even know the sought signal. The difficulty of reaching sufficient separation is inversely related to the amount and specifics of *a priori* information. In the case of speech separation, the complexity of modelling speech signals due to their inherent elusive various forms (pauses in speech, rate, pitch,

inter-human phoneme variation) means separating speech from elaborate background sounds is at most semi-blind source separation. [20]

The following sections will introduce the general problem formulation, its translation to the specific problem, a widely used solution method relying on concepts in information theory, exploitable audio characteristics that become relevant to these concepts, and finally, the final model for speech-from-background separation is presented. In addition, the performance metrics are presented in more detail.

2.1.1 General Problem Formulation

The BSS problem consists of separating and retrieving unobservable *source signals*, denoted in vector notation as

$$\boldsymbol{s}(t) = (s_1(t), \ldots, s_N(t))^T \in \mathbb{R}^N,$$

assuming zero mean and stationarity from *observed mixtures*

$$\boldsymbol{x}(t) = (x_1(t), \ldots, x_P(t))^T \in \mathbb{R}^P,$$

which can be expressed as

$$\boldsymbol{x}(t) = \mathscr{A}(\boldsymbol{s}(t)) + \bar{b} \tag{1}$$

where $\mathscr{A}$ is an unknown (mixing) mapping from $\mathbb{R}^N$ in $\mathbb{R}^P$, and $\bar{b}$ is some additive noise. Often, signals are pre-whitened and $\bar{b}$ is neglected. The symbol t denotes here time index, but can generally be considered as a sample index. The invertibility of the *mixture* $\mathscr{A}$ often necessitates $N \leq P$, meaning that identification of

$$\mathscr{B} := \mathscr{A}^{-1}$$

directly leads us to source separation, and provide us with the *estimated sources*

$$y_i = k_i\big(s_{\sigma(i)}(t)\big),\ i = 1, 2, \ldots, N, \tag{2}$$

where σ being a permutation in $\{1, 2, \ldots, N\}$, and k_i being a mapping corresponding to a residual distortion. The indeterminacy of BSS is displayed in Equation (2). The mixture's nature affects factors that in turn determine separation quality. Note also that this equation asserts the ambiguity of imposing an index set on the sources and observations; there is no guarantee that estimated source at index i corresponds to the ith unobservable source. This might be trivial in this theoretical setting but an essential insight that needs to be handled in real-time digital signal processing (DSP).

This ambiguous solution mapping is referred to as the **permutation problem**, and the separations are scaled as a result of separation, also of concern for a DSP setting. [20]

From this outlook, the BSS problem seems ill-posed since further information about the *mixture* itself and the relationship between observations themselves is not specified. This is where extraneous assumptions have to be made. Let us therefore begin to formulate some specifics of mixtures.

The mixing transformation can be either linear and non-linear, instantaneous or convolutive. [20] The linear instantaneous mixture, where signals arrive at the microphone simultaneously, is described by a matrix transformation

$$\bar{\boldsymbol{x}}(t) = \boldsymbol{A}\bar{\boldsymbol{s}}(t).$$

Linear convolutive mixtures, on the other hand, indicate sources at discrete time instance n exhibit a difference of arrival (DOA) as well as reverberations due to space confinement, resulting in replacing the mixing matrix $\boldsymbol{A}$ with a linear time-invariant system (LTI) with finite impulse response. Instead of a linear combination at each time instance, the sound signals' delayed values contribute to the mixture at a given time. Thus, for a given time instance n with some support set indicating the decaying reverberation $\mathbb{M} := \{-M, \dots, M\}$, one of the N sound sources are observed as

$$\boldsymbol{x}(n) = \sum_{k \in \mathbb{M}} \boldsymbol{A}(k)\boldsymbol{s}(n-k), \tag{3}$$

leading to a multichannel convolution model. Here the mixing is a convolutive finite impule response $(\boldsymbol{A(n)})_{n\in\mathbb{M}}$. The multiple-input-multiple-output (MIMO) LTI system's inverse system (henceforth referred to as the *separator*) with impulse response $(\boldsymbol{B}(n))_{n\in\mathbb{M}}$, retrieves the sources, provided the mixing filter is stable, causal and has finite impulse response. One should consult [21] for further details. The support set $\mathbb{M}$ enables the model to handle future and previous time instances values since the microphone ordering is not specified, meaning DOA leads to either positive or negative inter-microphone translation in time. See Section 2.2.2 regarding DOA detection. The separated outputs take the form for each estimate $(y_1(n), \dots, y_P(n)) = \boldsymbol{y}(n)$ at a given time instance n

$$\boldsymbol{y}(n) = \sum_{k \in \mathbb{M}} \boldsymbol{B}(k)\boldsymbol{x}(n-k). \tag{4}$$

The global output at the separator then becomes from (3) and (4)

$$\forall n \in \mathbb{M} \qquad \boldsymbol{y}(n) = \sum_{k \in \mathbb{M}} \boldsymbol{G}(k)\bar{\boldsymbol{s}}(n-k), \tag{5}$$

$$\forall n \in \mathbb{M} \qquad \boldsymbol{G}(n) = \sum_{k \in \mathbb{M}} \boldsymbol{A}(k)\boldsymbol{B}(n-k) \text{ and } \boldsymbol{G}[z] = \boldsymbol{B}[z]\boldsymbol{A}[z], \tag{6}$$

where $\boldsymbol{G}$ is the global system of the inverse MIMO-LTI and notation $[z]$ refers to a transform in the frequency domain, like a DFT or z-transform, as mentioned in [21].

2.1.2 Semi-Blind Speech-from-Background Separation

The translation of the problem description to an digital audio processing problem translates sources $\boldsymbol{s}(t)$ to sound sources, and observations $\boldsymbol{x}(t)$ to be the recorded audio of the N sounds as they propagate in air through direct and reflective paths to the P sensors, or *microphones*. The sources themselves can be classified to be speech or non-speech background audio,

$$\bar{\mathbf{s}}(t) = \left(s_k\right)_{k=1}^{N} = \left\{v_1, \ldots, v_{N_v}, b_1, \ldots, b_{N_b}\right\} \in \mathbb{R}^N \,, \tag{7}$$

having N_v number of *voiced signals* $\bar{v}(t)$, simultaneous with the composite *background signal* $\bar{b}(t)$ (N_b number of sources). The sources are observed by being recorded in an unknown mixture $\mathscr{A} : \mathbb{R}^N \mapsto \mathbb{R}^P$ at the P microphones, analogous to (1)

$$\bar{\boldsymbol{x}}(t) = \mathscr{A}(\bar{\mathbf{s}}(t)) \in \mathbb{R}^P \tag{8}$$

where the P *observed signals* $\bar{\boldsymbol{x}}(t)$ are the mechanical sound waves propagating into the microphones. The speech sources are assumed to be created by the process mentioned in Section 1.2. Audio mixtures are usually linear and convolutive [22], and if we infer that $\mathscr{A}$ to be time-invariant, Equation (5) and (6) are readily available. However, when dealing with moving sources of speech, the problem's complexity increases, causing the time-invariance property relied upon by the separator to be compromised. Especially when considering longer durations. More precisely, the approximation of an MIMO-LTI system breaks down as $\boldsymbol{G}$ changes over time, as described in [21], [23]. This introduces the notion of dynamic (also referred to as *adaptive*, *recursive*, *online*) MIMO systems.

Considering the single surveillance camera setting, we ought to expect that there are more sound sources than number of microphones, that $N_v + N_b \geq P$, which usually is between one and four. The single channel case is the most extreme example, often solved by neural nets, as described in [23], and is discarded for reasons presented below. Thus one should always assume the underlying real-world model to be *under-determined*, necessitating exploitation of useful audio characteristics into our next model, presented after covering the methods to solve BSS.

2.2 Solutions to BSS

2.2.1 Spatial whitening, Information Theory, Component Analysis

The noiseless model in (8) is achieved by spatial whitening the observations, which is often performed with *principal component analysis* (PCA), a dimension reduction technique linearly transforming real-valued data (our observations) such that directions capturing the largest variations in the data (principal components) are pairwise orthogonal. [24] Consider an $m \times n$ data matrix, being column-wise zero empirical mean, where each row represents an instance (e.g an observation from microphone one, two etc.). The iterative method yields the full principal components decomposition as

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$$

where $\boldsymbol{W}$ is $n \times n$ weight matrix whose columns are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. Performing PCA as a spatial whitening not only reduces search space for a separation matrix in BSS, it also reduces the effect of additive noise during separation. [20]

Thus pre-whitening observations through PCA ought to result in a noiseless model. Of course, the composite background sounds might possess additive noise, but they are characterised as less important to other non-stationary audio events and can be omitted at this point. We are left with what is sought to be noiseless observations. A question that might arise to the reader is what further assumptions one might impose on the sources. Indeed, the question arose for and was answered by the pioneers of what became known as *Independent Component Analysis* (ICA) and more interestingly the *FastICA* method in the 1980s and 1990s (algorithm presented in Section 3). Solutions to BSS using ICA-methods depend on an optimisation criterion, referred to as *contrast function*. Due to blindness, the solutions have to rely solely on observations and metrics like the mean-squared error with ground truth or similar are unavailable. In order to find suitable contrasts, one begins with imposing the following fundamental assumptions, formulating the signal sources as real-valued stochastic random variables:

- Sources $\bar{s}$ are mutually temporally statistically independent. The joint probability density function (pdf) can be factorised with the marginal pdf:s:

$$p_s(s_1, s_2, \ldots, s_N) = p_1(s_1)p_2(s_2)\ldots p_N(s_N). \tag{9}$$

- Sources $\bar{s}$ are non-Gaussian, or at most one source is Gaussian:

$$\begin{aligned} \forall i = \{1, \ldots, N\} \quad & s_i \not\sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{or} \\ \exists! j : \quad & s_j \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad \forall i \neq j, \quad s_i \not\sim \mathcal{N}(\mu_i, \sigma_i^2). \end{aligned}$$

The first assumption indicate that for two (independent) sources s_1, s_2 and given two functions f_1, f_2, that

$$\begin{aligned}\mathbb{E}[f_1(s_1)f_2(s_2)] &= \int\int f_1(s_1)f_2(s_2)p_s(s_1,s_2)d_{s_1}d_{s_2} \\ &\overset{(9)}{=} \int\int f_1(s_1)p_1(s_1)f_2(s_2)p_2(s_2)d_{s_1}d_{s_2} = \int f_1(s_1)p_1(s_1)d_{s_1}\int f_2(s_2)p_2(s_2)d_{s_2} \\ &= \mathbb{E}[f_1(s_1)]\mathbb{E}[f_2(s_2)]\end{aligned}$$

further implying when the functions are the identity operator $f_1(s) = f_2(s) = s$, that

$$\mathbb{E}[s_1, s_2] - \mathbb{E}[s_1]\mathbb{E}[s_2] = 0, \tag{10}$$

implying that the covariance between the sources is zero. This then indicates uncorrelatedness, which is a weaker form of independence. FastICA assumes independent sources, implying uncorrelatedness, but the reverse is not true in general. The second assumption is necessary to be able to obtain separation; Hyvärinen, the pioneer of FastICA, showed in [25] that more than one independent Gaussian source prohibits identifying the mixture $\boldsymbol{A}$, due to an symmetric ambiguity in the joint pdf between observations. Consequently it was shown that non-gaussianity can serve as a proxy for statistical independence in [26]. The distance to normality can be given by the *negentropy*, which is borrowed from the field of (Differential) Information Theory, as presented in [27]. It measures the difference in *entropy* between a given distribution and the Gaussian distribution with the same mean and variance. The following definitions are stated below and one can consult the description of certain *contrast functions* that become the metric which optimisation is performed with in [28].

Definition 2.1 (Shannon Entropy)**.** *Let X be a discrete random variable, with state-space $\mathcal{X}$ which is distributed $p : \mathcal{X} \mapsto [0, 1]$, then its entropy is*

$$H(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x),$$

where the choice of log-base determines unit; e.g base 2 yields 'bits' and base 10 yields 'dits'. In the case of a continuous random variable the

Definition 2.2 (Negentropy)**.** *Given a discrete random variable X, with given mean μ_X and variance σ_X^2 and with state-space $\mathcal{X}$ which is distributed $p : \mathcal{X} \mapsto [0, 1]$, then its negative entropy, its negentropy is*

$$J(X) = H(X^*) - H(X),$$

where H is the Shannon entropy and X^* is a random variable $X^* \sim \mathcal{N}(\mu_X, \sigma_X^2)$.

It is known for a specified mean and standard deviation that the Gaussian distribution has maximum entropy, resulting in $J(X) \geq 0$ with equality if and only if X share the same Gaussian distribution $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. The *FastICA* method utilises the following approximation.

Corollary 2.2.1 (Approximation of Negentropy). *[25] Assuming X is a r.v with zero-mean and unit variance and similarly $X^* \sim \mathcal{N}(0,1)$. Then the negentropy of X can be approximated with*

$$J(X) = [\mathbb{E}[G(X)] - \mathbb{E}[G(X^*)]]^2,$$

where G is a non-quadratic and preferably slowly increasing function, as suggestions for $G(x)$:

$$G_1(x) = \log\cosh \alpha x, \ \alpha \in [1,2] \tag{11}$$

$$G_2(x) = -\exp -\frac{x^2}{2} \tag{12}$$

$$G_3(x) = x^3 \tag{13}$$

The FastICA algorithm finds independent components in the mixture by iteratively finding weight vectors that maximise nongaussianity of separated signal estimates. The negentropy is a measure of nongaussianity serving as a proxy for independence between sources. The suggestions of the non-quadratic function G displayed in Equations (11)-(13) has all been shown to be useful in several applications but G_1 is considered by [25] to be robust.

2.2.2 Exploitable Audio Characteristics

Speech and Frame of Reference

Speech exhibits as mentioned natural speech pauses and local stationarity (consider vowels which are periodic or some phonemes corresponding to consonants which are transient) and therefore the independence assumption only holds when long enough consecutive time frames are considered. Validation of earlier claims on temporal independence in speech signals (audio recordings of speech) has been performed through simulations in [22]. With the Shannon Entropy defined, its link with the concept of *mutual information* (MI) is presented in accordance with [29].

Definition 2.3 (Mutual Information). *Let $s_1, \ldots, s_N$ be random variables $\mathbf{s}$ taking finite amount of values with defined pdfs $p_{s_1}, \ldots, p_{s_N}$ alongside their joint pdf $p_{s_1,\ldots,s_N}$, then their*

mutual information is given by

$$I(s_1, \ldots, s_N) = -\mathbb{E}[\log \frac{p_{s_1}(s_1) \ldots p_{s_N}(s_N)}{p_{s_1,\ldots,s_N}(s_1, \ldots, s_N)}] = \sum_{i=1}^{N} H(s_i) - H(\boldsymbol{s}) \geq 0.$$

The term $H(\boldsymbol{s}) = H(s_1, \ldots, s_N)$ is the joint entropy; the uncertainty associated with the set of the random variables. We note that when mutual independence is reached, that $\log 1 = 0$; their mutual information is zero. Corollary 2.2.1 was derived by [25] by first showing

$$I(s_1, \ldots, s_N) = \text{const.} - \sum_i J(s_i)$$

when s_i are uncorrelated. Thus minimisation of MI and maximisation of negentropy can both serve as proxies for statistical independence. Note also $\boldsymbol{s} = \bigcup_{i=1}^{N} s_i$; with Definition 2.3, one defines the sub-sequences (defined by window length) as random variables, whose MI can be measured, to represent the weakening dependence over time. A study of speech signals found that MI decreased sufficiently (MI< 0.05) when time-frames were greater than roughly 100ms for speech signals, suggesting speech signals display dependency under this threshold. This suggest a crude estimate on minimum duration of at least 100ms to perform FastICA on. Although a longer signal duration might be preferable due to the rough estimation and to discern speech contents. The limit can thus be considered for dynamic or online ICA methods. [20]

Direction of Arrival

As stated in Section 2.1.2, audio mixtures are usually linear and convolutive. The audio sources arrive at the microphones from different DOA:s, and one could detect a single source's DOA through a simple time-delay *beamformer*.

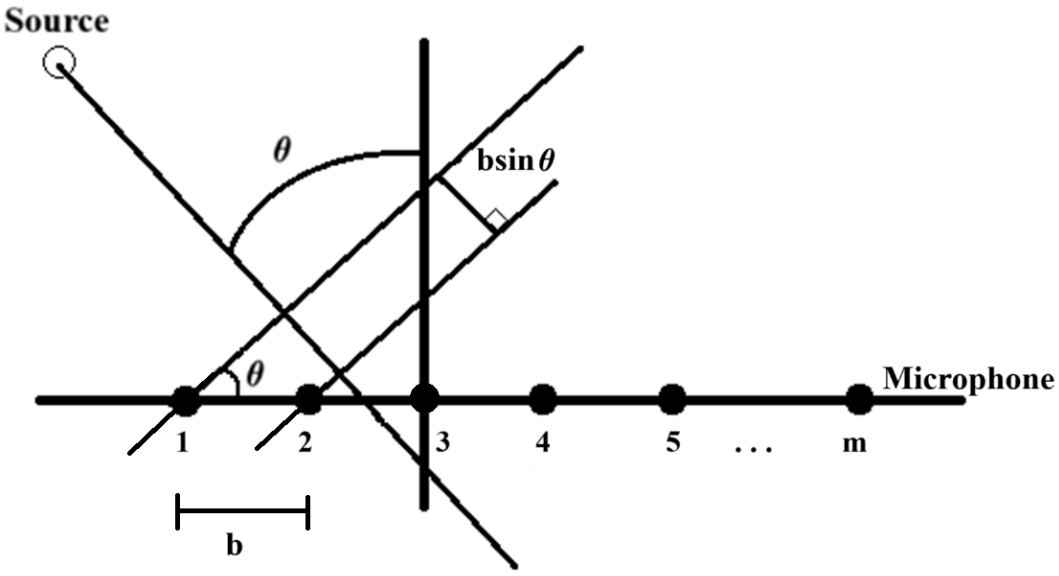


Figure 2: Sketch of DOA denoted by θ. Evenly placed m microphones along a single line are said to all have the same θ when the source is located far enough away.

Finding an expression for the DOA denoted by θ will be shown and Figure 2 displays the setting. The following is based on the problem outline in [30]. If one assumes the sound source to be located further away such that the arriving sound waves are at the m microphones assumed to be essentially parallel, then one would obtain a microphone *uniform linear array* (ULA) as shown in Figure (2). In this far-field scenario, we have the *time difference of arrival* (TDOA) or also referred to as the *time delay* between two microphones $\tau = t_1 - t_2$, where t_1, t_2 are the arrival times. With a constant wave propagation speed v one obtains θ through symmetry as:

$$\sin\theta = \frac{v\tau}{b} \leftrightarrow \theta = \arcsin\frac{v\tau}{b} \tag{14}$$

Equation (14) holds for an ULA of at least two microphones. When several sound sources (including noise) reach the ULA in the same time frame, the result of digital conversion is that the loudest signal possibly masks other signals during recording, and one would require for multi-tracking DOA to keep track of different DOA's over time, and distinguish between movement of one source and when a new source enters the system. In our scenario, we have one such speech target.

One should note that Equation (14) has two solutions for every τ due to symmetry, e.g the mirrored angle in the lower half plane in Figure 2. This ambiguity vanishes with

placement spanning the desired dimension, according to [30]. What is interesting to determine is whether the effects of time-delays can be neglected when using FastICA. Estimating τ from two audio channels of some duration $\bar{x}_1$, $\bar{x}_2$ e.g those produced by for example microphone 1 and 2 in Figure 2 is performed through cross-correlation

$$\hat{\tau} \xleftarrow{i \text{ array index}} \arg\max_i (\bar{x}_1 * \bar{x}_2)(i) \tag{15}$$

where $\hat{\tau}$ is obtained through finding the corresponding time of the maximally correlated index which is dependant on the sample rate. Figure 3 displays the observable DOA when sampling with 48 and 16kHz respectively, with $b = 0.10$m.

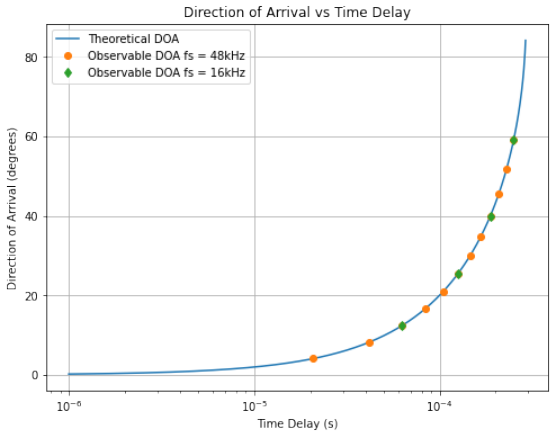


Figure 3: Theoretical and accessed DOA values from (14) and (15), for a quarter-plane $\theta \in [0, 90)$, with $b = 0.1$m.

The sign of the $\hat{\tau}$ translates the θ to the domain $(-90, 90)$, and it is noted that sampling audio linearly implies one has limited visibility. For the discussed common sampling rates, the maximum observable lag is 250μs as seen in Figure 3. This is to illustrate that DOA-values numerically estimated from (15) at sampling rate 48kHz only gives a range in $(4°, 68°)$, corresponding to one through thirteen indices in a lag vector. This means that a maximum lag index of thirteen at this sampling rate may refer to an angle beyond $68°$.

Alongside an inter-microphone intensity difference it has been stated that time-domain (over)determined convolutive ICA is consistent to audio data, however for underdetermined convolutive mixtures, one has to employ acoustic mixing filters, e.g. finite room-impulse-response filters to mitigate reverberation. Again such filters become dynamic in the presence of moving sources. The separation performance over stereo

audio mixtures has been observed in [22] to decrease dramatically with increasing reverberation time, decreasing microphone distance and diffuse interfering sources - even sound wave propagation in a room, leading to an equal reverberation time at any listening position.

2.2.3 Final Offline MIMO model

A reminder is in order:

The separation of voiced audio from background sounds need not be perfect; the aim is that one of the separator outputs is perceived by human ears as containing only background sounds, implying that another channel should contain the ongoing speech.

Thus if one separated channel contains a majority of the speech signal alongside some background audio, while the other contains only background sound with artefacts due to speech signal, the performance is deemed sufficient. The balance between separation quality and computational efficiency is chosen to favour the latter , and the final model can now be presented.

The mixture, with a duration of several seconds, is assumed to be linear and convolutive after pre-processing, where effects of reverberation and DOA of sources are neglected, assuming the inherent maximum time delay of 250 microseconds does not deteriorate source estimation severely. The sources are assumed to become classifiable analogous to (7) such that $\bar{\boldsymbol{s}} = (\bar{\boldsymbol{v}}, \bar{\boldsymbol{b}})$ (voiced audio and background audio) and the number of microphones $P \geq 2$ now equal the number of assumed sources (or is greater than them), forming a (over)determined system. It is assumed $\bar{\boldsymbol{v}}$ and $\bar{\boldsymbol{b}}$ are temporally independent and $\bar{\boldsymbol{v}}$ non-Gaussian and $\bar{\boldsymbol{b}}$ having at most one Gaussian noise component. Invertibility of mixing matrix is assumed guaranteed. The resulting model that FastICA is performed on becomes

$$\begin{aligned} \bar{\boldsymbol{x}} &= \boldsymbol{A}\bar{\boldsymbol{s}} \\ \bar{\boldsymbol{y}} &= \boldsymbol{W}\bar{\boldsymbol{x}} \\ \boldsymbol{W} &= \boldsymbol{A}^{-1} \end{aligned}$$

where $\bar{\boldsymbol{y}}$ are the estimated sources that are obtained with an unknown permutation. Furthermore, the assumption that nongaussianity of $y_1, y_2 \in \bar{\boldsymbol{y}}$ can be measured by Corollary 2.2.1 and is assumed to be greater than all other estimation candidates. Due

to sufficient separation, one column vector in $\bar{\boldsymbol{y}}$ is assumed to contain $\bar{v}$, and analogous to (2) the speech signal is obtained as

$$y_i = k_i\big(\bar{v}(t)\big),$$

where the distortion k_i is assumed to be due to the Gaussian component in $\bar{\boldsymbol{b}}$.

2.3 Performance Metrics

The separator yields scaled separated signals, possibly a signal can be scaled with a negative sign, which would imply minimum correlation approaching -1. Such a scaled separation signal would still imply successful separation, thus to disambiguate the concept of background intactness (BI) the measure is defined below.

Definition 2.4 (Background Intactness)**.** *Let $\boldsymbol{b}$ be the ground truth background signal vector of length N in the reference window of a separation. Let one of the separator outputs $\boldsymbol{y} \in \boldsymbol{y_1} \dots \boldsymbol{y_P}$ be the estimate of $\boldsymbol{b}$. Then the Background Intactness (BI) is the magnitude of the normalised cross-correlation, taking the form*

$$BI(\boldsymbol{b}, \boldsymbol{y}) := \left| \frac{\sum_i^N b_i y_i}{\sigma_{\boldsymbol{b}} \sigma_{\boldsymbol{y}}} \right| \mapsto [0, 1],$$

where $\sigma_{\boldsymbol{b}}$ and $\sigma_{\boldsymbol{y}}$ denote standard deviation.

The detailed calculations of STOI and ESTOI will be presented next and are the ones described in [12] and [13]. Suppose we have access to the time-aligned clean and processed speech, denoted by x and y, which have been resampled to 10 kHz.

STOI: A STFT is calculated by segmenting both signals with 50% overlap, using Hanning windows of length 256, where each frame is zero-padded with 512 further elements. Consequently a grouping of DFT-bins obtains 15 one-third octave bands, where the lowest frequency becomes 150 Hz. Let $\hat{x}(k, m)$ be the kth DFT-bin of the mth frame of the clean speech signal. Then, we define the ℓ^2-norm of the jth one-third octave band as a *unit*,

$$X_j := \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k, m)|^2}, \tag{16}$$

with k_1 and k_2 denoting the one-third octave band edges, rounded to the nearest DFT-bin. Similarly we substitute $X_j(m)$ with $Y_j(m)$ for the processed speech signal. With this defined, we can formulate a intermediate intelligibility measure for one unit, $\rho_{X,Y'}(j,m)$, where $Y \mapsto Y'$ by a normalisation to the clean speech and clipping to lower bound the signal-to-distortion ratio (See Equation (2) and (3) in [12]). We now define $\rho_{X,Y'}(j,m)$ thoroughly as

$$\begin{aligned}\rho_{X,Y'}(j,m) &= \frac{\mathbb{E}[(X-\mu_X)(Y'-\mu_{Y'})]}{\sigma_X \sigma_Y} \\ &= \frac{\sum_n \left(X_j(n) - \frac{1}{N}\sum_l X_j(l)\right)\left(Y'_j(n) - \frac{1}{N}\sum_l Y'_j(l)\right)}{\sqrt{\sum_n \left(X_j(n) - \frac{1}{N}\sum_l X_j(l)\right)^2 \sum_n \left(Y'_j(n) - \frac{1}{N}\sum_l Y'_j(l)\right)^2}},\end{aligned}$$

where $n, l \in \mathcal{M} := \{(m-N+1), (m-N+2), \ldots, m-1, m\}$ are N consecutive frames and where μ_X denotes the arithmetic mean as an estimate of $\mathbb{E}[X]$, and finally σ_X denotes the standard deviation. Analogously the mean and standard deviation holds for the processed speech. Finally, the STOI measure is computed as the average of the intermediate intelligibility measure over all bands and time frames, as

$$STOI(x,y) := \frac{1}{JM}\sum_{j,m} \rho_{X,Y'}(j,m) \mapsto [-1,1],$$

where the intermediate speech intelligibility is calculated for the mth frame out of totally M frames, the jth one-third octave band out of J bands, due to initial DFT. The resulting STOI ranges from -1 to 1, indicating no intelligibility to perfect intelligibility of processed speech.

ESTOI: With x and y as before, Equation 16 is unaltered but instead of transforming Y to Y' as before, one collects spectral values for every frequency band $j = 1, \ldots, J$ across N consecutive time segments ($\mathcal{M}$ as before) - yielding

$$\begin{aligned} X_m &= \begin{pmatrix} X_1(m-N+1) & \ldots & X_1(m) \\ \vdots & & \vdots \\ X_J(m-N+1) & \ldots & X_J(m) \end{pmatrix} \\ Y_m &= \begin{pmatrix} Y_1(m-N+1) & \ldots & Y_1(m) \\ \vdots & & \vdots \\ Y_J(m-N+1) & \ldots & Y_J(m) \end{pmatrix}, \end{aligned}$$

where the jth row of one of the matrices correspond to the temporal envelope of the signal in sub-band j. Both these short-time spectrogram matrices of the clean and

processed speech signals are subsequently mean- and variance-normalised according to first rows and then columns. Let

$$x_{j,m} = \begin{pmatrix} X_j(m-N+1) & X_j(m-N+2) & \dots & X_j(m) \end{pmatrix}$$

denote the jth row of X_m. The jth mean- and variance- normalised row becomes then

$$\bar{x}_{j,m} = \frac{1}{\sqrt{(x_{j,m} - \mu_{x_{j,m}})^T (x_{j,m} - \mu_{x_{j,m}})}} (x_{j,m} - \mu_{x_{j,m}} \mathbb{1}),$$

where $\mathbb{1}$ is a ones-vector and $\mu_{x_{j,m}}$ is the sample mean $\frac{1}{N} \sum_{m'=0}^{N-1} X_j(m-m')$. We obtain the normalised matrices $\bar{X}_m$ and $\bar{Y}_m$ in this fashion, and the intermediate intelligibility index related to the mth time segment d_m is defined as the projection of the processed signal onto the clean vector

$$d_m = \frac{1}{N} \sum_{n=1}^{N} \bar{s}_{n,m}^T \bar{x}_{n,m}$$

and finally the temporal average of d_m gives the resulting ESTOI measure

$$ESTOI(x,y) := \frac{1}{M} \sum_{m=1}^{M} d_m(x,y) \mapsto [-1,1],$$

where M is the number of time segments in the signal(s). Further intuition are given by details mentioned in [13].

2.4 Voice Extraction and Masking

Figure 4 display the simple offline process of applying speech masking to the separated speech signal, referred to as *Voice Extraction and Masking* (VEM).

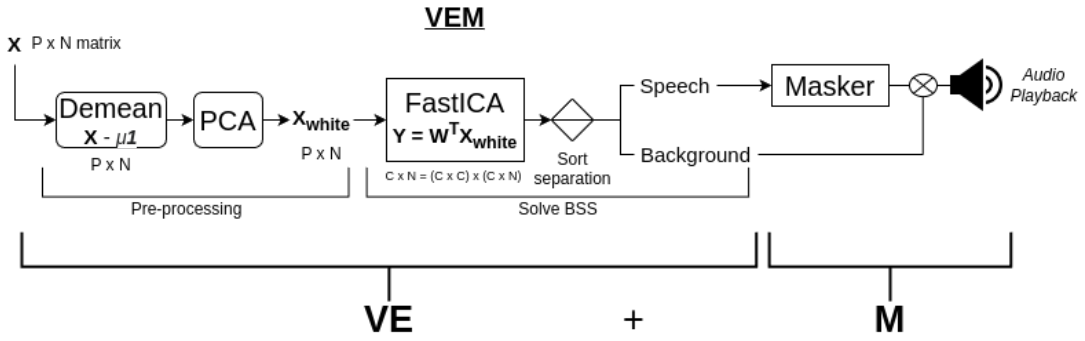


Figure 4: Flow chart of VEM. Observations from the mixture undergoes pre-processing before being separated as background and speech signal estimates. Here $\boldsymbol{\mu 1}$ denotes the row-wise mean, and the number of microphones equal the number of components to extract, $P = C = 2$. Speech signal is passed to the masker, generating a masking sound which is multiplexed back with background audio for playback.

The masker is here defined to have two main features. One is to anonymise speech online by utilising three *pitch shifts*. A pitch shift scales the pitch up or down by altering the short-time frequency contents. A simple method relies on frequency translation performed in the frequency plane. Consider a real-valued continuos $x(t)$ which is to be shifted in frequency with scalar value w_0. Then it is well-known that the connection to its Fourier transform $\mathcal{F}\{x(t)\}(w) = X(w)$ is [31]

$$e^{iw_0 t}x(t) \overset{\mathcal{F}}{\longleftrightarrow} X(w - w_0).$$

Here the angular frequency w is used. The desired factor w_0 can be represented by a factor c in *cents* where 1200 cents comprise the octave in the twelve tone equal temperament scale in music theory, meaning $w_0 = \frac{c}{1200}$, or, one can just use the ratio within or outside the octave of the implicit fundamental frequency F_0 in the signal, e.g. where the pitch shifting by η octaves is equivalent to $w_0 = 2^\eta$.[32] Using several unique pitch shifts (also restricting $w_i \neq 0$) obeys linearity of the Fourier transform as

$$\sum_{j=0}^{N-1} e^{iw_j t}x(t) \overset{\mathcal{F}}{\longleftrightarrow} \sum_{j=0}^{N-1} X(w - w_j) \tag{17}$$

where $N - 1$ is the total number of pitch shifts. If $N = 1$ then reversing the pitch shift effect is done by multiplying the back shift $e^{-iw_0 t}$,

$$e^{i(-w_0)t}e^{iw_0 t}x(t) = e^0 x(t) = x(t).$$

For the discrete case one has a real valued audio signal sampled at some sampling frequency f_s, yielding a vector $\boldsymbol{x} = \{x_n\}_0^{N-1}$, say N is even. The DFT becomes for the frequency shift of m frequency bins

$$\mathcal{F}\big(\{x_n\}\big)_k = X_k = \sum_{n=-N/2}^{N/2-1} x_n e^{i2\pi \frac{k}{N} n} \tag{18}$$

$$\mathcal{F}\big(\{x_n e^{\frac{i2\pi}{N} nm}\}\big)_k = X_{k-m}, \tag{19}$$

where the subscript $k - m$ should be understood as modulo N, the vector's length, and the $N/2 - 1$ bin is the Nyquist frequency $f_s/2$. One finds an m that corresponds to desired frequency bin shift using that $\frac{f_s}{N}$ is the bin width.

Indeed, pitch shift factors close to the original pitch might perceptually expose identity of the speaker, not needing to solve some back-shifting of frequencies. It is assumed, one can fine-tune w_0, w_1, w_2 to obtain speaker anonymisation when BSS is successful. In the case of faulty BSS the separated speech signal, contaminated with background noise, sound will still be present and one can then has to anonymise speaker identity with $w_0 = 0$, w_1, w_2, w_3 when multiplexing back.

The other masker aims to obfuscate the content by removing the defining cues in speech. The *time-reversing masker* locally reverses audio frames as $x(t) \to x(-t)$, for $t \in [t_0, t_1]$, where the claim that similarity of frequency contents is expected, which was made in Section 1.4.2, is here confirmed since scaling with -1 of a real-valued signal x yields

$$x(-t) \overset{\mathcal{F}}{\longleftrightarrow} X(-w) = X(w)^*$$

where * denotes the complex conjugate. [31] Thus, both temporal and tonal masking are present when reversed frame segments are placed (somewhat) time-aligned with the original audio segment, where imperfect time-alignment leads to forward and backward temporal masking. The drawback of this method is the necessity for several seconds for an efficient masker, since reversing stationary sounds like vowels with typical duration being less than a second. For example reversing a sinusoid with constant amplitude in the reference window alters not the resulting heard longitudinal wave. It is here noted that the psychoacoustical phenomena of auditory masking, mentioned in Section 1.2, is documented to occur when consecutive speech frames correlate with the increasing independence after circa 100ms, mentioned in Section 2.2.2.

Another simplistic method of removing cues in speech in real-time is by altering the phase of the corresponding spectrum content. Consider the frequency domain of a

signal. The component of the signal at any given frequency is given by a complex number, which has a real and imaginary part as $a + bi$, where a and b are real-valued. These parts expressed in polar coordinates becomes $r\cos\phi$ and $r\sin\phi$ respectively, where ϕ is the phase of each sinusoid. The real part corresponds to the magnitude of that frequency and the phase corresponds to where in time the frequency's peaks are located. Thus taking the inverse Fourier transform of the real part of a continuous speech signal's Fourier transform

$$\int_{-\infty}^{\infty} \Re\left\{ \int_{-\infty}^{\infty} f(t)e^{-i2\pi\xi t}dt \right\}(\xi)e^{i2\pi\xi t}d\xi$$

is analogous to assuming all phases ϕ of the sinusoids in the frequency domain are zero when returning to the time domain. The term *phase-less signal* is here used to describe this procedure. When the setting is discrete, magnitude and phase of frequency bins represent the discrete signal's short-time Fourier transform (STFT), as briefly mentioned above. This effect is proposed as a speech masking technique which hinders access to cues in speech. The window length determines how much information gets displaced in time in overlapping windows, where displacement of content in one frame is confined to the window itself. An offline approach to obtaining a phase-less signal, where the window length can vary in duration, can achieve more intricate masking sounds, increasing computational load for devices, although risking to surrender its autonomy to function on any speech signal of various length. The sought speech masker should therefore be computationally lightweight to have any usefulness in real-world surveillance.

2.5 Dynamic VEM and Voice Activity Detection

In a real-world surveillance implementation, deciding whether to activate the separator when speech activity is present becomes a necessity. There is no need to separate the background sounds from themselves which would happen since FastICA itself is a blind separator. Therefore one would have to trigger the separator when speech is detected and deactivate it after some duration of consecutive non-speech duration. In an offline setting, this binary classification could yield timestamps in the recording for which segments that should be masked or separated, and in the online setting, one would have to start and stop masking or separation in real-time, which themselves is a considerably harder problem, not performed her. This is in part due to a conflict of demanding minimal latency when streaming versus successful separation and masking, stated below. Consider latency from a few milliseconds to a few seconds, then from previous sections up until this point, one should expect:

- Increasing latency → weakening source dependence and gaussianity → increasing FastICA success rate
- Increasing latency → Speech-like maskers have better tonal-temporal masking effect → Increasing speech masking performance.

Detection of voice activity (*Voice Activity Detection* or *Speech Activity detection*) has been performed either through pure signal processing or via deep learning, in [33], [34]. The method usually consists of feature extraction from signal spectrum, decision-making module followed by decision-smoothing. Methods are based, among other things, on energy thresholds, pitch detection, spectrum analysis, zero-crossing rate, and periodicity. Convolutional Neural Networks (CNN) have also been used as a model for VAD, and utilised some of the methods from the aforementioned features above as input to the network. In a surveillance setting, it is most likely that VAD is integrated earlier in the pipeline as a feature in audio classification, necessitating robust detection in noisy and vibrant audio settings.

3 Methodology

The FastICA method was performed and evaluated sequentially on each data set. For every $N \times P$ test sample in each data set, reference clean speech and background sound were used in the corresponding aforementioned performance metrics. The number of microphones were $P = 2$ for the first two data sets, and $P = 4$ for the third (see Section 3.1.2 below). The permutation problem was handled by first finding separated channel with maximum ESTOI to reference, which would imply the other channel was the candidate to measure background intactness with. Thus, each test sample generated a separation with two metrics that were documented. The mean and standard deviation was calculated for both metrics. The results were also listened back to, in order to detect possible features that the metrics could not convey. In addition, to obtain a reference for the bounds of STOI/ESTOI, these measures were computed for clean speech and and each background in Data Set 1. The process can be seen in Figure 5.

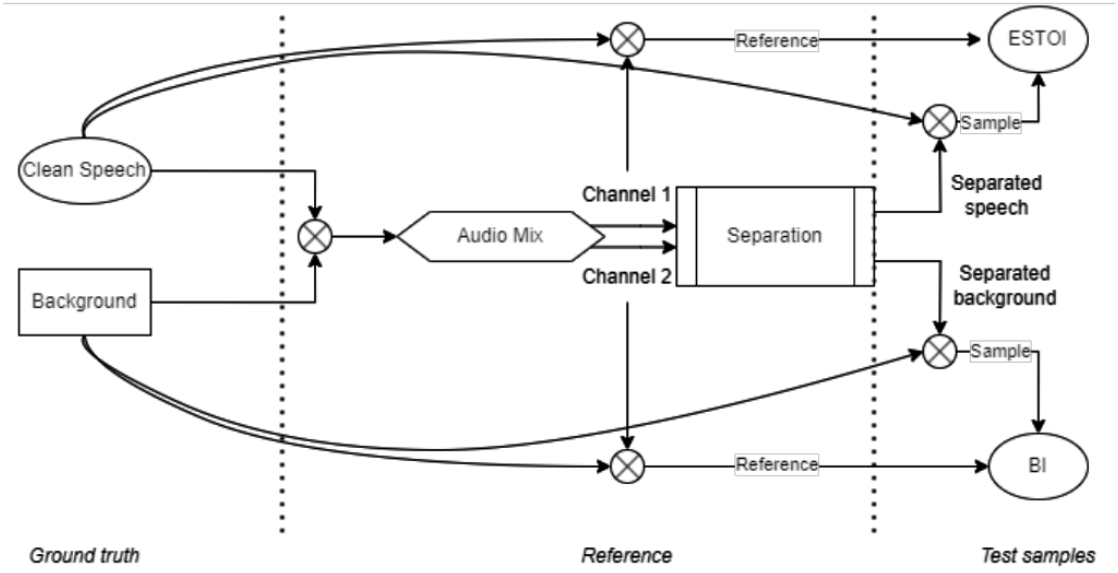


Figure 5: Methodology of Creating and Evaluating Data Set 1-3

Speech masking and speaker anonymisation was performed on the reference clean speech in Data Set 1. The audio were subjected to two types of maskers, one of which masked speech content, one of which masking identity. These were

- Three Pitch-Shift threads (speaker anonymization)
- Time-reversal of Phase-less STDFT (speech masking), plus reverberation.

The speaker anonymisation was done through introducing three separate pitch shifts on three copies of the audio data. The three frequency shifts were tuned until deemed sufficient, having one lower shift, and two higher, resulting in angular frequency shifts $w_0 = \frac{3}{5}\ \frac{rad.}{s}$, $w_1 = \frac{17}{10}\ \frac{rad.}{s}$, and $w_2 = \frac{11}{5}\ \frac{rad.}{s}$. The process was performed offline using an overlap-add method; by windowing frames of 32ms at sample rate 48kHz with 50% overlap, using a Hann window. The process was also performed in real-time, with the difference being that data was read from a ring buffer and put in an output queue after pitch shifting. The output queue played back resulting audio. This was to show the possibility of a real-time implementation.

The second masker sound was generated as follows. First an offline phase-less manipulation on a speech signal was performed to demonstrate its general effect. Subsequently, the second masker was implemented in a stream imitation setting, where manipulated

time frames were multiplexed back to a masking sound of equal duration as original speech signals. This was performed by reading in frames without overlap, computing the STDFT, setting the complex part to zero and computing the inverse STDFT. Then frames were reversed. Also, repeating the process with 400ms of reverberation effect on the masking sound was implemented to see if even a short reverberation effect affected intelligibility. The time frames were chosen to be 500ms, 1000ms and 2000ms in duration. To compare the speech intelligibility due to masking, STOI and ESTOI was computed on the resulting masking sounds. The resulting masker sounds was then listened back to, to see if this new context of using the provided OIMs were meaningful.

3.1 FastICA extracting several components

FastICA can be described as an algorithm seeking an orthogonal rotation of pre-whitened data, through a fixed-point iteration scheme, maximising negentropy of rotated components. A fixed-point iteration is a method that computes an approximate fixed point x^* of a function. For example a real-valued function f the fixed point x^* has the property $x^* = f(x^*)$. For a sequence $x_0, x_1, \ldots x_n, x_{n+1}, \ldots$

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} : \forall n > N, \quad |x_{n+1} - f(x_n)| < \epsilon,$$

where further mapping of x_{n+1} through f approximates x_n up to a tolerance limit, for example $\epsilon = 10^{-4}$. FastICA for MIMO-systems perform fixed-point iterations according to the number of components to extract.

Pre-processing

The pre-processing steps are to subtract the sample mean $\boldsymbol{m}$ (centering) and pre-whiten using PCA. Centering yields zero mean of the observations $\boldsymbol{x}$

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \boldsymbol{m}$$

and pre-whitening can be done using PCA to obtain a covariance matrix $\boldsymbol{C}$ based on (10) where we seek to obtain new column vectors $\tilde{\boldsymbol{x}}$

$$\boldsymbol{C} = \mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T] = \boldsymbol{I}.$$

Pre-whitened data become after PCA

$$\tilde{\boldsymbol{x}} = \boldsymbol{E}\boldsymbol{D}^{-1/2}\boldsymbol{E}^T\boldsymbol{x} = \boldsymbol{E}\boldsymbol{D}^{-1/2}\boldsymbol{E}^T\boldsymbol{A}\boldsymbol{s} = \tilde{\boldsymbol{A}}\boldsymbol{s},$$

where $\boldsymbol{D} = diag(d_1^{-1/2}, d_2^{-1/2}, \ldots, d_n^{-1/2})$, $\boldsymbol{E}$ holds the eigenvectors, and a new sought mixing matrix $\tilde{\boldsymbol{A}}$ is obtained, which is orthogonal. [25]

Algorithm

The iteration scheme updates weight vectors $\boldsymbol{w}_i$ to find respective directions in observation space such that each projection $\boldsymbol{w}_i^T \tilde{\boldsymbol{x}}_i$ maximises nongaussianity $\boldsymbol{J}_G$. Since data is pre-whitened, one has to restrict the norm of all $\boldsymbol{w}$ to be unity. The nongaussianity is measured with the approximation of negentropy, Corollary 2.2.1. In addition the first and second derivative, g and g', of G are used in updating. This is since negentropy maxima of $\boldsymbol{J}_G$ coincide with the maxima of its term $\mathbb{E}[G(\boldsymbol{w}^T\boldsymbol{x})]$, stated in [25]. The update rule for one $\boldsymbol{w}$ becomes eventually

$$\begin{aligned} \boldsymbol{w}^+ &= \mathbb{E}\left[\tilde{\boldsymbol{x}} g(\boldsymbol{w}^T \tilde{\boldsymbol{x}})\right] - \mathbb{E}\left[g'(\boldsymbol{w}^T \tilde{\boldsymbol{x}})\boldsymbol{w}\right] \\ \boldsymbol{w}^+ &\leftarrow \frac{\boldsymbol{w}^+}{||\boldsymbol{w}^+||} \end{aligned}$$

for the single unit extraction. Now since, we aim to separate several components, one must prevent different $\boldsymbol{w}$ from converging to the same maxima. The method of *decorrelation* achieves this, where different projections $\boldsymbol{w}^T\tilde{\boldsymbol{x}}$ find different optimal directions in observation space. The method used for decorrelating the weights is done here by gathering weight vectors in a matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$ and computing

$$\boldsymbol{W} \leftarrow (\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^{\boldsymbol{T}})^{-1/2}\boldsymbol{W}$$

after each update. The inverse square root above is thus equal to the eigenvalue decomposition, given by PCA at each iteration. That is $(\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}^{\boldsymbol{T}})^{-1/2} = \boldsymbol{E}\boldsymbol{D}^{-1/2}\boldsymbol{E}^{\boldsymbol{T}}$. Note that covariance matrix $\boldsymbol{C}$ for whitened data is just the identity matrix and can be omitted. The extra requirement from the weights having unity norm is now that they must be orthogonal and thus $\boldsymbol{W}$ must be unitary. Now the main algorithm can be presented in Algorithm (1), with $G(x) = \log\cosh x$ and $\mathbb{E}[X] = \frac{1}{N}\sum_{x_i \in X} x_i$, in accordance with [26].

Algorithm 1 FastICA for Several Components Extraction

1: **Input:** $N \times P$ pre-whitened data matrix $\tilde{\boldsymbol{X}}$
2: **Input:** Desired number of independent components $M \leq P$
3: **Output:** $M \times M$ matrix $\boldsymbol{W}$ of unmixing matrix estimate
4: **Initialisation:** Random initialisation of the unmixing matrix $\boldsymbol{W}$
5: **Repeat until convergence:**
6: Compute projections $\boldsymbol{W}^T\tilde{\boldsymbol{X}}$
7: Compute $g(\boldsymbol{W}^T\tilde{\boldsymbol{X}}) = \tanh(\boldsymbol{W}^T\tilde{\boldsymbol{X}})$ and $g'(\boldsymbol{W}^T\tilde{\boldsymbol{X}}) = 1 - \tanh^2(\boldsymbol{W}^T\tilde{\boldsymbol{X}})$
8: Compute new estimate $\boldsymbol{W}^+$

$$\boldsymbol{W}^+ = \mathbb{E}\big[\tilde{\boldsymbol{X}} g(\boldsymbol{W}^T\tilde{\boldsymbol{X}})\big] - \mathbb{E}\big[g'(\boldsymbol{W}^T\tilde{\boldsymbol{X}})\boldsymbol{W}\big]$$
$$\boldsymbol{W}^+ \leftarrow \frac{\boldsymbol{W}^+}{||\boldsymbol{W}^+||}$$

9: Decorrelate $\boldsymbol{W}^+$ with respect to previously unmixing matrix estimates $\boldsymbol{W}$

$$\boldsymbol{E}, \boldsymbol{D} \leftarrow PCA(\boldsymbol{W})$$
$$\boldsymbol{W}^+ \leftarrow \boldsymbol{E}\boldsymbol{D}^{-1/2}\boldsymbol{E}^{\boldsymbol{T}}$$

10: Update unmixing matrix: $\boldsymbol{W} \leftarrow \boldsymbol{W}^+$
11: **End**

Post-processing

When $\boldsymbol{W}$ is obtained, one has a solution to $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}}\boldsymbol{s}$, one still has to transform back to the original observation space, by dewhitening and adding back the mean. The estimated sources $\boldsymbol{Y}$ gets retrieved as

$$\tilde{\boldsymbol{Y}} = \boldsymbol{W}\tilde{\boldsymbol{X}}$$
$$\boldsymbol{Y} = \boldsymbol{W}\big(\boldsymbol{E}^T\boldsymbol{D}^{1/2}\boldsymbol{E}\boldsymbol{x} + \boldsymbol{m}\big).$$

3.2 Tools and Resources

3.2.1 Software

Almost all code was written in Python using popular packages NumPy, SciPy, Matplotlib for computational efficiency. The PCA and FastICA algorithm used were from the Scikit-learn package, which has several machine learning tools in Python. The STOI measure were computed using the package PySTOI, the normalised cross-correlation measure between background sounds with SciPy signal's *correlate*-function.

The TASCAR *scenes* (Section 3.2.2 Data Set 2), [35] the designed virtual acoustic environments, were of course written in TASCAR, meaning that HTML-files were written to configure sound sources and their movement and then a Python class that would run TASCAR commands in a command window to render said scenes, in order to create the last two data sets.

Aforementioned speech maskers were first implemented in Python, but wanting to stream audio in real-time with low latency, speaker identity anonymisation was implemented as Linux Audio Developer's Simple Plugin API (LADSPA) plugins that fed into the low-level multimedia framework PipeWire's *filter-chain* module. These plugins were configured in UNIX configuration files.

3.2.2 Data Sets

Data set 1**;** 6156 **seconds of Synthetic Mixtures**

Clean speech was recorded by two supervisors at Axis Communications AB and comprised two audio files of similar sentences phrased through various emotions. Reverberations were not audible when subjectively examined. Each wav-recording totalled roughly 5 minutes and 30 seconds and was sampled at 48000 Hz. The audio files were cut to shorter sentences with some natural pauses in speech, mostly around eight to nine seconds. The total amount of clips was 83, 41 of which female, 42 of which male. Figure 6 displays the histogram of the clean speech recordings, alongside the histograms of the male and female speaker.

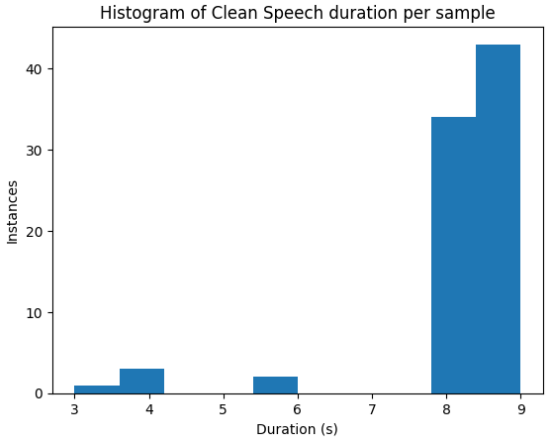
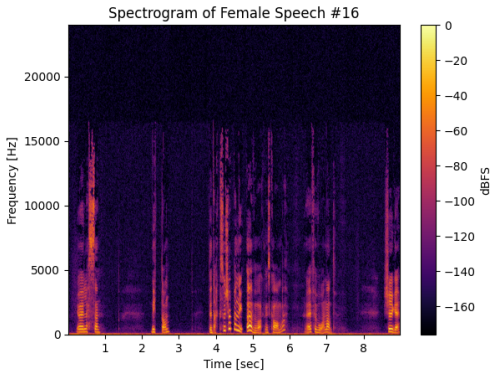
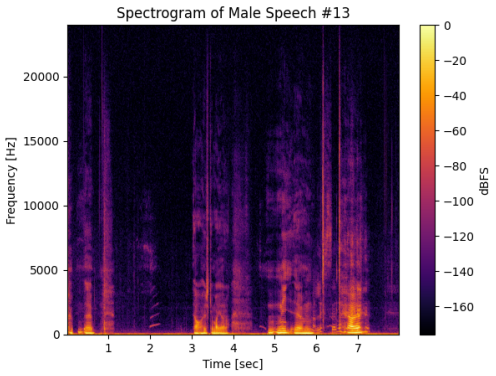


Figure 6: Histogram of clean speech recordings's duration in seconds.

The spectrograms of two clean speech samples can be shown in Figure 7, where Figure 7a displays a sample from the female supervisor. Figure 7b displays a sample from the male supervisor.



(a) Spectrogram of 16th female sample. Frequencies over 16kHz seems to have been removed from the signal.



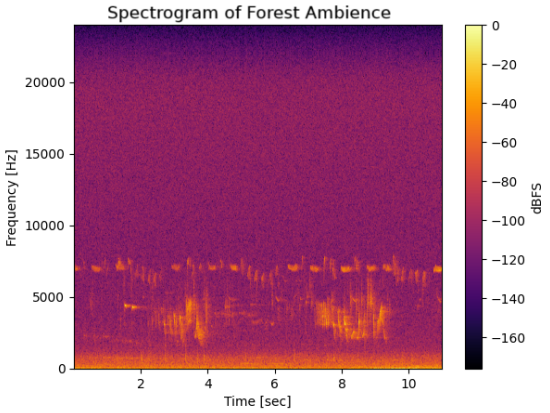
(b) Spectrogram of 13th male sample. In comparison to 7a, the audio appears not to have been notably altered.

Figure 7: Spectrograms of two speech samples. The absence of noise if visible through low intensities inter speech, in decibel relative to full scale (dBFS).

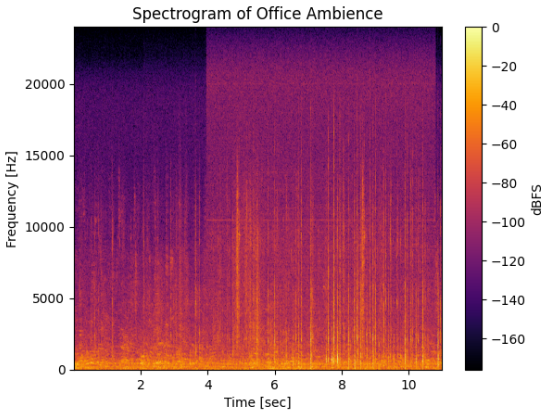
Three different background sounds was obtained from [36]. They were resampled to the same sampling frequency as clean speech. The background sounds were

- European spring forest ambience
- Office ambience
- Urban city sounds and light car traffic

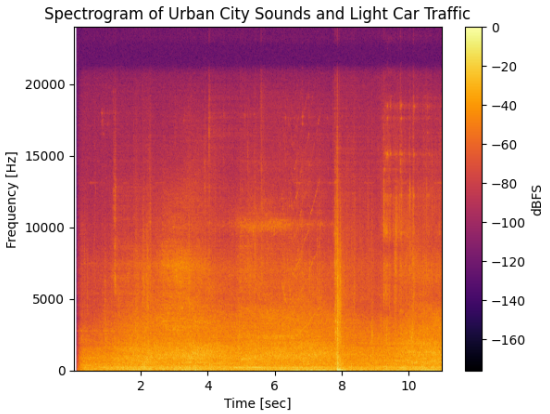
and were chosen since they were deemed realistic background sounds for surveillance technology. The spectrograms for the background sounds can be seen in Figure 8a, 8b, 8c.



(a) Several birds chirping alongside a mild breeze can be heard.



(b) Several instances of typing on keyboards can be heard.



(c) Urban ambience sounds can be heard. A car door is slammed shut at roughly eight seconds.

Figure 8: Background sounds' spectrograms for the first eleven seconds in decibels relative to full scale (dBFS).

The bird chirps in Figure 8a display complex rhythmic pattern of rapid rising and falling pitch with pauses, and although the visible dominant frequencies do not fully overlap with the vocal frequency range, this background sound's spectrogram is most similar to the typical human speech spectrograms displayed in Figure 7a and 7b.

Each clean speech sample were mixed in with each background sound on two channels, meaning the background sounds were the same duration as every individual speech audio. Three SNR levels of 0, -6 and -12dB were created for each speech-background sample pair, in order to measure the effectiveness of speech separation as speech were masked by increasingly louder background noise. The clean speech was emulated to originate from 10 metres away to the centre between two omnidirectional microphones with spacing 0.1 metres apart, at a fixed DOA of roughly $-59°$. This was to simulate a decrease in intensity and a inter-microphone time delay between the two microphones. The background sounds were emulated to be *omnipresent*, *diffuse* sources; even though each background sound had two channels, only one of them was chosen to be mixed in each channel. This was due to the unknown nature of the recordings, meaning different microphones most certainly had been used and the audio contained different reflection patterns and room impulse responses, and the distance between microphones for these sounds were unknown. To avoid several types of errors present in this data set model, the crude data set was deemed to be a suitable beginning of measuring the possibility to separate speech.

Data Set 2; TASCAR's Acoustic Scene Renderings

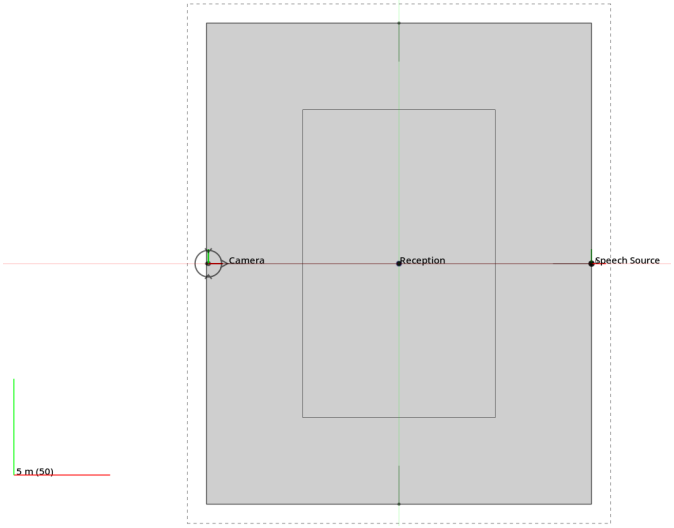


Figure 9: Overview of the scene in TASCAR from above. Red and green axis lines correspond to directions in the xy-plane according to the Right-Hand-Rule convention. The rectangle inside the room is the path the moving speech source traces

The second data set were rendered in Toolbox for Acoustic Scene Creation And Rendering (TASCAR), [35] and it comprised the same clean-speech recordings but with new background sounds, namely

- Train Station
- Crowd Talking Murmur
- Outdoor wind and bird song

provided by supervisors at Axis Communications AB. The setup can be seen in Figure 9 and Figure 10. The room was a shoe-box with dimensions $20 \times 25 \times 4\ m^3$ with origin in the midpoint of the room, with the camera placed at coordinates $\begin{pmatrix} -9.9 & 0 & 2.7 \end{pmatrix}$. The clean speech test file is a source that moves around in 10 seconds, where it approaches the camera, backs up a bit and traces a rectangle before returning to the initial position across the room from the camera. The idea of quick movement was to examine how FastICA handled rapid changes in DOA. The data set was generated by first rendering, in TASCAR, the clean speech and background sound as reference for evaluation and then rendering the mix. The camera had two microphones and was modelled as

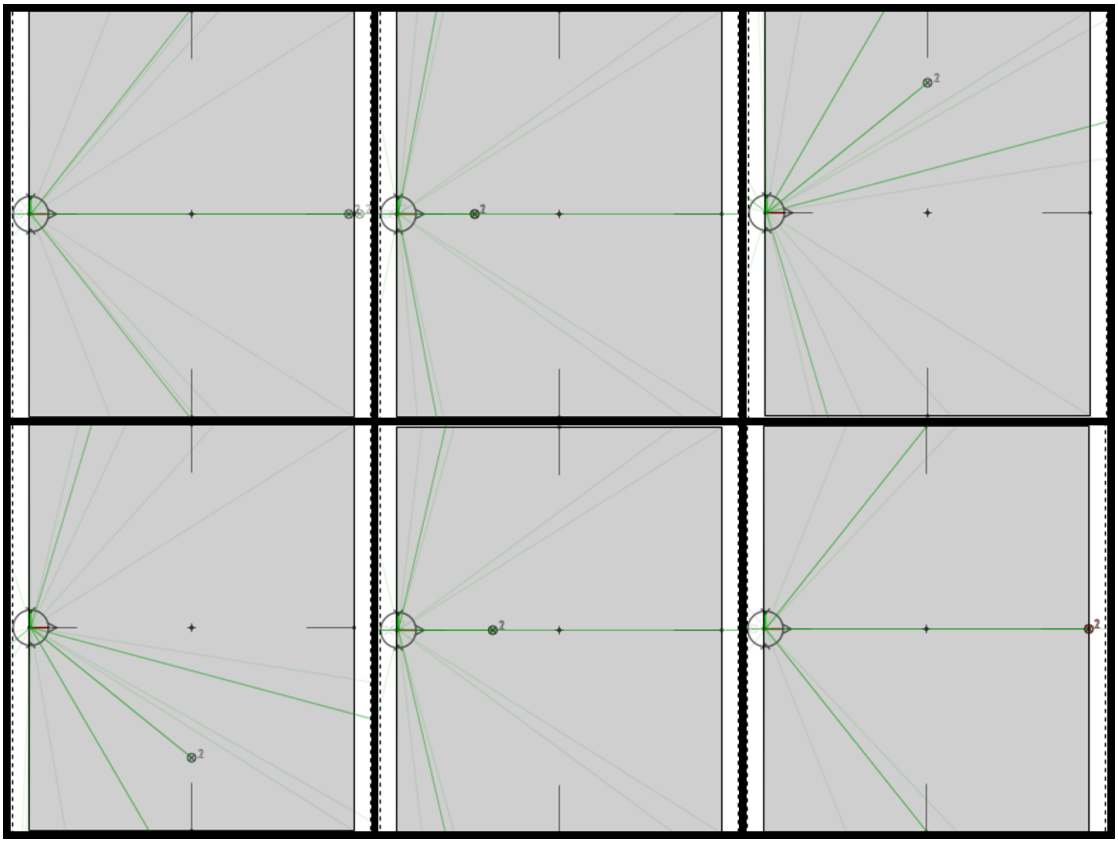
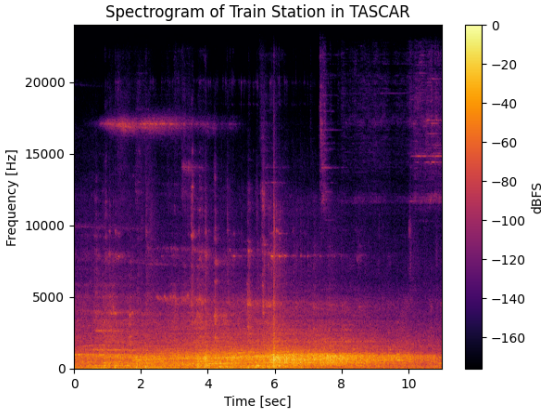
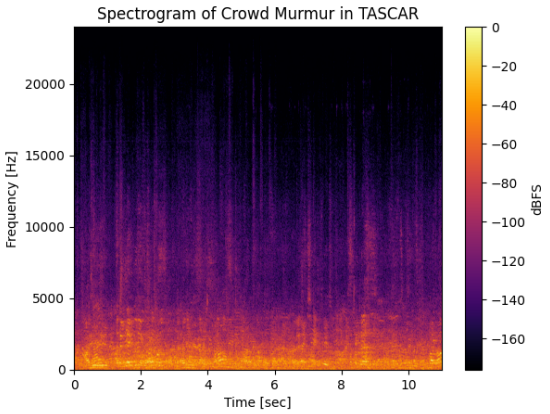


Figure 10: The scene in TASCAR for Data set 2, seen from z-axis. The dot labelled '2' is the moving speech source, the head shape that is positioned centre left is the camera. The background sounds are played equally loud and reverberant throughout the room. The speech source moves centre-right to centre-left, backs up, then travels clock-wise in a rectangle shape back to being in front of camera, then backs up to its starting position. The green lines indicate its sound paths.

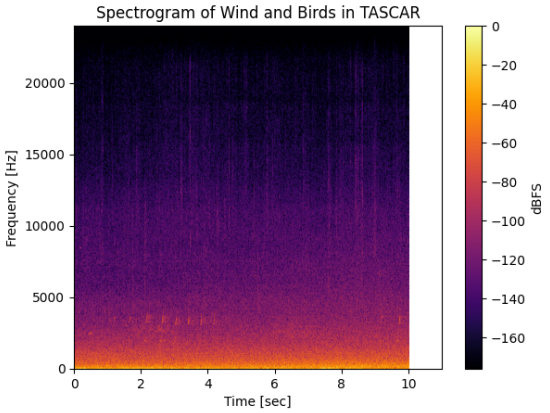
Brown and Duda's Spherical Head Model (1998), with 8cm radius, with the associated Head Related Transfer Function (HRTF), to emulate how sound waves would reach the microphones on a physical camera, similar to human ears. The SNR was set at +20dB internally in TASCAR (60dB and 40dB), since speech was not audible at equal loudness parameters. TASCAR's SNR parameter was thus approximately set to 0dB and no further SNR changes was performed. The spectrogram of the new background sources can be seen below in Figure 11.



(a) Several train wheels on a moving train are heard grinding and lots of reverberation.



(b) Several people are talking simultaneously from different positions, and sometimes one clear voice can be heard in the foreground.

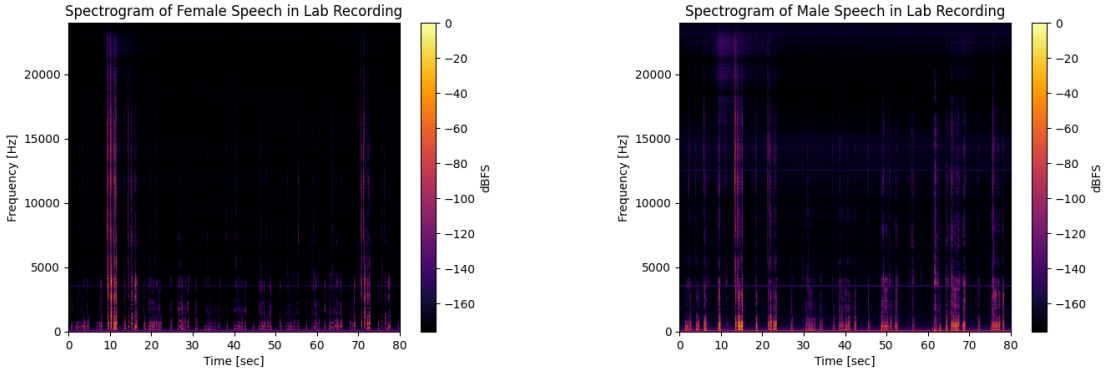


(c) A wind breeze is present and birds are heard chirping, possibly the sound of rain too.

Figure 11: Second Data Set: Background sounds' spectrograms for the first eleven seconds in decibels relative to full scale (dBFS).

Data set 3 - Camera Recordings in Lab

Returning to the original clean speech files of the female and male speaker, one chunk of 80 seconds of speech was chosen from each speaker to be the new prolonged clean speech recordings. Figure 12 displays the spectrograms. Both speakers asked a question in a loud irritated voice roughly after ten seconds of recording, often resulting in inaudible background sound under the same duration. SNR was set at +10dB internally in TASCAR (60dB and 50dB), since speech was not audible at equal loudness parameters. TASCAR's SNR parameter was thus approximately set to 0dB and no further SNR changes was performed.



(a) Spectrogram of female speaker

(b) Spectrogram of male speaker

Figure 12: Lab recording of moving clean speech, the loud peaks at 10 and 70 seconds can be explained by that the speech source was close to the camera.

The background sounds themselves were the same as the former data set but in addition

- Park sound
- Traffic sound

were included. The park sound included a leaf blower and the traffic sound consisted of cars swooshing past the camera every few seconds. The lab setup can be seen in Figure 13, were the TASCAR animation was emulated through playback through five loudspeakers around and a sub-woofer under the camera. The camera had two pairs of microphones and the pairs were separated by 0.1m. The animation was analogous to the former data set's animation, the difference being that the speech source traced the same path in a rather slow walking speed. Again references were recorded by the microphones followed by a mix-recording and was saved. The sound level at the

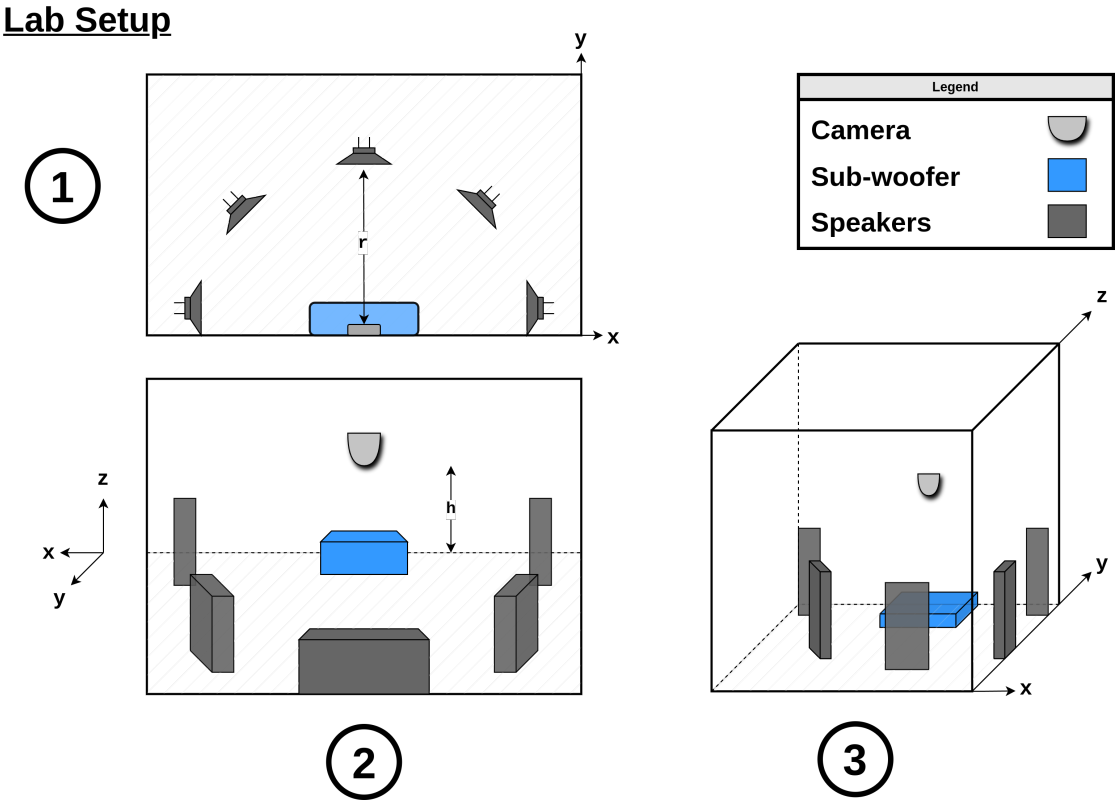


Figure 13: Sketch of lab setup. ①: Upper figure displays view from above. ②: Lower figure displays the view standing behind the speakers and facing the camera. ③: Three-dimensional view of the setup. The five loudspeakers were placed in half-circle with radius $r = 2$m and the camera with 4 microphones were placed roughly $h = 1m$. Below a sub-woofer was placed to emulate the low-frequency components in the mixture.

camera was documented with a sound level meter for both clean speech and mixed audio before recording. The procedure of recording was to start and stop recording before and after the playback of the TASCAR scenes, then manually cutting all audio to obtain synchronised test and reference results manually by choosing correct indices of reference and mix recordings such that the waveforms of the first channel aligned. Alignment was performed by plotting references and mixes in increasing zoom until a window of 100 samples were deemed to be aligned. Lastly the clipped recordings were written as new files using software.

4 Results

4.1 FastICA's results on Data Sets

FastICA always converged with tolerance 10^{-4} indicating number of iterations was within 200 max iterations. The Extended STOI of separated speech signals can be seen along with their reference (ESTOI between ground truth clean speech and the mixed unseparated left channel audio) in Figure 14 for Data Set 1. The corresponding background intactness can be seen in Figure 15. The performance metrics can be seen in Figure 16 for Data Set 2. Separation for Data set 3 was unsuccessful when performing separation on the full length audio, resulting in the same ESTOI scoring as the references in Data Set 2. The difference between TASCAR sound setting and the recorded sound level in the lab (Data set 3) was documented and can be seen in Table (1). When computing STOI and ESTOI between the three background sounds in Data Set 1 and male and female speech recordings, the STOI was in $0.28 - 0.34$ and ESTOI in $-0.01 - 0.01$, displaying lower values than all reference ESTOI mean, yet that ESTOI-scoring with a sound not containing speech did not approach the theoretical lower bound of -1.

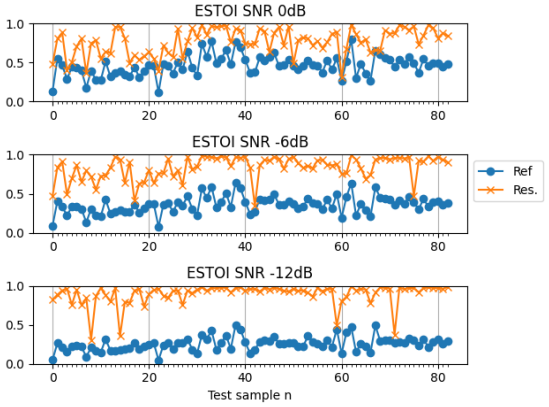
Audio type	Recorded dBSPL	TASCAR dBSPL
*Clean speech Male	43-80	60
*Clean speech female	45-80	60
Background Crowd Murmur	58-63	50
Background Train station	52-65	50
Background Windy Birds	50-54	50
Background Traffic	50-70	50

Table 1: Data set 3: Recorded decibel relative to sound pressure level at the camera throughout the recording and TASCAR programme's internal setting. *The measured sound level is only for voiced segments of the audio.

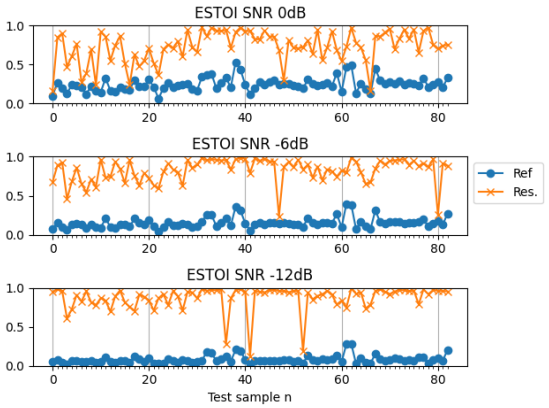
Background ESTOI	**SNR** 0**dB** mean	std	**SNR** -6**dB** mean	std	**SNR** -12**dB** mean	std
Forest result	0.77	0.17	0.84	0.15	0.90	0.13
------- reference	0.46	0.13	0.36	0.11	0.26	0.09
Office result	0.73	0.20	0.83	0.15	0.87	0.16
------- reference	0.25	0.09	0.15	0.07	0.16	0.05
Urban result	0.75	0.18	0.83	0.13	0.90	0.08
------- reference	0.26	0.09	0.16	0.07	0.09	0.05
Background BI	**SNR** 0**dB** mean	std	**SNR** -6**dB** mean	std	**SNR** -12**dB** mean	std
Forest result	0.61	0.08	0.85	0.06	0.94	0.06
------- reference	0.57	0.04	0.81	0.03	0.94	0.01
Office result	0.75	0.06	0.92	0.04	0.97	0.05
------- reference	0.71	$<$0.01	0.90	$<$0.01	0.97	$<$0.01
Urban result	0.75	0.06	0.92	0.03	0.98	$<$0.01
------- reference	0.71	$<$0.01	0.90	$<$0.01	0.97	$<$0.01

Table 2: Data set 1: Mean and standard deviation of ESTOI and BI for different Signal-to-Noise ratios.

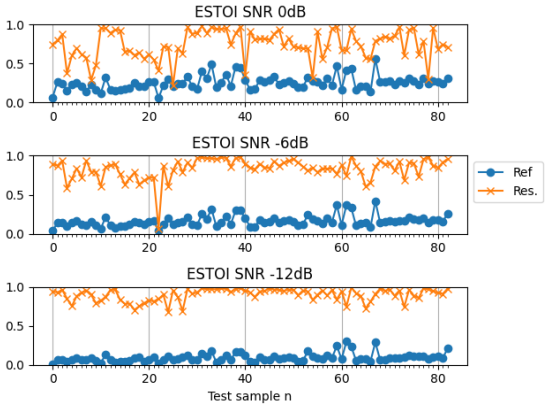
The corresponding mean and variance can be seen in Table (2) for first data set. When listening back to separations of Data Set 1, the separation was deemed near perfect; all separated speech was audible and only at -12dB SNR did some static white noise enter the speech-separated audio. A slight tendency of higher ESTOI scores at negative SNR in decibels is noted in Figure 14. Background separation was always clear when listened back to and as Figure 15 indicates, louder background sounds displayed higher BI score. The lower scores in BI still only contain the background audio events but with some audible difference in spectrum quality. Table (2) indicate similar performance for all background sounds. When listening back to the separations of Data Set 2 the low performance scores could not convey the fact that for all separations, one channel contained only background sound while the other channel contained both speech and background sounds. This indicates successful background separation of Data Set 2. However, some background separations suffers from altered spectrum quality. This is displayed in corresponding scores of BI in Figure 16. When listening back to Data Set 3 test samples, reference and separations, separation of the considerably longer audio files were always unsuccessful. When the moving speech source was further away from the camera, it was barely audible, in contrast to the loud raised voice asking a question standing in front of the camera. Attempts at achieving sufficient separation was performed for durations of roughly ten seconds for different locations in the scene, they too were unsuccessful. The resulting separations of both longer audio and shorter segments always had at least one Gaussian component each; static background noises.



(a) Separated speech from forest ambience.

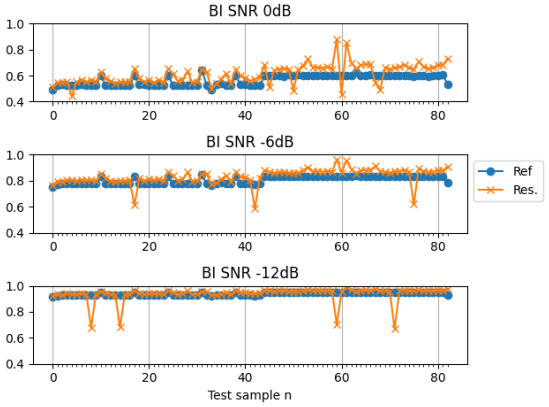


(b) Separated speech from office ambience.

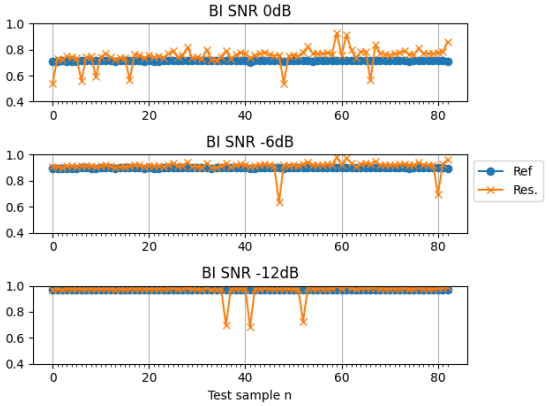


(c) Separated speech from urban city sounds.

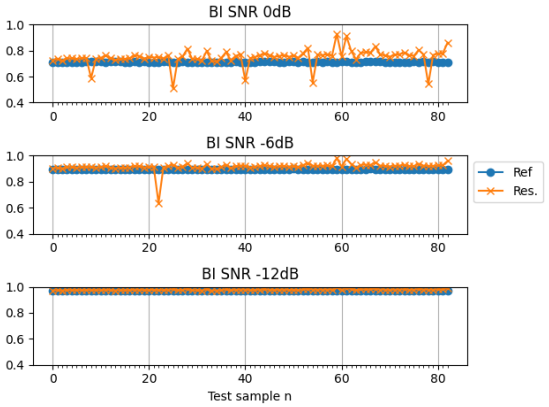
Figure 14: Extended STOI of separated speech in Data Set 1. Reference is the scoring of the input mixture to clean speech.



(a) Separated background from forest ambience.

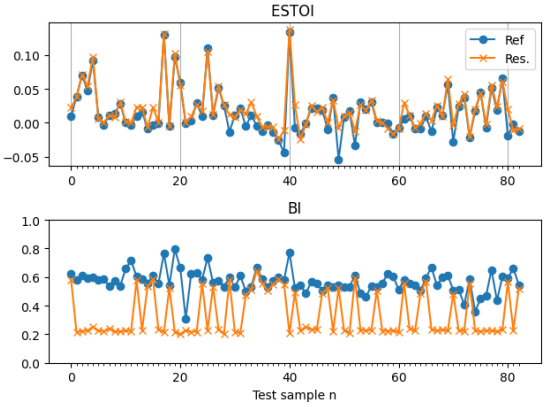


(b) Separated background from office ambience.

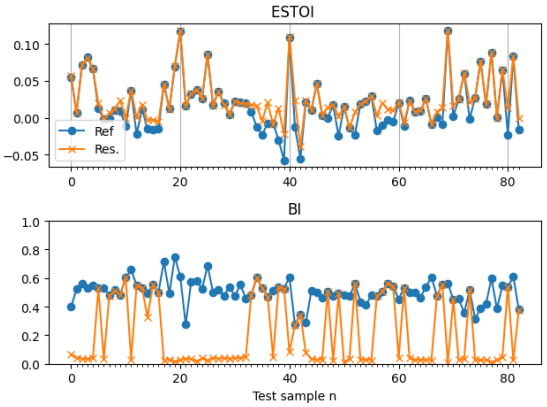


(c) Separated background from urban city sounds.

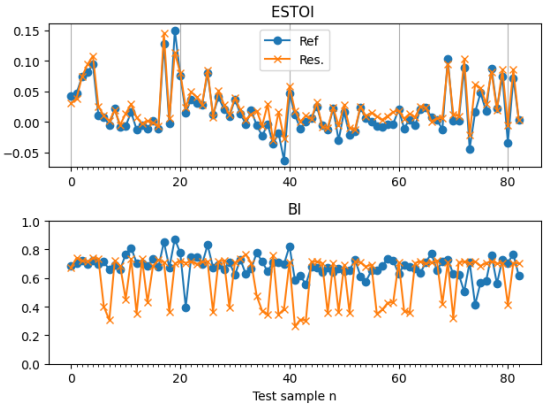
Figure 15: BI of separated speech in Data Set 1. Reference is the scoring of the input mixture to background audio.



(a) Separation from Crowd Murmur.



(b) Separation from train station ambience.



(c) Separation from outdoor wind and bird song.

Figure 16: Extended STOI of separated speech and Background Intactness in Data Set 2.

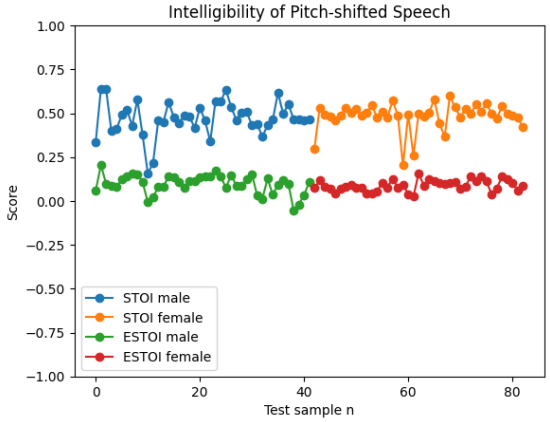


Figure 17: Speech intelligibility of Three-threaded Pitch Shifting on clean speech recordings from Data Set 1.

4.2 Results of Speaker Anonymisation and Speech Masking

The STOI and ESTOI of the voice anonymiser performed on the clean speech recordings from Data Set 1 can be seen in Figure 17. The mean and standard deviation for STOI was 0.48 and 0.09 respectively. The mean and standard deviation for ESTOI was 0.093 and 0.04 respectively. Dividing the data into male and female speaker, the corresponding values become for male STOI 0.47 and 0.10, male ESTOI 0.10 and 0.05, female STOI 0.48 and 0.08 and female ESTOI 0.09 and 0.03. When listening back to the anonymised speeches, the speech content was deemed intelligible.

The act of nullifying the phase of the spectrum content of a whole signal can be seen in Figure 18. The waveform is visibly altered, displaying lower resemblance to the original signal and seems to be symmetric around the half-time mark of its duration in seconds. The computed intelligibility of the second masker sounds can be seen in Figure 19. It shows that each masker sound performed similarly for all time frame durations and that adding reverberation significantly lowered the intelligibility score. As previously seen, ESTOI scores are never greater than its STOI counterpart. The mean and standard deviation of the OIMs for all frame durations for the masker sound without reverb was roughly 0.60 and 0.12 for STOI, the corresponding ESTOI scores were 0.46 and 0.12. The effect of short reverberation altered the corresponding scores to 0.16 and 0.08 for STOI, 0.02 and 0.03 for ESTOI. When listening back to the results of the masker sound without the reverberation, the lower frame duration corresponded to greater clipping effect, irritating discontinuities. The speech content was seldom fully masked, especially when single words were uttered. Extending the duration of frames to one and two seconds increased the masking effect, but discontinuities were

still present as the rhythm of the speech was changed. Especially at the two second frame duration, the effect of informational masking increased. When listening back to the results of adding 400ms reverberation, the masker sound's clipping sounds were mitigated, not as irritating but did not increase the masking effect significantly. To clarify, the original speech was still present after time-reversing the phase-less audio, though, it was accompanied by time-reversed fragments of the original speech that resulted in auditory masking effects.

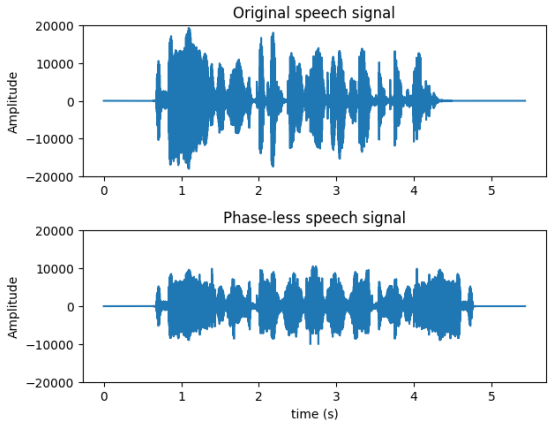


Figure 18: Result of nullifying phase of speech signal. The original sentence was five seconds long.

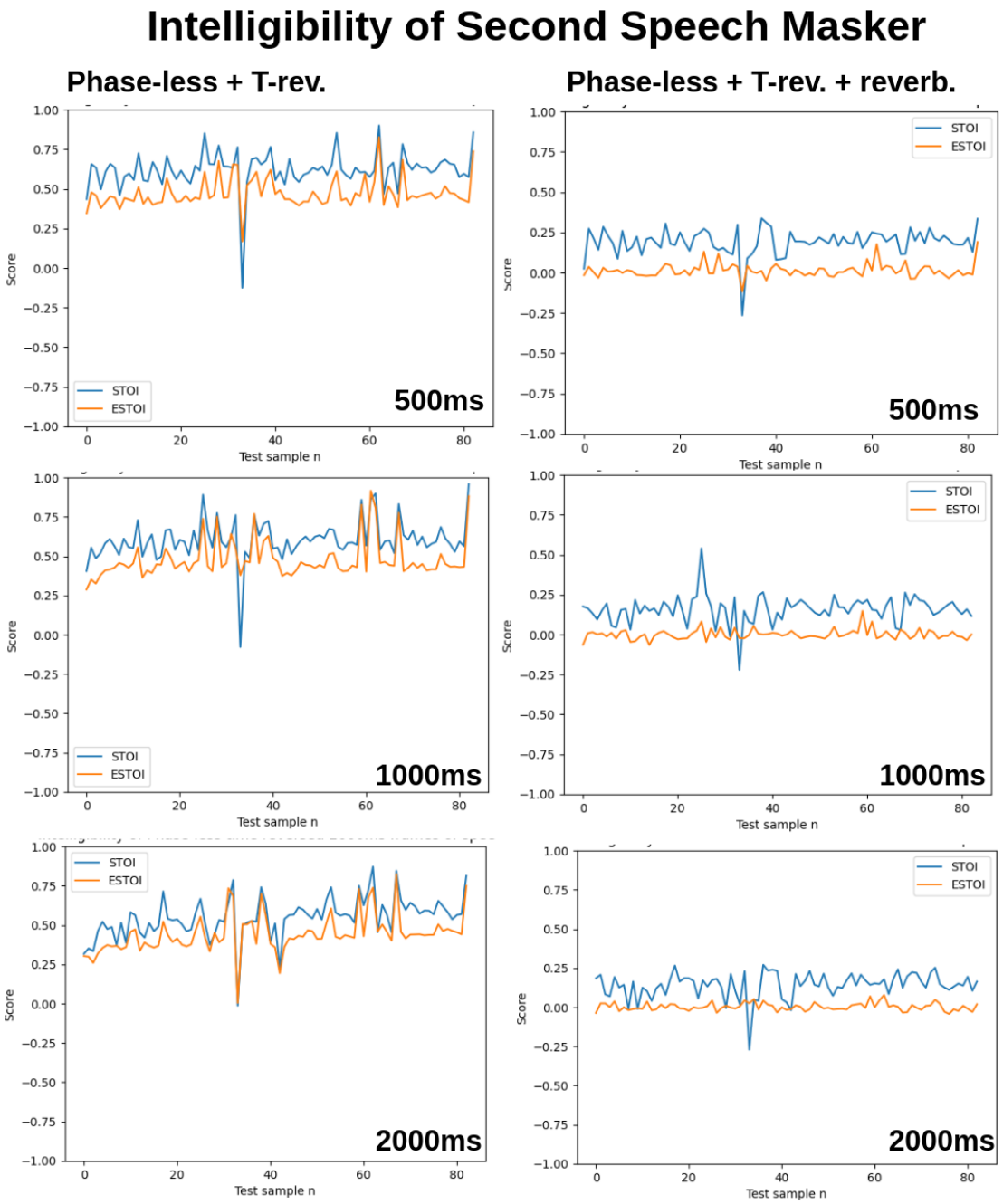


Figure 19: STOI and ESTOI of second masker sound.
Left column: OIMs result on Phase-less time-reversed masker sound for 500ms, 1000ms and 2000ms.
Right column: OIMs result on Phase-less time-reversed with 400ms reverberation masker sound for 500ms, 1000ms and 2000ms.

5 Discussion

The complexity of the speech separation problem becomes increasingly apparent as the data sets progressively align to real-world conditions. This complexity is reflected in the deterioration of separation quality, as observed in the results. For Data Set 1, the synthetic mixtures which featured static sources, the FastICA algorithm performed well, achieving high ESTOI scores and maintaining BI across various SNRs. The results suggest that FastICA is effective when dealing with stationary sources in a controlled environment, as evidenced by the clear and audible separated speech and background sounds in all SNR conditions, with only minor degradation at -12 dB SNR.

In contrast, Data Set 2 of acoustic scene renderings, where the speaker traced a rectangle over ten seconds, presented a more dynamic scenario. The separation results, although showing lower performance metrics, demonstrated successful background separation. One channel consistently contained only background sounds while the other mixed speech and background sounds, indicating that FastICA managed to isolate background audio despite the speaker's movement. However, the separation suffered from alterations in the spectrum, affecting the overall clarity somewhat.

Data Set 3, which involved two 80-second real-world recording with a slowly moving speech source, posed the greatest challenge. SNR between the background and the speaker was highly variable, leading to difficulties in achieving consistent separation. In addition, the fact that every recorded mixture were contaminated with static background noise, indicated that pre-processing did not contribute a sufficiently noiseless model for ICA; the number of Gaussian sources were greater than one. In conjunction to extended mixture duration and dynamic nature of the source movement, these are concluded to be the central factors explaining the poor separation. Also, these features seem to have resulted in, with the number of components to extract being two, that certain background audio events together with speech were more dependent than some other component of the background audio, and consequently separated as such. The separation did not significantly improve even when revisiting smaller time segments (with audible speech), suggesting that the method struggled more with the noisy and dynamic convolutive nature of real-world recordings than longer duration.

The difference in sound levels between the TASCAR settings and the actual recorded levels, as shown in Table 1, further highlights the challenge of replicating real-world conditions in a controlled environment. This discrepancy likely impacted the algo-

rithm's performance, suggesting that more accurate calibration of sound levels was needed.

The FastICA results indicate that positive SNR conditions are particularly challenging for separation, possibly due to the higher energy overlap between speech and background sounds. The success of the fast-paced rectangle path tracing in Data Set 2 might be attributed to sufficient SNR, allowing the algorithm to distinguish between sources despite the rapid movement. Data set 3 further consolidates fluctuating SNR levels and presence of several Gaussian components in the mixture, as the two main factors that deteriorates speech-from-background separation.

Measuring the background intactness as the magnitude of normalised cross-correlation between ground truth and separated background audio seems valid, since when separation with FastICA succeeded during evaluation of Data Set 1 in Figure 15 the BI scores were significantly higher than when separation quality decreased in Data Set 2, seen in Figure 16. Although its performance in relation to human classifiability has not been thoroughly studied, nor has AI classifiability on separated background been evaluated.

The speaker anonymisation technique displayed considerable difference in scoring intelligibility, indicating that when speech contents are audible to human ears, the STOI and ESTOI measures cannot present this and other measures should be used. When considering the division between male and female speakers, the STOI and ESTOI values for both genders remain relatively consistent, with minor variations. This may suggest that the anonymisation process may perform similarly across different speaker characteristics, maintaining intelligibility and similarity to the original speech regardless of gender. However since only two peoples' voices were subjected to this process, this claim will not be elaborated upon further.

The speech masking technique implemented was only partially successful in removing speech cues in the masker sound itself. The difficulty resides in wanting the masker to use the target speech signal as its seed while the masker sound is generated computationally efficient in real-time. The notion of a successful offline speech content masker is the ability to generate it autonomously, which is feasible but not studied here. The deterioration in STOI and ESTOI when adding a slight reverberation effect to the speech masking sound indicate that these scores are misleading and cannot be used in this scenario. This is since the actual perceived change in masking effect is

minimal. Thus, the disproportional disparity of the computed scores does not correlate with subjective intelligibility scoring. Therefore, even though ESTOI is stated by its authors to give a reasonable measure for intelligibility of target speech contaminated with another speech-like signal, this result could not be replicated in this Master's Thesis.

5.1 Future work and conclusion

Given the performance limitations observed, it is clear that the current metrics and methods are not entirely sufficient for comprehensive evaluation and separation in more realistic and dynamic scenarios. Although, the ability to preserve background audio is has been demonstrated to be more frequent than to isolate single voices. Future work should focus on developing online ICA methods that can handle dynamic convolutive MIMO systems. Additionally, identifying suitable contrast functions for speech separation in varying acoustic environments will be imperative. Contrasts might be dynamic and different depending on acoustic environment where surveillance takes place, allowing to tailor to the typical sources present. The need for extensive noise reduction suggests PCA might be complemented with further noise reduction techniques.

In conclusion, while FastICA shows promise in controlled static environments, its performance diminishes with increased realism and source movement. Future advancements in dynamic ICA methods and better acoustic modelling are essential for improving speech separation in complex, real-world audio scenes. In those scenarios, the permutation problem would need to be solved after each separation in real-time, to link background sounds and speech between consecutive audio frames.

As for the voice anonymisation, it can be concluded that online voice anonymisation as described in Section 2.3 is a feasibility in surveillance. Though, whether the effects of the technique is reversible has not been proven here, and such an investigation could be the subject of future research, favourably combining complex analysis and digital signal processing. One can then begin in a setting of complete insight about the original speech signal and pitch shift factors and subsequently decrease the number of known parameters to investigate the possibility of recovering the original speech. In such a case, the validity of the speaker identity masking technique used in this Master's Thesis would be compromised.

The conclusion for the proposed masker sound is this; removing speech cues by producing a phase-less and uniformly time reversed signal is not a sufficient online masker sound. A more intricate procedure is needed if one wants to use the speech signal itself as the seed for the masker to reduce intelligibility of speech. Simultaneously, producing a non-irritating masker sound should be investigated further. Future work should perhaps consider two factors. One is expanding the available memory allocation when streaming such that imposition of masking segments can occur on other parts than their respective seeds. The other is to produce a masker sound that is constant during speech, not unlike babble noise.

The presented objective measures of speech intelligibility lack transferability to objective intelligibility measures for speech-like masker sounds. Similarly for the measure of background intactness, the objective measures depend on availability of ground truth references, limiting their use. Furthermore as results suggests, the similarity in STOI/ESTOI scoring between ground truth clean speech and

- Unsuccessful separation
- Anonymised voice signal
- Speech masked by speech-like masker sound
- Completely other background sounds without speech

show that these measures are ambiguous and only appropriate here for successfully separated speech.

6. References

[1] H. Masuda, Y. Hioka, C. J. Hui, J. James, and C. I. Watson, "Performance evaluation of speech masking design among listeners with varying language backgrounds," *Applied Acoustics*, vol. 201, p. 109 122, 2022, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2022.109122. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X22004960.

[2] A. R. B. Lauren Calandruccio Sumitrajit Dhar, "Speech-on-speech masking with variable access to the linguistic content of the masker speech," *The Journal of the Acoustical Society of America*, vol. 128, pp. 860–9, 2010.

[3] Y. Hioka, J. W. Tang, and J. Wan, "Effect of adding artificial reverberation to speech-like masking sound," *Applied Acoustics*, vol. 114, pp. 171–178, 2016, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2016.07.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16302080.

[4] J. Qian, H. Du, J. Hou, *et al.*, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv e-prints*, arXiv:1711.11460, arXiv:1711.11460, Nov. 2017. DOI: 10.48550/arXiv.1711.11460. arXiv: 1711.11460 [cs.CR].

[5] M. Koutsogiannaki, S. M. Dowall, and I. Agiomyrgiannakis, "Gender-ambiguous voice generation through feminine speaking style transfer in male voices," *ArXiv*, vol. abs/2403.07661, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268364313.

[6] D. S. J. Kreiman, "Producing a voice and controlling its sound," in *Foundations of Voice Studies*. John Wiley Sons, Ltd, 2011, ch. 2, pp. 25–71, ISBN: 9781444395068. DOI: https://doi.org/10.1002/9781444395068.ch2. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444395068.ch2. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444395068.ch2.

[7] W. A. Yost, *Fundamentals of Hearing: An Introduction, Fourth Edition*. Academic Press Limited, 2000, ISBN: 0-12-775695-7.

[8] R. Cristina Oliveira, A. C. Gama, and M. D. Magalhães, "Fundamental voice frequency: Acoustic, electroglottographic, and accelerometer measurement in individuals with and without vocal alteration," *Journal of Voice*, vol. 35, no. 2, pp. 174–180, 2021, ISSN: 0892-1997. DOI: https://doi.org/10.1016/j.jvoice.2019.08.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199719302851.

[9] B. C. J. Moore, *An Introduction to the Psychology of Hearing, Sixth Edition*. Emerald Group Publisher Limited, 2012, ISBN: 978-1-78052-038-4.

[10] ITU-T, "Subjective test methodology for assessing speech intelligibility," International Telecommunications Union's Telecommunication Standardization Sector, Standard, 2016.

[11] IEC, "Iec 60268-16:2020 sound system equipment – part 16: Objective rating of speech intelligibility by speech transmission index," International Electrotechnical Commision, International Standard, 2020.

[12] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217. DOI: 10.1109/ICASSP.2010.5495701.

[13] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. DOI: 10.1109/TASLP.2016.2585878.

[14] A. C. C. Warnock, "Acoustical privacy in the landscaped office," *Journal of the Acoustical Society of America*, pp. 1535–1543, 1973.

[15] M. Zaglauer, H. Drotleff, and A. Liebl, "Background babble in open-plan offices: A natural masker of disruptive speech?" *Applied Acoustics*, vol. 118, pp. 1–7, 2017, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2016.11.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X16304339.

[16] T. Renz, P. Leistner, and A. Liebl, "Auditory distraction by speech: Can a babble masker restore working memory performance and subjective perception to baseline?" *Applied Acoustics*, vol. 137, pp. 151–160, 2018, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2018.02.023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X17311003.

[17] IEC, "Iec 61672-2:2013+amd1:2017 csv electroacoustics - sound level meters - part 2: Pattern evaluation tests," International Electrotechnical Commision, International Standard, 2013.

[18] Y. Hioka, J. James, and C. I. Watson, "Masker design for real-time informational masking with mitigated annoyance," *Applied Acoustics*, vol. 159, p. 107 073, 2020, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2019.107073. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X19305195.

[19] B. Jiang and J. Yang, "Effect of similarity between target speech and time-reversed masker on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 131, no. 4*Supplement*, pp. 3515–3515, Apr. 2012, ISSN: 0001-4966. DOI: 10.1121/1.4709292. [Online]. Available: https://doi.org/10.1121/1.4709292.

[20] C. Jutten and P. Comon, "Chapter 1 - introduction," in *Handbook of Blind Source Separation*, P. Comon and C. Jutten, Eds., Oxford: Academic Press, 2010, pp. 1–22, ISBN: 978-0-12-374726-6. DOI: https://doi.org/10.1016/B978-0-12-374726-6.00006-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123747266000060.

[21] M. Castella, A. Chevreuil, and J.-C. Pesquet, "Chapter 8 - convolutive mixtures," in *Handbook of Blind Source Separation*, P. Comon and C. Jutten, Eds., Oxford: Academic Press, 2010, pp. 281–324, ISBN: 978-0-12-374726-6. DOI: https://doi.org/10.1016/B978-0-12-374726-6.00013-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123747266000138.

[22] E. Vincent and Y. Deville, "Chapter 19 - audio applications," in *Handbook of Blind Source Separation*, P. Comon and C. Jutten, Eds., Oxford: Academic Press, 2010, pp. 779–819, ISBN: 978-0-12-374726-6. DOI: https://doi.org/10.1016/B978-0-12-374726-6.00024-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123747266000242.

[23] V. Zarzoso, P. Comon, and D. Slock, "Chapter 15 - semi-blind methods for communications," in *Handbook of Blind Source Separation*, P. Comon and C. Jutten, Eds., Oxford: Academic Press, 2010, pp. 593–638, ISBN: 978-0-12-374726-6. DOI: https://doi.org/10.1016/B978-0-12-374726-6.00020-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123747266000205.

[24] I. T. Jolliffe, *Principal Component Analysis* (Springer Series in Statistics). New York, NY: Springer New York, 1986, ISBN: 9781475719062. DOI: 10.1007/978-1-4757-1904-8. [Online]. Available: http://link.springer.com/10.1007/978-1-4757-1904-8.

[25] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000, ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(00)00026-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608000000265.

[26] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999. DOI: 10.1109/72.761722.

[27] "The negentropy principle of information," *Journal of Applied Physics*, vol. 24, no. 9, pp. 1152–1163, 1953.

[28] E. Moreau and P. Comon, "Chapter 3 - contrasts," in *Handbook of Blind Source Separation*, P. Comon and C. Jutten, Eds., Oxford: Academic Press, 2010, pp. 65–105, ISBN: 978-0-12-374726-6. DOI: https://doi.org/10.1016/B978-0-12-374726-6.00008-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123747266000084.

[29] D. Pham, "Chapter 2 - information," in *Handbook of Blind Source Separation*, P. Comon and C. Jutten, Eds., Oxford: Academic Press, 2010, pp. 23–63, ISBN: 978-0-12-374726-6. DOI: https://doi.org/10.1016/B978-0-12-374726-6.00007-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123747266000072.

[30] P. Stoica and R. Moses, "Spectral analysis of signals," in PHI Learning, 2011, ch. 6, ISBN: 9788120343597. [Online]. Available: https://books.google.se/books?id=POSbjwEACAAJ.

[31] S. Spanne, "Lineära system," in Kompendieförmedlingen Sigma Aktiebolag, 1997, ch. 12,13.

[32] T. E. of Encyclopædia Britannica, "Equal temperament," in *Encyclopædia Britannica*, Britannica, 2019. [Online]. Available: https://www.britannica.com/art/equal-temperament.

[33] J. Ramírez, J. Gorriz, and J. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in Jun. 2007, vol. 6(9), ISBN: 978-3-902613-08-0. DOI: 10.5772/4740.

[34] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, https://github.com/snakers4/silero-vad, 2021.

[35] G. Grimm, J. Luberadzka, T. Herzke, and V. Hohmann, "Toolbox for acoustic scene creation and rendering (tascar): Render methods and research applications," 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:53060207.

[36] Mixkit, "Free sound effects," Open-source data, 2024, url https://mixkit.co/free-sound-effects/.

Master's Theses in Mathematical Sciences 2024:E52
ISSN 1404-6342

LUTFMS-3501-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/