FACULTY OF ENGINEERING
AND ARCHITECTURE

"I can 't understand you, there is someone speaking over you!" Sound separation using distributed microphone arrays

Martijn Meeldijk
Student number: 02111587

Supervisor: Prof. dr. ir. Nilesh Madhu
Counsellors: Stijn Kindt, Alexander Bohlender

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Information Engineering Technology

Academic year 2022-2023

GHENT
UNIVERSITY

Acknowledgements

Vul aan…

Toelichting in verband met het masterproefwerk

Deze masterproef vormt een onderdeel van een examen. Eventuele opmerkingen die door de beoordelingscommissie tijdens de mondelinge uiteenzetting van de masterproef werden geformuleerd, werden niet verwerkt in deze tekst.

Melding van vertrouwelijkheid (enkel indien van toepassing)

Bekijk hiervoor de informatie op de facultaire website - **Nota in verband met de vorm van de masterproef (alle opleidingen)**

Abstract

Meer informatie op https://masterproef.tiwi.ugent.be/verplichte-taken/ - Korte abstract (Nederlands en/of Engels)

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

List of Acronyms

ASN Acoustic Sensor Network.

FFT Fast Fourier Transform.

NMF Nonnegative Matrix Factorization.

WASN Wireless Acoustic Sensor Network.

List of Code Fragments

1

Introduction

1.1 Problem definition

Many devices such as Amazon's Alexa and Google Home require the processing of human speech to function. This is not always as straightforward as one would expect. Especially considering the input of said devices often contains a great deal of background noise, such as reverberation or other unwanted sound sources. One solution is to make use of microphone arrays in order to help extract the wanted speech component from the signal. A microphone array consists of multiple microphones placed close to each other, typically inside the same device. From the microphone array's signals, time-frequency masks can be created. These are subsequently used in combination with the short-time Fourier transform of the signal to extract the speech component.

To further improve on this solution, it is possible to make use of (ad hoc) distributed microphone arrays. This way, several microphones or microphone arrays are placed at different locations, opening up new possibilities for signal processing such as utilizing the different amplitudes of the signals received in different locations. Signal capture using ad hoc distributed microphones, or acoustic sensor networks (ASNs), is an active and rapidly expanding field of research. With the inclusion of microphones in an increasing variety of smart devices, distributed audio capture is becoming increasingly available - with potential for application in a wide range of fields such as surveillance for assisted living and healthcare, hearing aids, communications. The challenges, however, are also manifold. Compared to traditional, compact microphone arrays, where multiple microphones are placed close to each other with predefined geometries, the relative locations of sensors are not known a priori, and their placement with respect to audio sources of interest can be arbitrary. The processing power and bandwidth available to each node can also be limited - constraining on-edge processing and data communication with a central hub.

1.2 Motivation

1.3 Summary of results

2

Background

2.1 Signal Model

The acoustic environment considered in this thesis consists of N acoustic sources and D microphones which are distributed, typically in an unknown arrangement, within the boundaries of the environment. The signal received by a microphone d may be described in continuous time t as:

$$x_d(t) = \sum_{n=1}^{N} \int_0^\infty h_{nd}(\tau) s_n(t-\tau) d\tau \tag{2.1}$$

- $s_n(t)$: the n-th source signal
- $h_{nd}(t)$: the impulse response from source n to microphone d
- $x_d(t)$: the resulting microphone signal

This can displayed by making use of the convolution operator like so:

$$x_d(t) = \sum_{n=1}^{N} h_{nd}(t) * s_n(t) \tag{2.2}$$

The microphone signals are sampled and transformed to the short-time discrete Fourier domain.

$$X_d(k,b) = \text{STFT}[x_d(l)] \tag{2.3}$$

- $x_d(l)$: the sampled signal of microphone d
- l: the time sample index
- k: the frequency bin index
- b: the time frame index

As of now, this model doesn't take into consideration possible interference or noise. This will be discussed in the next paragraph.

2.1.1 Interference

Taking into account an interference signal … TODO

2.2 Source separation

Source separation aims to obtain a signal containing only the target source, effectively suppressing all other sources except the target. The aforementioned principle is recognized as adaptive 'nulling'.[1] In this part, an emphasis will be put on source separation in ad hoc setups, so any techniques that require positional information on microphone nodes will not be discussed.

2.2.1 Time-Frequency masking

By assuming that sources are approximately disjoint in the short-time-frequency plane, it is reasonable to assume that only one source is dominant at any time-frequency point. This allows for the estimation of a spectral mask $\mathscr{M}_n(k, b)$ which suppresses time-frequency points that do not belong to the target source, effectively suppressing interferers. TODO

$$\mathscr{M}_n(k, b) = \begin{cases} 1 & \text{if source } n \text{ is dominant at } (k, b) \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

TODO

2.3 Clustering source dominated microphones

2.3.1 Fuzzy clustering

TODO

2.3.2 Federated Learning

TODO

2.3.3 Coherence-based clustering

A relatively novel and slightly different approach [2] to clustering in ad-hoc microphone arrays proposes a method based on the magnitude-squared-coherence between microphones' observations, which measures their degree of linear dependency by analyzing similar frequency components.

Subsequently, a non-negative matrix (NMF) based approach is utilized, with the goal of obtaining an optimal clustering, whereby nodes are assigned into subnetworks based on their respective microphone observations.

The suggested method offers the capability to dynamically perform clustering while imposing a low computational burden, rendering it highly applicable to various audio signal processing applications. Consequently providing a notable advantage in terms of processing efficiency, making the method an attractive option for real-world scenarios where computational resources may be limited or where rapid processing is required.

2.3.3.1 Signal model

To include interference in the signal model, the observed signal $x_d(t)$ at microphone d can be represented like so:

$$x_d(t) = s_d(t) + v_d(t) \tag{2.5}$$

Where $v_d(t)$ denotes the noise signal plus interference at time instant t. The linear signal model in equation 2.5 can be conveniently restated to denote the collection of a frame of samples into a vector form:

$$\begin{aligned} \mathbf{x}_d(t) &= [x_d(t)x_d(t-1)\cdots x_d(t-T+1)]^T \\ &= \mathbf{s}_d(t) + \mathbf{v}_d(t) \end{aligned} \tag{2.6}$$

- T: frame size
- $\_^T$: matrix transpose
- $\mathbf{x}_d(t)$: observed signal vector
- $\mathbf{s}_d(t)$: clean speech vector
- $\mathbf{v}_d(t)$: noise signal vector

2.3.3.2 Clustering algorithm

Magnitude-squared coherence

By utilizing the magnitude squared coherence, it is possible to conduct an analysis of the linear relationship between two signals $x(t)$ and $y(t)$. First, the Fast Fourier Transform (FFT) of both signals is computed. After which the coherence is measured as a function of the center frequency of the filter. The magnitude-squared coherence can be obtained with the following formula [3]:

$$\Gamma_{xy}(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)} \tag{2.7}$$

- f: The center frequency of the filter
- S_{xx}: The auto spectral density of x
- S_{yy} The auto spectral density of y
- S_{xy} The cross-spectral density

The power spectra $S_{xx}(f)$ and $S_{yy}(f)$ describe the distribution of power into frequency components composing the signals $x(t)$ and $y(t)$. [4] To compute the cross-spectral density, the following equation can be used:

$$S_{xy}(f) = \sum_{k=1-T}^{T-1} R_{xy}(k)e^{-i2\pi fk} \tag{2.8}$$

- $R_{xy}(k)$: The cross-correlation between $x(t)$ and $y(t)$
- T: The frame size

For the special case $x(t) = y(t)$, equation 2.8 reduces to $S_{xx}(f)$, $R_{xy}(k)$ can be estimated with:

$$R_{xy}(k) = \begin{cases} \frac{1}{T}\sum_0^{T-1-k} x(t)y(t+k) & k = 0, \dots, T-1 \\ R_{xy}(-k) & k = -(T-1), \dots, -1 \end{cases} \tag{2.9}$$

Now, sufficient information is provided to be able to compute $\Gamma_{xy}(f)$. This calculation yields a value between 0 and 1, with higher values denoting a stronger linear correlation. By calculating

$$C_{xy} = \frac{\sum_{f=0}^{F} \Gamma_{xy}(f)}{F} \in [0, 1] \tag{2.10}$$

all frequency bins are assigned the same weight regardless of their power. By arranging all coherence measures C_{xy} between the audio signals, a non-negative coherence matrix **C** can be obtained.

$$\boldsymbol{C} = \begin{bmatrix} 1 & \cdots & \cdots & C_{1M} \\ C_{12} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ C_{1M} & C_{2M} & \cdots & 1 \end{bmatrix} \in \mathbb{R}_+^{M \times M} \tag{2.11}$$

2.3.3.3 Non-negative matrix factorization

The matrix $\boldsymbol{C}$ contains values that represent the degree of correlation between the signals observed by each combination of microphones, meaning that each row (or column) j of $\boldsymbol{C}$ represents the degree of correlation between the j-th microphone signal and all other signals. As a result, groups of microphones close to a specific source will be highly correlated.

The next step consists of exploiting the inherent clustering property of NMF [5]. $\boldsymbol{C}$ can be considered as a linear subspace of dimension M. By downgrading this subspace into a linear subspace with dimension corresponding to the amount of sources K, a clustering can be achieved. The matrix $\boldsymbol{C}$ is non-negative and can be modelled as:

$$\boldsymbol{C} = \boldsymbol{B}\boldsymbol{B}^T \odot (\mathbf{1} - \boldsymbol{I}) + \boldsymbol{I} \tag{2.12}$$

- $\boldsymbol{B} \in \mathbb{R}^{M \times K}$: The cluster matrix, where K is the amount of speakers (the amount of clusters)
- $\odot$: Element-wise product
- $\boldsymbol{I}$: The identity matrix
- $\mathbf{1}$: The all-ones matrix

The latter two are introduced because the main diagonal of $\boldsymbol{C}$ does not provide any relevant information in the learning process of $\boldsymbol{B}$. Because $\boldsymbol{C}$ is symmetric, we model it as $\boldsymbol{B}\boldsymbol{B}^T$. It is possible to estimate $\boldsymbol{B}$ using iterative multiplicative update rules based on Euclidian divergence [6]:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \odot \frac{(\boldsymbol{C} \odot (\mathbf{1} - \boldsymbol{I}))\boldsymbol{B}}{(\boldsymbol{B}\boldsymbol{B}^{\boldsymbol{T}} \odot (\mathbf{1} - \boldsymbol{I}))\boldsymbol{B}} \tag{2.13}$$

Now each column of $\boldsymbol{B}$ contains the contribution of a microphone to each cluster. We can obtain the clustering result with:

$$\gamma_m = \{j \in [1, K] : B_{mj} \geq B_{mk}, \forall k \in [1, K]\} \tag{2.14}$$

The value γ_m denotes the cluster assigned to the m-th microphone. This is simply the largest value of column m.

2.4 Proposed methods

3

Implementation

3.1 Sectie titel

Vul aan...

4

Nothing yet

Conclusion

References

[1] N. Madhu, “Acoustic source localization: Algorithms, applications and extensions to source separation,” Ph.D. dissertation, Ruhr-Universität Bochum, 2009.

[2] M. G. C. Antonio J. Munoz-Montoro, Pedro Vera-Candeas, “A Coherence-based Clustering Method for Multichannel Speech Enhancement in Wireless Acoustic Sensor Networks,” pp. 1–5, 2018.

[3] W. A. Gardner, “A unifying view of coherence in signal processing,” *Signal Processing*, vol. 29, no. 2, pp. 113–140, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165168492900150

[4] P. Stoica and R. Moses, *Spectral analysis of signals*. Prentice Hall, 2004. [Online]. Available: https://user.it.uu.se/~ps/SAS-new.pdf

[5] H. D. S. Chris Ding, Xiaofeng He, “On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering,” p. 606–610, 2005.

[6] D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf

Appendices

Bijlage A

Toelichting bijlage.

Bijlage B

Toelichting bijlage.