An adaptive Speech Denoising System based on ICA with Voice Activity Detection

by

António Rui Ferreira Rebordão,

Thesis

Presented to the Faculty members of

The University of Tokyo

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Department of Information and Communication Engineering

Graduate School of Information Science and Technology

The University of Tokyo

February 2008

To my family

Acknowledgments

My sincere thanks go to Professor Keikichi Hirose and Dr. Md. Khademul Islam Molla for their generosity in providing me with ideas, comments, suggestions, constant support and materials.

I acknowledge the University of Tokyo and the Japanese Ministry of Education for their precious support and, last but not least, I want to express my deepest gratitude to my family and friends because life without them would be meaningless.

ANTÓNIO RUI FERREIRA REBORDÃO

An adaptive Speech Denoising System based on ICA with Voice Activity Detection

António Rui Ferreira Rebordão, MSc
The University of Tokyo, 2008

Supervisor: Professor Keikichi Hirose

This contribution presents an innovative system for adaptive speech denoising using Independent Component Analysis (ICA) and Voice Activity Detection (VAD) in low dB-SNR environments. The implemented experiments consider instantaneous mixtures (two sources and two microphones) where the proposed system identifies the noise contained in each noisy mixture, applies the most suitable block ICA method among 3 methods (FastICA, Kernel ICA and JADE) and, after source separation, automatically identifies the estimated speech signal. The ICA suitability is in accordance with the detected noise, the signal mixtures are non-linear and the proposed system extracts information that can be used for further pre and/or post-processing and for improving the block ICA's output. The process is completely automatic from the source recording to its output and such system has a wide range of applications and significant potential over the conventional approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General introduction

For Huang (2001) since human prehistory that speech communication has been and will be the dominant mode of human social bonding and information exchange. The spoken word is now extended, through technological mediation such as telephony, movies, radio, television, and the Internet. This trend reflects the primacy of spoken communication in human psychology. In addition to human-human interaction, this human preference for spoken language communication finds a reflection in human-machine interaction as well. Most computers currently utilize a graphical user interface (GUI), based on graphically represented interface objects and functions such as windows, icons, menus, and pointers and their operating systems also depend on a users keyboard strokes and mouse clicks, with a display monitor for feedback. It is easily seen that todays computers lack the fundamental human ability to speak, listen, understand, and learn. However, speech, supported by other natural modalities, will be one of the primary means of interfacing with computers, changing the way we live and work [1].

In some environments (e.g. inside a car, public mobile phone usage, etc.), higher flexibility and safety standards can be achieved by using human voice commands to retrieve information from navigation systems or execute simple command tasks [2]. A number of commercial speech recognition systems are already available (e.g. speech support in Microsoft Office 2007). However the performance of those systems degrade substantially under real-world conditions because the

physical channels used in the process are prone to signal distortion and musical noise. Consequently the speech signal is recorded as a mixture of several signals mixed together and so it arises the need for speech enhancement.

1.2 Speech enhancement overview

Speech enhancement aims at improving the performance of speech communication systems in noisy environments by suppressing the noise and improving the perceptual quality and intelligibility of the speech signal [3], mainly through noise reduction algorithms. Such types of processes may be applied to a mobile radio communication system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired.. This problem remains a challenging task in real-world environments [4] and to solve it several approaches were made. However, such approaches can be significantly improved if they take into account some factors [5] related with the nature and properties of noise.

For extracting information about the noise contained by the noisy mixture it is important to identify the noise source(s) and how they behave through time. For such aim it is possible to use techniques like Voice Activity Detection algorithms that allow the identification of the voiceless components contained in the noisy mixtures and by analyzing them is possible to extract valuable information that can be used for effective speech enhancement.

Due to its significant importance in todays information technology, the topic is widely researched. The performance of such systems is mainly evaluated according to the quality and intelligibility. The quality of the enhanced signals refers to its clarity, distorted nature and the level of the residual noise in that signal. Most speech enhancement methods improve the quality of the signal however degrades its intelligibility, which refers to the understandability of the enhanced speech; the percentage of words that could be correctly identified by the listener. Human listeners can usually extract more information from the noisy signal than from the enhanced signal by careful listening. Since quality and intelligibility require live listening sessions, they are both time consuming and expensive to measure. That is why; researchers mostly use some mathematical measures which are believed to be correlated with the quality and intelligibility of the enhanced speech. For this

purpose, signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) tests are widely used to show the performance of the proposed algorithms in terms of their quality. For assessing the intelligibility of the enhanced signal, automatic speech recognition tests are commonly used.

1.3 Speech enhancement methods

Speech enhancement schemes currently available can be subdivided into single microphone and multiple microphone methods. Single-microphone algorithms are most commonly encountered and are based on temporal-spectral information about the recorded signals [6]. A variety of schemes combining different time-spectral and cepstral domain based speech processing methods have been proposed for robust speech recognition purposes [7, 8]. The traditional framework used in single-microphone enhancement techniques is a probabilistic one with statistical models of a speech signal corrupted by additive Gaussian noise [9]. The noise signal estimate is commonly adapted from the most recent recording, i.e. a few seconds before the command is spoken, or voice activity detection algorithms are used to estimate noise power from noisy speech silence intervals. This approach works well when the noise signal is reasonably stationary. Perceptually inspired processing techniques [10] and variations of cepstral mean subtraction approaches for speech recognition [11] have been successfully applied to handle convolutional noise and speech reverberation as well. However performance is unsatisfactory when strongly reverberated speech signals recorded in non-stationary noise environments are considered or the desired speaker signal is corrupted by highly interfering speech sources [6].

To overcome the limitations of single microphone temporal processing methods, spatial information can be exploited by using multiple microphones. In beamforming [12, 13] for example, an array of microphones with a known geometry enabling both spatial and temporal measurements of sounds is used to suppress interfering signals. Acoustic room modeling and source localization can be performed as well as reverberation be handled to some extent with adaptive algorithms [12, 13].

Multiple-microphone configurations for speech processing play an ever increasing role in multimedia systems, video-conferencing facilities, computer inter-

faces, etc. [12, 14, 15, 16]. Probabilistic denoising approaches using multiple models for both speech and noise sources [17] would require a very large model database to work reliably in such unknown and challenging acoustic environments. In these cases we could achieve good performance with large microphone arrays but its implementation is difficult and expensive. However, the number of microphones can be reduced by using second or higher-order decorrelation based, Blind Source Separation algorithms [18, 19, 20]. These signal processing algorithms exploit spatial information about signal mixtures recorded at a limited number of microphone locations to explicitly separate interfering noise signals from the desired source signal. Since they assume no a priori information about the interfering sources, they are particularly suited for environments where the number of disturbance scenarios is virtually unlimited (e.g. The Cocktail Party problem) [21]. This thesis focus on ICA methods that aim at speech denoising based on noise identification.

Many Speech Enhancement techniques go beyond the scope of this thesis but some popular approaches are as follows:

- Spectral subtraction (a traditional method for single mixtures)
- Independent Component Analysis (ICA is one of the popular approaches to BSS)
- ICA with reference (an interesting approach that combines ICA and EMD)

1.4 Speech enhancement applications

In the current information technology, there are many areas that speech enhancement is used in order to improve the performance of the system:

- **Robust Automatic Speech Recognition (RASR):** The accuracy of automation speech recognition degrades in the presence of background noise or other interfering sources. Noise reduction for speech signals has therefore critical importance as a pre-process of such types of systems, including human-computer interactions, robotics and audio driven systems, etc.

- **Telecommunication:** Background noise is a common problem which degrades the quality of the communication for the human listener. Speech enhancement may be applied to such systems in order to remove the unwanted noise sources. Another problem in telecommunication is the channel noise. By enhancing the speech signal before it goes into the channel, it is also possible to reduce the effect of the channel noise.

- **Digital Hearing Aids:** The digital hearing aid users often complain of difficulty in understanding speech in the presence of background noise. Therefore, speech enhancement is an important process to improve the speech perception in a noisy environment for the human listener.

1.5 Thesis overview

This research presents an innovative system for adaptive speech denoising using ICA and Voice Activity Detection (VAD) and it is described in fig. 5.1.

Designed for instantaneous speech mixtures with two sources and two microphones the proposed system identifies the noise contained in each noisy mixture, applies the most convenient block ICA method among 3 methods (JADE, KERNEL ICA and FastICA) and, after source separation, identifies the estimated speech signal. The ICA suitability is in accordance with the detected noise.

The mixing process is non-linear and the information extracted on the first stage can be used for later post-processing and further system extension with significant potential over the conventional approaches. The process is completely automatic from the source recording to its output and such system has a wide range of applications. This approach and its experimental data significantly provides auspicious potential over the systems currently available.

1.5.1 Problem statement

In many speech related systems, the desired signal is not easily available; usually is contaminated with some interference sources. These background noise signals degrade the quality and intelligibility of the original speech, resulting in a severe drop in the performance of the applications. There are different types of

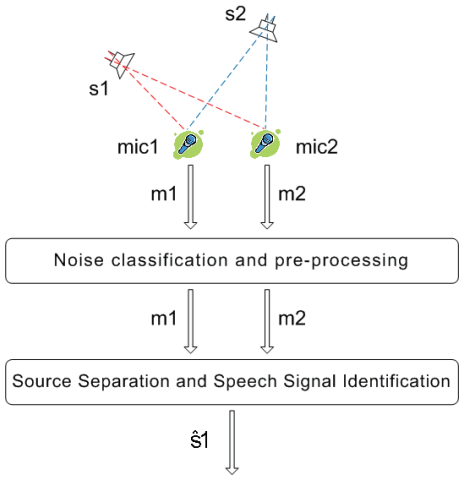


Figure 1.1: The proposed system

noise signals which affect the quality of the original speech. It may be a wide-band noise in the form of a white or colored noise, a periodic signal such as in hum noise, room reverberations, or it can take the form of fading noise. It is also possible that the speech signal may be simultaneously attacked by more than one noise source. The most common type of noise in time series analysis and signal processing is the white noise, pink noise and other noises contained by Noisex. That is why; this thesis is mainly concerned with these type of noises.

The degradation of the speech signal due to the background noise is a severe problem in speech related systems and therefore should be eliminated through speech enhancement algorithms. Speech enhancement aims at improving the perceptual quality and intelligibility of a speech signal in noisy environments, mainly through noise reduction algorithms. Such types of processes may be applied to a mobile radio communication system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired. Figure 1.1 shows an illustration of the usage of speech enhancement. It can be observed that enhancement may also be applied directly to the clean speech signal in order to reduce the effect of the channel noise in communication systems.

1.5.2 Motivation

The degradation of the speech signal due to the background noise is a severe problem in speech related systems and therefore should be eliminated through speech enhancement algorithms. Combining such methods with BSS methods can improve the existence of the existing systems. This is our main goal and motivation.

1.5.3 Outline of the thesis

This research presents an innovative system for adaptive speech denoising aimed for blind noisy environments (the ones where both the sources and the mixing process are unknown and only recordings of the mixtures are available). The proposed system extracts information that can be used for ICA's automatic suitability and further pre and/or post-processing. Such approach improves the performance of the current available methodologies.

The structure of this thesis is organized as follows:

- **Chapter 1** is an overview of our research, motivation and outline.
- **Chapter 2** introduces Voice Activity Detection (VAD), its problem analysis, algorithms and some generalities about additive noise.
- **Chapter 3** briefly explains Blind Source Separation (BSS) and its applications.
- **Chapter 4** describes a popular BSS methodology used in our research.
- **Chapter 5** presents our research, its setup, experimental data and its analysis.
- **Chapter 6** concludes the paper.

1.6 Publications derived from this work

Peer-reviewed conference papers:

[22] António R. F. Rebordão; M. K. Islam Molla; Keikichi Hirose; Minematsu Nobuaki "A Speech Denoising System based on ICA and Voice Activity Detection", *Proc.*

of Acoustical Society of Japan (ASJ 2008 Spring Meeting), Chiba, Japan; 17-19 March, 2008.

[23] António R. F. Rebordão; M. K. Islam Molla; Keikichi Hirose; Minematsu Nobuaki ”An adaptive Speech Denoising System based on ICA and Voice Activity Detection”, *Proc. of International Workshop on Nonlinear Circuits and Signal Processing (NCSP08)*, Gold Coast, Australia; 6-8 March, 2008.

Conference papers waiting for acceptance:

[24] António R. F. Rebordão; M. K. Islam Molla; Keikichi Hirose; Minematsu Nobuaki ”Adaptive ICA usage for signal enhancement”, *International Conference on Audio, Language and Image Processing 2008 (ICALIP 2008)*, Shangai, China; 7-9 July, 2008.

Chapter 2

Voice Activity Detection

For [25], in speech communications, noise is fluctuations and the addition of external factors to the stream of target information (signal) being received at a sensor. It may be deliberate as for instance jamming of a radio or video signal, but in most cases it is assumed to be merely undesired interference with intended operations. Many speech processing systems users are familiar with the amount of background noise present in loud environments. This is because their hands free instruments amplify environment noise just as much as the conversation that they are trying to follow. Work is ongoing to suppress background noise as much as possible to positively influence the intelligibility of the speech in noisy environments.

Although speech processing in artificially constrained conditions has recently reached high levels of performance, problems still remain in the deployment of speech recognition technology in the real world. One of the problems is the performance degradation of speech detection when they are used in noisy environments such as offices, automobile cabins, streets and computer rooms. Many reasons account to eliminate or reduce noise from speech signals. However one of the biggest challenges is to avoid removal of speech components in this process.

Speech or Voice Activity Detector (VAD), aims to distinguish between speech and several types of acoustic background noise even with low signal-to-noise ratios (SNRs). In the field of multimedia applications, a VAD permits simultaneous voice and data applications. Similarly, in Universal Mobile Telecommunications Systems (UMTS), it controls and reduces the average bit rate and enhances overall coding quality of speech. In cellular radio systems (GSM and CDMA systems) based

on Discontinuous Transmission (DTX) mode, this facility is essential for enhancing system capacity by reducing co-channel interference and power consumption in portable digital devices.

It is very difficult to distinguish between noise and silence, in the presence of background noise, so more efficient and self-sustaining algorithms are needed for speech activity detection and noise reduction is changing and adverse noise acoustic background. There are different metrics used for speech detection in VAD algorithms, but recently Higher-order statistics (HOS) have shown potential results in a number of signal processing applications, and are of particular value when dealing with a mixture of Gaussian and non-Gaussian processes and non linear systems [25].

The system presented in this research is variance based (instead of HOS) and the results are satisfactory enough for the intended aim. Thus, our approach deviated from HOS to VAD's Variance Based because this approach can deal perfectly well with the noise types considered (NOISEX database).

2.1 Problem Analysis

The main question for this section is to explain how additive noise (in the form of gaussian noise) corrupted with clean speech can be suppressed or isolated.

The process of separating conversational speech and silence is called the voice activity detection (VAD). It was first investigated for use on Time Assigned Speech Interpolation (TASI) systems. VAD is an important enabling technology for a variety of speech-based applications including speech recognition, speech encoding, and hands-free telephony. For these purposes, various types of VAD algorithms were proposed that trade off delay, sensitivity, accuracy and computational cost.

The primary function of a voice activity detector is to provide an indication of speech presence in order to facilitate speech processing as well as possible provide delimiters for the beginning and end of a speech segment. For a wide range of applications such as digital mobile radio, Digital Simultaneous Voice and Data (DSVD) or speech storage, it is desirable to provide a discontinuous transmission of speech-coding parameters. The advantage can be a lower average power consumption in mobile handsets, or a higher average bit rate for simultaneous services

like data transmission or even a higher capacity on storage chips. However, the improvement depends mainly on the percentage of pauses during speech and the reliability of the VAD used to detect these intervals. On on hand, is advantageous to have a low percentage of speech activity but, on the other hand, clipping of active speech should be avoided to preserve the quality. This is a crucial problem for a VAD algorithm under heavy noise conditions.

Voice activity detection is important for speech transmission, enhancement and recognition. The variety and the varying nature of speech and background noise makes it challenging. Earlier algorithms for VAD are based on the Itakura LPC distance measure, energy levels, timing, pitch and zero crossing rates, cepstral features, adaptive noise modeling of voice signals and the periodicity measure. Unfortunately, these algorithms have some problems for low SNR values, especially when the noise is non-stationary. Consistent accuracy cannot be achieved since most algorithms rely on a threshold level for comparison. This threshold level is often assumed to be fixed or calculated in the silence intervals. During the last decade numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD decision on speech processing systems.

2.2 VAD Algorithm: The Principle

The basic function of a VAD algorithm is to extract some measured features or quantities from the input signal and to compare these values with threshold, usually extracted from the characteristics of the noise and the speech signals. Voice-active decision is made if the measured values exceed the thresholds. VAD in non-stationary noise requires a time-varying threshold value. This value is usually calculated in the voice-inactive segments.

A representative set of recently published VAD methods formulates the decision rule on a frame by frame basis using instantaneous measures between speech and noise. The different measures which are used in BAD methods include spectral slope, correlation coefficient, log likelihood ratio, cepstral, weighed cepstral, and modified distance measures.

A VAD can be decomposed in two steps: the computation of metrics and the application of a classification rule. Independently from the VAD method, the

operation is a compromise between having voice detected as noise or noise detected as voice. A VAD operating in a mobile environment must be able to detect speech in the presence of a range of very diverse types of acoustic background noises. In these difficult detection conditions it is vital that a a VAD should ”fail-safe”, indicating ”speech detected” when the decision is in doubt so that no clipping is introduced. The biggest difficulty in the detection of speech in this environment is the very low signal-to-noise ratios (SNRs) that are encountered. It is impossible to distinguish between speech and noise using simple level detection techniques when parts of the speech utterance are buried below the noise.

Robust voice activity detection algorithms are required, as traditional solutions present a high misclassification rate in the presence of the background noise typical of mobile environments. One important aspect of recent digital cellular systems is the robustness of the speech coding algorithms needed for the channel to be used efficiently. They have to be robust, not only to channel degradation, but also to the background noise typical of mobile environment. The underlying definition of the robustness can be formulated as *”a VAD is robust if it gives decisions close to a reference in quiet as well as in adverse environments*. There is introduced a new definition claiming that a VAD is robust when it gives similar decisions for clean speech and noisy speech. The robustness can be estimated by taking the VAD’s decision on clean speech as a reference and computing error statistics of the same VAD applied to noisy speech. The more robust the VAD, the scarcer the errors.

2.3 Generalities about Noise

Noise can be defined as the contamination of the desired signal or the unwanted signal. Natural and deliberate noise sources can provide both or either of random interference or patterned interference. Only the latter can be cancelled effectively in analog systems; however, digital systems are usually constructed in such a way that their quantized signals can be reconstructed perfectly, as long as the noise level remains below a defined maximum, which varies from application to application. There are many forms of noise with various frequency characteristics that are classified by ”color”.

White noise is a signal (or process) with a flat frequency spectrum. In other

words, the signal has equal power in any band, at any frequency, having a given bandwidth. In practice a signal can be ”white” with a flat spectrum over a defined frequency band. A signal that is ”white” in the frequency domain must have zero autocorrelation with itself over time, except at zero time shift. The figures 2.1 and 2.2 show that car noise taken for 10000 samples is not white. The periodogram shows that the spectrum is not uniform where as the randomly generated Gaussian noise has a uniform distribution. The power spectral density is the smoothed version of the periodogram.

Noise having a continuous distribution, such as a normal distribution, can be white. Gaussian noise is sometimes misunderstood to be white gaussian noise, but this is not so. Gaussian noise only means noise with pdf of the Gaussian distribution, which says nothing to correlation of the noise in time. Labeling Gaussian noise as white describes the correlation of the noise.

The next most commonly used colored noise is pink noise. Its frequency spectrum is not flat, but has equal power in bands that are proportionally wide. Pink noise is perceptually white. That is, the human auditory system perceives approximately equal magnitude in all frequencies. The power density decreases by -3 dB per octave with increase in frequency (density proportional to $\frac{1}{f}$). There are also many ”less official” colors of noise such as brown, blue, purple, violet, grey, red, orange, green and black.

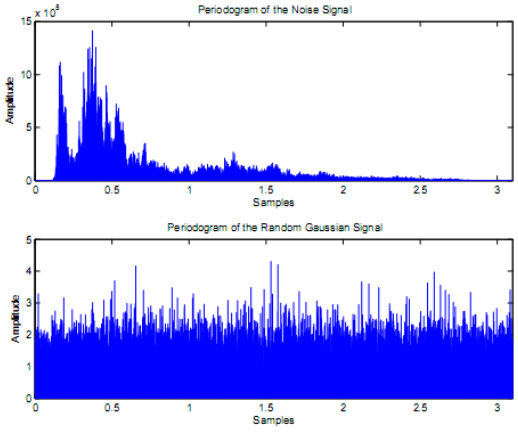


Figure 2.1: Test for Whiteness of Noise in the CAR Noise

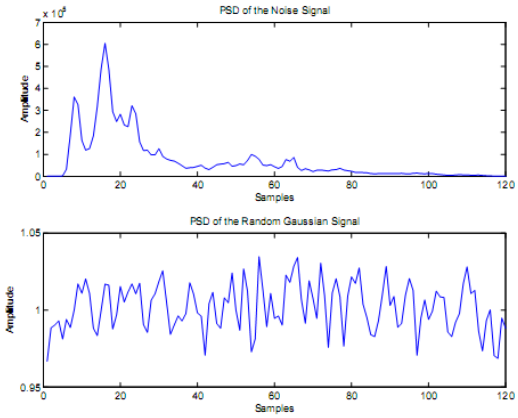


Figure 2.2: Test for Type of Noise in the CAR Noise

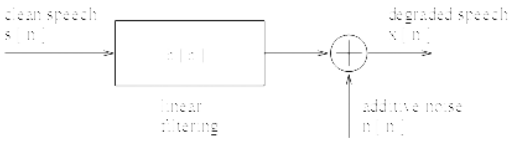


Figure 2.3: Representation of Additive Noise

2.3.1 Additive Noise

There are many sources of acoustic distortion that can degrade the performance of speech recognition systems. For many speech recognition application the most important source of acoustical distortion is the additive noise. Much research effort in robust speech recognition has been devoted to compensate the effects of additive noise.

Is the speech signal $s(k)$ affected by uncorrelated noise $n(k)$, then the observed signal in the frequency domain can be expressed as

$$Y(e^{jw}) = X(e^{jw}) + N(e^{jw}) \tag{2.1}$$

If $s(t)$ is the original clean speech signal, the received speech signal $y(t)$ in time domain can be represented as

$$y(t) = s(t) * h(t) + n(t) = x(t) + n(t) \tag{2.2}$$

where $h(t)$ is the impulse response of channel distortion and $n(t)$ the ambient noise. (*) denotes the convolution operation, and $x(t)$ the noise-free speech as shown in the figure XXX. Typical structural models for adaptation to variability assume that speech is corrupted by a combination of additive noise and linear filtering.

In speech processing, the speech is considered as useful data and all other signals are assumed to be noise. Many algorithms and applications are created to reduce or eliminate noise from signals, such as Voice Activity Detector.

2.4 Choice of VAD's Variance Based

A VAD's Variance Based is an early approach to VAD algorithms as others like short-term energy, zero-crossing rate and LPC coefficients. All of them can be used for speech detection even if more recent metrics can also be used (like cepstral features, formant shape and least-square periodicity measures).

The short-time energy or spectral energy has been conventionally used as the major feature parameters to distinguish the speech segments from other waveforms. However, these features become less reliable and robust in noisy environments, especially in the presence of non-stationary noise and sound artifacts such as lip snacks, heavy breathing and mouth clicks, etc.

HOS has shown good results in a number of signal processing applications and are of particular value when dealing with a mixture of Gaussian and non-Gaussian processes and system nonlinearity. Its application in speech processing is Gaussian suppression and phrase preservation properties.

Chapter 3

Blind Source Separation

We all know the problem: we are in a party, people are talking and it is quite hard to understand each other because of all the interference. You can imagine, it is even harder for a machine to separate individual speeches. This is the well known problem often referred to as the ”Cocktail Party Problem”.

The solution to these kinds of problems is called ”Blind Source Separation” (BSS). Blind source separation attempts, as the name states, to separate a mixture of signals into their different sources. The word ”blind” is used because we have no prior knowledge about the statistics of the source in general.

In this project we take an information theoretical approach and make use of the recently popularized statistical method called the independent component analysis. Our inspiration for this choice is of course the success of human brain to solve the problem. It has been hypothesized that brain is a ultimately a sophisticated statistical engine, where thought is modeled by statistical inference rather than logic and learning results from accumulation of massive date from interactions with the world. So a statistical method that claims to model information decomposition and encoding in the brain is certainly worthy of examination.

3.1 Different Approaches for BSS

There exist different approaches for ”Blind Source Separation”:

- Bayesian Approach. Basic idea: forming a model that describes a particular

source separation problem. The result is a mixing matrix. This approach can lead to algorithms known from ICA.

- TDSEP (Temporal Decorrelation source separation) nd It uses the temporal structure of signals in order to compute the time-delayed 2 order correlation for the source separation. The best results are achieved if the autocorrelations are as different as possible. The goal is to minimize the cost function:

$$\ell(C_{ij}) = \sum_{i \neq j} [y_i(t) y_j(t)]^2 + \sum_k \sum_{i \neq j} [y_i(t) y_j(t + \tau_k)]^2$$

  (y will be pre-whitened). Algorithm makes a rotation in order to simultaneously diagonalize the set of time-lagged correlation matrices. This algorithm sometimes delivers better results than ICA, especially with respect to Gaussian signals. Compared to ICA, it is computationally less expensive.

- Blind Separation of disjoint orthogonal signals. It uses only 2 mixtures of N sources, but the sources have to be pair-wise disjointly orthogonal. The algorithms are based on the Short Time Fourier Transform.

- Principal component analysis (PCA). Use of second-order methods in order to reconstruct the signal in the mean-square error sense. The results are independent in the second order statistics. In some areas, this is called KL-transform. PCA basis vectors are mutually orthogonal.

- Independent Component Analysis (ICA)

3.2 What is difficult in BSS

Before starting to discuss measures that indicate the degree of separation achieved we will discuss what conditions could increase the difficulty of a BSS task. These conditions will thus be candidates for parameters to vary when constructing the test cases.

Convolutive mixing of the sources is inherent in al most all imaginable audio and acoustic BSS applications. In addition, we also enumerate some aspects that are related to instantaneous mixing. As the whole paper, the enumerated conditions are geared towards audio situations:

1. The closer the mixing is to a singular matrix the harder the separation task is for algorithms that do not exhibit the equivariant behavior. In the presence of noise the task becomes harder also for equivariant algorithms. The level of difficulty can be controlled by adjusting the eigen value spread of the mixing filter matrix.

2. There is a continuum from instantaneous mixing to delayed mixing, i.e. convolutive mixing with only one nonzero coefficient per filter. This can be used to measure the ability of an algorithm to deal with simple convolutive mixing.

3. There is also a continuum from delayed mixing to real world convolutive mixing, which can be explored by changing the sparseness and the du ration of the mixing filters. This, tested, can rate an algorithm's ability to deal with increasingly complicated mixing filters. In real recordings these aspects can be controlled to some extent by changing the positions of the microphones and sources. The easiest cases are in general when the mixing matrix has strong direct paths with little crosstalk; i.e. every source is close to it's microphone. Also the acoustical characteristics of the recording room can be con trolled (anechoic vs. hard walled chamber). Introducing more reverberation makes the separation task more difficult in general.

4. In any kind of a mixing situation the probability density functions (pdf) of the sources have an effect. Usually the closer they are to Gaussians, the harder the separation gets.

5. The spectra of the sources may vary from narrow-band to wide-band which can have great influence on the performance of the algorithm. Tests should include sounds of both classes since some algorithms might rely on these qualities.

6. Some algorithms make use of the difference of the spectra of different source signals. Therefore it is useful to include test cases with distinct source spectra and test cases with similar source spectra.

7. Also in any kind of mixing the available amount of data needed to successfully learn to separate a static mixing situation characterizes how well the algorithm might perform in dynamic mixing circumstances. There is a continuum from

static mixing to rapidly varying mixing. This can be used to vary the level of difficulty when test ing an algorithm's tracking capabilities. When there is no comprehensive data set available with dynamic mixtures tracking capabilities can be judged from the convergence of the algorithm on static mixtures.

8. The ability to deal with silences is also needed, at least for static algorithms. Sections of silence from a source should not cause the algorithm to diverge. For example a case with a speaker with background noise little sections of silence should not cause a wildly different estimate so that re convergence is necessary when the speaker ap pears again.

9. Increasing the number of sources together with the number of mixtures increases the degree of difficulty significantly. For example, algorithms that work well in the 2by2 case might fail miserably in the 4by4 case. At the limit of convolved unmixing we have a 1by1 case which corresponds to blind deconvolution.

10. Keeping the number of sources fixed but vary ing the number of available mixtures can greatly influence the behavior of the algorithm. In general, at least the same number of mixtures as the number of sources is required. If there is further information available lesser number might suffice. By using more mixtures than there are sources, the capabilities of the algorithm to tolerate noise or to improve the separation performance could be characterized.

11. The amount and the quality of noise in the mixtures can be controlled using:

    - A single noise signal independent of all sources mixed to each sensor signal.
    - Different noise components, independent of all sources and each other, mixed to each sensor signal.
    - Similarity of the noise pdf/spectrum to the source signals.

The chosen BSS approach for our research is ICA and it is explained in the next chapter.

Chapter 4

Independent Component Analysis

4.1 Basics of ICA

Usually sensors record mixtures of different signals and, under certain conditions, those underneath signals can be recovered by making use of ICA [26]. ICA is a statistical technique that represents a multidimensional random vector as a linear combination of non-Gaussian random variables and is sought as a linear transformation of the original data [27]. It can be applied to Biological data, Speech, Image processing, EEG and functional magnetic resonance imaging (fMRI).

This methodology separates signals that are mixed together without possessing significant information about the nature of the signals (only relying on their statistical independence) [28]. ICA's model is as follows:

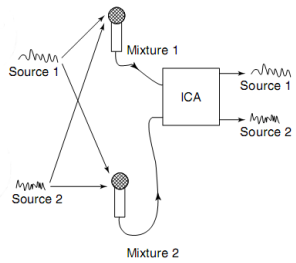


Figure 4.1: A graphical illustration of ICA for 2 sources and 2 microphones

$$X = A(t)S \tag{4.1}$$

S corresponds to the source signals, $A(t)$ is the mixing matrix and X is the recorded signals.

Basically, ICA's approach consists in finding an unmixing matrix $W(t)$ such that Y represents the estimation of S and $W(t)$ is an approximated inverse of the matrix $A(t)$ of equation (4.1). $W(t)$ is chosen so that the output signals Y are as statistically independent as possible.

$$Y = W(t)X \tag{4.2}$$

- Matrix $A(t)$ in equation (4.1) must be full rank, invertible and its columns linearly independent;
- As maximum, only one element of $S(t)$ can be gaussian.

Non-Gaussianity, motivated by the central limit theorem, is one method for measuring the independence of the sources and it can be measured, for instance, by kurtosis or negentropy [29]. Mutual information is another popular criteria for measuring statistical independence of signals. The later is not discussed in this paper.

In a certain sense ICA can be divided into two parts: an objective function plus an algorithm that maximizes it.

4.1.1 General definition

Let us consider the instantaneous case of equation equation (4.1). The task is to transform the observed data $x(t)$, using a linear transformation $y(t) = W(t)x(t)$, into maximally independent components $y(t)$ measured by some function $F(y_1, \ldots, y_n)$ of independence. The estimation of the data model of ICA is usually performed by formulating an objective function and then minimizing or maximing it [29]. Such function is called a contrast function and its optimization enables the estimation of the independent components. Thus

ICA method = objective function + algorithm optimization

4.1.2 Definition of statistical independence

Pierre Comon (1994) [30] define the concept of independence as follows. Two random variables y_1 and y_2 are said to be independent if information on y_1 does not give any information on the value of y_2 and vice versa. This independence can be defined by the probability densities. If we denote $p(y_1, y_2)$ the joint probability density function (pdf) of y_1 and y_2 and $p_1(y_1)$ the marginal pdf of y_1 (i.e. the probability of y_1 when it is considered alone), we define

$$\mathbf{p}_1(y_1) = \int p(y_1, y_2) dy_2, \tag{4.3}$$

and similarly for y_2. Then we define that y_1 and y_2 are independent, if and only if, the joint pdf is factorizable in the following way:

$$\mathbf{p}(y_1, y_2) = p_1(y_1) p_2(y_2). \tag{4.4}$$

This definition extends naturally for any number n of random variables, in which case the joint density must be a product of n terms. By deriving such definition we can obtain an important property of independent random variables. Given two functions, h_1 and h_2, we have

$$\mathbf{E}\left\{h_1(y_1) h_2(y_2)\right\} = E\left\{h_1(y_1)\right\} E\left\{h_2(y_2)\right\} \tag{4.5}$$

Aapo Hyvarinen and Erkki Oja (2000) [31] assure that the key to estimate the ICA model is nongaussianity. Without nongaussianity the estimation is not possible at all. Is by maximizing nongaussianity that we find the independent components. The optimization landscape for nongaussianity in the n-dimensional space of vectors w has $2n$ local maxima, two for each independent component. To find several independent components, we need to find all these local maxima [31].

4.2 Contrast Functions for ICA

To use nongaussianity in ICA estimation, we must have a quantitative measure of nongaussianity of a random variable y (with zero mean and variance equal to one). For measuring nongaussianity we have two popular functions:

- Kurtosis
- Negentropy

Kurtosis

The classical measure of nongaussianity is kurtosis or the fourth-order cumulant. The kurtosis of y is classically defined by

$$\mathbf{k}urt(y) = E\left\{y^4\right\} - 3\left(E\left\{y^2\right\}\right)^2 \tag{4.6}$$

Kurtosis is zero for a gaussian random variable and for almost all nongaussian random variables, kurtosis is nonzero [29]. Kurtosis can be both positive or negative. Random variables that have a negative kurtosis are called subgaussian (usually with a flat pdf), and those with positive kurtosis are called supergaussian (these ones usually have a spiky pdf with heavy tails). Typically nongaussianity is measured by the absolute value of kurtosis and has been widely used as a measure of nongaussianity in ICA and related fields. It is simple, both computationally and theoretically. Computationally, kurtosis can be estimated simply by using the fourth moment of the sample data [31]. Theoretical analysis is simple because of the following linearity property. If x_1 and x_2 are two independent random variables, it holds $kurt\left(x_1 + x_2\right) = kurt\left(x_1\right) + kurt\left(x_2\right)$ and $kurt\left(\alpha x_1\right) = \alpha^4 kurt\left(x_1\right)$, where α is a scalar

example in [31]

Let us look at a 2-dimensional model $x = As$. Assume that the independent components s_1 and s_2 have kurtosis values $kurt\left(s_1\right))$ and $kurt\left(s_2\right))$, both different from zero. We seek for one of the independent components as $y = W^T x$. Let us make the transformation $z = A^T x$. Then we have $y = w^T x = w^T As = z^T s = z_1 s_1 + z_2 s_2$. Thus $kurt\left(y\right) = kurt\left(z_1 s_1\right) + kurt\left(z_2 s_2\right) = {z_1}^4 kurt\left(s_1\right) + {z_2}^4 kurt\left(s_2\right)$. As the variance of y is equal to 1, this implies a constraint on z and $E\left\{y^2\right\} = {z_1}^2 + {z_2}^2 = 1$. Geometrically, this means that vector z is constrained to the unit circle on the 2-dimensional plane. The optimization problem now is to find the maxima of the function $\left\|kurt\left(y\right)\right\| = \left|{z_1}^4 kurt\left(s_1\right) + {z_2}^4 kurt\left(s_2\right)\right|$ on the unit circle. The maxima are at the points when exactly one of the elements of the vector z is zero and the other nonzero; because of the unit circle constraint, the nonzero element must be equal to 1 or -1. But these points are exactly the ones when y equals one of the independent components and the problem is solved.

However following Hyvarinen (2000) kurtosis is not a robust measure of non-gaussianity because kurtosis is very sensitive to outliers and if its value depends on only a few observations in the tails of the dsitribution, then it may be erroneous or irrelevant.

Negentropy

A second very important measure of nongaussianity is given by negentropy. It is based on the information theoretic quantity of differential entropy.

Entropy is the basic concept of information theory. The entropy of a random variable can be interpreted as the degree of information that the observation of the variable gives. The more random, i.e., unpredictable that the variable is, the larger its entropy. The differential entropy H of a random vector y with density $f(y)$ is defined as:

$$\mathbf{H}(Y) = -\int f(y) \, log f(y) \, dy \tag{4.7}$$

A fundamental result of information theory is that a gaussian variable has the largest entropy among all random variables of equal variance. This means that entropy could be used as a measure of nongaussianity [31].

To obtain a measure of nongaussianity that is zero for a gaussian variable and always nonnegative, one uses a modified version of the definition of differential entropy, called negentropy. Negentropy J is defined as follows:

$$\mathbf{J}(y) = H(y_{gauss}) - H(y) \tag{4.8}$$

A problem with negentropy consists of the difficulties that arise to compute it [29].

Previously we introduced objective functions for the ICA estimation. But we also need an algorithm for its implementation (maximizing the contrast function) Thus after choosing an objective function for ICA we have to optimize it [31]. Are several algorithms on the market with different characteristics but this paper just focus on the FastICA. It is as follows: It finds a unit vector w such that the projection $w^T x$ maximizes nongaussianity (measured by negentropy). Basically is a fixed-point iteration that finds the maximum of the nongaussianity of $w^T x$ [31].

4.3 ICA's algorithms

After having a metric for measuring the nongaussianity, i.e. objective functions for ICA estimation we also need an algorithm for maximizing the contrast function. In the next subsections, we introduce some methods of maximization suited for this task.

Typical algorithms for ICA use centering, whitening and dimensionality reduction as preprocessing steps in order to simplify and reduce the complexity of the problem for the actual iterative algorithm. Whitening and dimension reduction can be achieved with Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) [31]. Popular algorithms for ICA include Infomax, FastICA (that operates in the time-domain), Kernel ICA and JADE.

4.3.1 FastICA

The fast fixed point algorithm (FastICA) is a computational efficient method for performing the estimation of ICA, e.g., separates linearly mixed independent source signals and it was presented by Hyvärinen and Oja [32]. It is a fixed-point iteration algorithm and it has been found by independent experiments to be 10-100 times faster than conventional gradient descent methods for ICA. Another advantage of the FastICA algorithm is that it can be used to perform projection pursuit as well, thus providing a general-purpose data analysis method that can be used both in an exploratory fashion and for estimation of independent components (individual signals) [27].

FastICA can be used for finding a unit vector w such that the projection $w^T x$ maximizes the non-Gaussianity (measured by negentropy) of $w^T x$ [33].

Let us denote the function g by the derivative of a non-quadratic nonlinearity. Then the basic form of FastICA is as follows [33]:

1. Choose a initial (e.g. random) weight vector w
2. Let $w^+ = E\left\{xg\left(w^T x\right)\right\} - E\left\{g'\left(w^T x\right)\right\} w$
3. Let $w = w^+ / \|w^+\|$
4. If not converged, go back to 2

Note: Here convergence means that the old and new values of w point in the same direction.

Properties of the FastICA Algorithm

The FastICA algorithm and the underlying contrast functions have a number of desirable properties when compared with existing methods for ICA.

- The convergence is cubic (or at least quadratic), under the assumption of the ICA data model. This is in contrast to ordinary ICA algorithms based on (stochastic) gradient descent methods, where the convergence is only linear. This means a very fast convergence, as has been confirmed by simulations and experiments on real data.
- Contrary to gradient-based algorithms, there are no step size parameters to choose. This means that the algorithm is easy to use.
- The algorithm finds directly independent components of (practically) any non-Gaussian distribution using any nonlinearity g. This is in contrast to many algorithms, where some estimate of the probability distribution function has to be first available, and the nonlinearity must be chosen accordingly.
- The performance of the method can be optimized by choosing a suitable nonlinearity g. In particular, one can obtain algorithms that are robust and/or of minimum variance. In fact, the two nonlinearities in have some optimal properties; for details see.
- The independent components can be estimated one by one, which is roughly equivalent to doing projection pursuit. This es useful in exploratory data analysis, and decreases the computational load of the method in cases where only some of the independent components need to be estimated.
- The FastICA has most of the advantages of neural algorithms: It is parallel, distributed, computationally simple, and requires little memory space. Stochastic gradient methods seem to be preferable only if fast adaptivity in a changing environment is required.

A Matlab implementation of the FastICA algorithm is available on the World Wide Web free of charge.

4.3.2 Kernel ICA

Kernel ICA works by defining a contrast function on reproducing kernel Hilbert space. This essentially means that mixtures (or observations) are projected to a higher dimensional space (even infinite dimensional space) and then we try to find the mixing matrix such that pair wise correlations are minimized in this space because once this is achieved it can be proven that for reproducing kernel Hilbert spaces based on Gaussian kernels this ensures that the sources are independent.

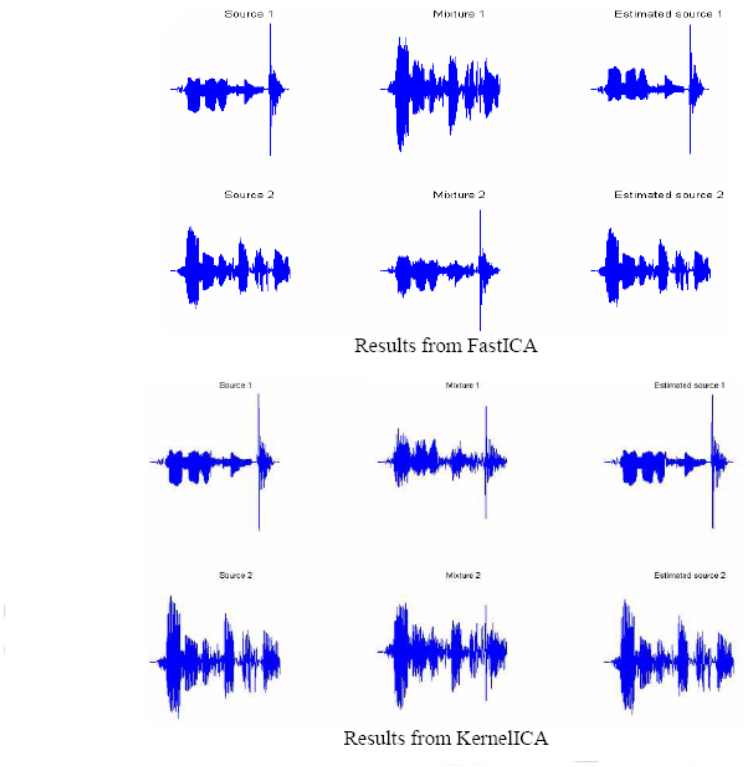


Figure 4.2: Preliminary results. The first column shows the two sources, the next column shows the mixture and the last column shows the separated channels.

4.3.3 JADE

JADE ICA was presented by Jean Francois Cardoso [34] and this subsection outlines its main characteristics. Basically JADE can se seen as a statisticbased technique and summarized as follows:

- *Initialization.* Estimate a whitening matrix W and set Z = WX.
- *Form statistics.* Estimate a maximal set Q Z i of cumulant matrices.
- *Optimize an orthogonal contrast.* Find the rotation matrix V such that the cumulant matrices are as diagonal as possible, that is, solve V = arg min P i Off(V Q Z i V).
- *Separate.* Estimate A as A = V W -1 and/or estimate the components as S = A -1 X = V Z.

This is a Jacobi algorithm because the joint diagonalizer at step 3 is found by a Jacobi technique. However, the plane rotations are applied not to the data (which are summarized in the cumulant matrices) but to the cumulant matrices themselves; the algorithm updates not data but matrixvalued statistics of the data. As with MaxKurt, the Givens angle at each step can be computed in closed form even in the case of possibly complex matrices.

A key issue is the selection of the cumulant matrices to be involved in the estimation. As explained in section 3.2, the joint diagonalization criterion P i Off(V + Q Z i V) is made identical to the contrast function, equation 3.11, by using a maximal set of cumulant matrices. This is a bit surprising but very fortunate. We do not know of any other way for a priori selecting cumulant matrices that would offer such a property (but see the next section). In any case, it guarantees equivariant estimates because the algorithm, al though operating on statistics of the sphered data, also optimizes implicitly a function of Y = V Z only.

Before proceeding, we note that true cumulant matrices can be exactly jointly diagonalized when the model holds, but this is no longer the case whenwe process real data. First, only sample statistics are available; second, the model X = AS with independent entries in S cannot be expected to hold accurately in general. This is another reason that it is important to select cumulant matrices such that P i Off(V Q Z i V) is a contrast function. In this case, the impossibility of an exact joint

diagonalization corresponds to the impossibility of finding Y = BX with independent entries. Making a maximal set of cumulant matrices as diagonal as possible coincides with making the entries of Y as independent as possible as measured by (the sample version of) criterion 3.11.

There are several options for estimating a maximal set of cumulant matrices. Recall that such a set is defined as Q Z (M i)—i = 1, n 2 where M i i = 1, n 2 is any basis for the n 2 dimensional linear space of nn matrices. A canonical basis for this space is e p e q —1 p, q n, where e p is a column vector with a 1 in pth position and 0's elsewhere. It is readily checked that [Q Z (e p e q)] ij = Cum(Z i , Z j , Z p , Z q). (4.6)

In other words, the entries of the cumulant matrices for the canonical basis are just the cumulants of Z.Abetter choice is to consider a symmetric/skew symmetric basis. Denote M pq an n n matrix defined as follows: M pq = e p e p if p = q, M pq = 2 -1/2 (e p e q +e q e p) if p ¡ qand M pq = 2 -1/2 (e p e q -e q e p) if p ¿ q. This is an orthonormal basis ofR nn . Wenote that because of the symmetries of the cumulants Q Z (e p e + q) = Q Z (e q e + p) so that Q Z (M pq) = 2 -1/2 Q Z (e p e + q) if p ¡ q and Q Z (M pq) = 0 if p ¿ q. It follows that the cumulant matrices Q Z (M pq) for p ¿ q do not even need to be computed. Being identically zero, they do not enter in the joint diagonalization criterion. It is therefore sufficient to estimate and to diagonalize n+n(n-1)/2(symmetric) cumulant matrices.

There is another idea to reduce the sizeof the statistics needed to represent exhaustively the fourthorder information. It is, however, applicable only when the model X = AS holds. In this case, the cumulant matrices do have the structure shown at equation 3.18, and their sample estimates are close to it for large enough T. Then the linear mapping M Q Z (M) has rank n (more precisely, its rank is equal to the number of components with nonzero kurtosis) because there are n linear degrees of freedom for matrices in the form UU +, namely, the n diagonal entries of . From this fact and from the symmetries of the cumulants, it follows that it exists n eigenmatrices E 1 , . . . , E n , which are orthonormal, and satisfies Q Z (E i) = i E i where the scalar is the corresponding eigenvector. These matrices E 1 , . . . , E n span the range of the mapping M Q Z (M), and any matrix M orthogonal to them is in the kernel, that is, Q Z (M) = 0. This shows that all the information contained in Q Z can be summarized by the n eigenmatrices associated with the n nonzero eigenvalues. By inserting M = u i u i in the expressions 3.18 and using the

orthonormality of the columns of U (that is, u i u j = ij), it is readily checked that a set of eigenmatrices is E i = u i u i .

The JADE algorithm was originally introduced as performing ICA by a joint approximate diagonalization of eigenmatrices in Cardoso and Souloumiac (1993),where we advocated the joint diagonalization of only the n most significant eigenmatrices of Q Z as a device to reduce the computational load (even though the eigenmatrices are obtained at the extra cost of the eigen decomposition of an n 2 n 2 array containing all the fourthorder cumulants). The number of statistics is reduced from n 4 cumulants or n(n + 1)/2 symmetric cumulant matrices of size nn to a set of n eigenmatrices of size nn. Such a reduction is achieved at no statistical loss (at least for large T) only when the model holds. Therefore, we do not recommend reduction to eigen matrices when processing data sets for which it is not clear a priori whether the model X = AS actually holds to good accuracy. We still refer to JADE as the process of jointly diagonalizing a maximal set of cumulant matrices, even when it is not further reduced to the n most significant eigenmatrices. It should also be pointed out that the device of truncating the full cumulant set by reduction to the most significant matrices is expected to destroy the equivariance property when the model does not hold.

Chapter 5

Proposed System

This research presents an innovative system for adaptive speech denoising using ICA and Voice Activity Detection (VAD) and it is described in fig. 5.1.

Designed for instantaneous speech mixtures with two sources and two microphones the proposed system identifies the noise contained in each noisy mixture, applies the most convenient block ICA method among 3 methods (JADE, KERNEL ICA and FastICA) and, after source separation, identifies the estimated speech signal. The ICA suitability is in accordance with the detected noise.

The mixing process is non-linear and the information extracted on the first stage can be used for later post-processing and further system extension with significant potential over the conventional approaches. The process is completely automatic from the source recording to its output and such system has a wide range of applications. This approach and its experimental data significantly provides auspicious potential over the systems currently available.

5.1 Setup

As it can be observed the system considers 2 sources and 2 microphones. The source signals are 3 seconds in length and one of them is a speech signal extracted from TIMIT and the other is a noise signal from NOISEX (it can be White Noise, Pink Noise, HF Channel Noise, Car Noise, Tank Noise, Factory Noise, etc.). All signals are sampled with 16KHz sampling frequency. Our mixing process simulates fig. 5.2 where the sources' properties change through time. The mixing

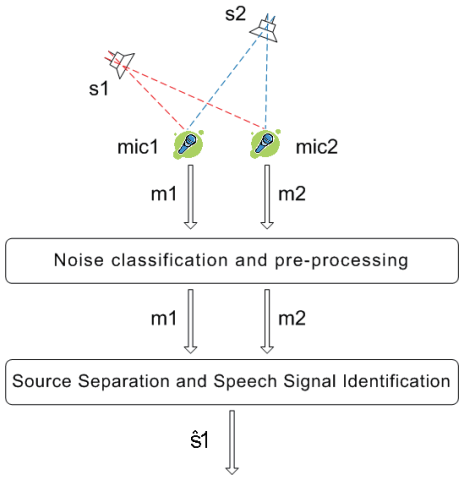


Figure 5.1: The proposed system

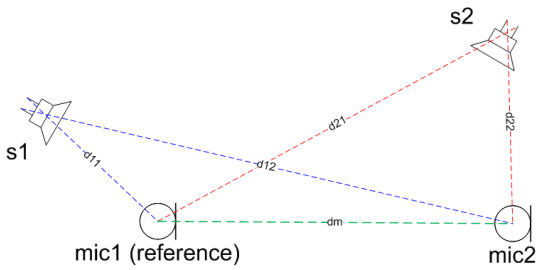


Figure 5.2: The proposed model

process follows eq. 5.1 where S corresponds to the source signals, A_α are the mixing matrices and m^1 and m^2 are the noisy mixtures (with respective mixing matrices A_2 and A_1). The sources' contribution to each microphone evolve through time alternating every 400ms. In a formal way:

$$\begin{pmatrix} m^1 \\ m^2 \end{pmatrix} = A_\alpha \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \wedge A_\alpha = \begin{pmatrix} 1 & \frac{d_{11}}{d_{21}} \\ \frac{d_{11}}{d_{12}} & \frac{d_{11}}{d_{22}} \end{pmatrix} \tag{5.1}$$

and

$$A_1 = \begin{pmatrix} 5.4 & 0.5 \\ 2.4 & 0.4 \end{pmatrix} \wedge A_2 = \begin{pmatrix} 4.5 & 0.2 \\ 3.4 & 0.002 \end{pmatrix}$$

and the noisy mixtures are:

$$X = \begin{cases} m^1 \\ m^2 \end{cases} = \begin{cases} {m^1}_1, {m^1}_2, ..., {m^1}_n \\ {m^2}_1, {m^2}_2, ..., {m^2}_n \end{cases}$$

$${m^1}_i = \begin{cases} A_1 S_i & \text{for } (400j - 399) \leq i \leq 400j, \quad j = 1, 3, ..., 2p + 1 \\ A_2 S_i & \text{for } (400j - 399) \leq i \leq 400j, \quad j = 1, 2, ..., 2p \end{cases}$$

and

$${m^2}_i = \begin{cases} A_2 S_i & \text{for } (400j - 399) \leq i \leq 400j, \quad j = 1, 3, ..., 2p + 1 \\ A_1 S_i & \text{for } (400j - 399) \leq i \leq 400j, \quad j = 1, 2, ..., 2p \end{cases}$$

Where $(2p + 1) \leq n$, for $i, j, p \in \mathbb{N}$.

5.2 Implementation

As it can be observed in fig. 5.2 the first step is the identification of the noise contained in the noisy mixtures and, after that, performing source separation based on ICA and speech signal identification. This approach paves the way for further post processing before and after ICA's application.

The main novelty of this approach is that extracts information that can be used for additional system module implementation. Such information allow us to implement effective speech enhancement for the aimed goal.

5.2.1 Detection of the noisy Frames using VAD's Variance Based

The noisy sources recorded by the microphones may contain a wide-band noise in the form of a white or colored noise, a periodic signal such as in hum noise, room reverberations or it can take the form of fading noise. The first two examples represent additive noise sources, while the other two examples represent convolutional and multiplicative noise sources, respectively [5]. The system presented by this paper refers to the first two examples and other types that share common features (all files contained by NOISEX).

Sometimes *à priori* information about the signals is not available but knowing such information can be extremely valuable for some pre and/or post-processing.

With this aim and based on the proposed model of fig. 5.2 a VAD was made and furthermore optimized for efficient recognition. The goal is to identify the speechless components of a noisy mixture. Concatenating all detected noise-dominant frames allows the extraction of its spectrum and its comparation with a system built database of template noises based on NOISEX. Following such comparation, the nature of the noise can be robustly asserted and is possible to apply automatic processing that suits such type of noise.

5.2.2 Estimation of the spectrum of the speechless detected components

The Voice Activity Detection that was built is as follows: the noisy speech is segmented into 42 ms frames. For each frame its variance is calculated and all of them are sorted by crescent order. Such vector provides us with an reference value in the form of the mean of the first 60 values. Such reference value is used for adaptive thresholding aiming at frame classification as either signal or noise-dominant frame. All speechless frames are merged together, its spectrum calculated and compared by Euclidean distance with the spectrums of the noise database.

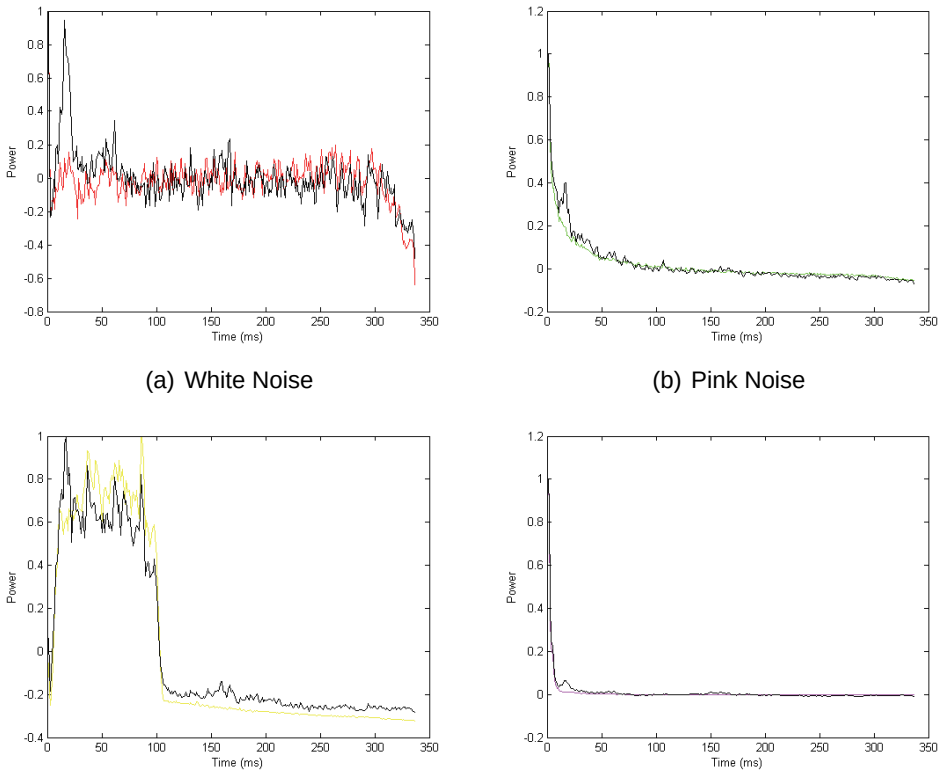
5.2.3 Noise identification

For the considered setup this approach allows the identification of the noise contained in the noisy mixture. Consequently, it is possible to know the noise's behavior and to perform effective speech denoising or adaptive use of ICA.

On figure 5.3, for some noise types, it is possible to analyze the performance of the VAD (variance based) that is made for this research. This VAD works efficiently for all noise types contained by NOISEX.

5.2.4 Source Separation - ICA based

After identifying the noise contained in the noisy mixture, the noise signal and the speech signal are separated by applying block ICA. The ICA method used can be FastICA, Kernel ICA or JADE. Such choice is up to the noise that is detected in the noisy mixture and follows table 6.1.



(a) White Noise

(b) Pink Noise

(c) HF channel Noise

(d) Car Noise

Figure 5.3: Spectrum of the Speechless Detected Components (black), White Noise (red), Pink Noise (green), HF channel Noise (yellow) and Car Noise (magenta)'

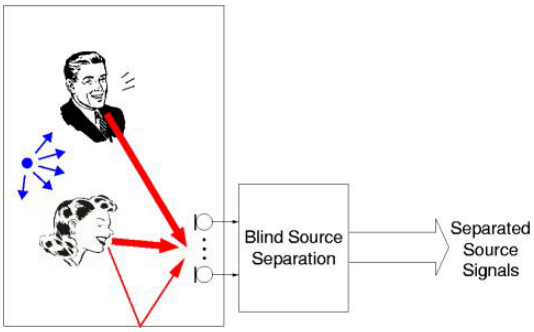


Figure 5.4: ICA's application

Table 5.1: ICA suitability per type of noise

	WN	**PN**	**HF**	**Car**	**Factory1**
KERNEL ICA	73%	19%	0%	86%	78%
FAST ICA	27%	53%	76%	14%	22%
JADE	0%	28%	24%	0%	0%

5.2.5 Speech Signal Identification

With the information extracted in the first stage it is possible to identify the estimated noise signal and consequent identification of the speech signal. For this goal is only necessary to compare the spectrum of the speechless detected components with ICA's output. Such comparation is possible by using a metric as Euclidean distance.

Note: The Euclidean distance between points $(p_1, p_2, ..., p_n)$ and $(q_1, q_2, ..., q_n)$ in Euclidean n-space, is defined as:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

5.3 Analysis of the results

Through extensive experiments it come out that the VAD that was built provides an accurate and reliable output (see fig 5.3). The major problem in this algorithm is that it is not robust if the speech signal is simultaneously attacked by more than one noise source and this still stands as a challenge that it is going to be taken

into account in further extensions of the proposed system.

Decurrent from the way the reference value is obtained, not all speechless frames are correctly identified as noise dominant. However, this does not interfere with the current goal for the VAD in the proposed system.

The input dB-SNR is 10dB and 5dB for mixture 1 and mixture 2, respectively and the Source Separation is made by applying FastICA for the separation of a speech signal (that is the phrase *"Ducks have colorful feet and white feathers"*) from a noise signal extracted from NOISEX. For different block FastICA length values values the achieved dB-SNR improvement can be seen at table 5.2. The best output db-SNR are for divisors (200 and 400 ms) of the chosen mixing length interval (400 ms) and show enough potential data for further developments. The SNR is measured by Average Segmental SNR and those values are obtained by running FastICA method 16 times, ignoring the tail values (the first and last two values) and calculating their mean.

A problem that needs to be solved is finding a method of detecting automatically where the mixing process changes. Knowing those points means that it is possible to ignored them for evaluation purposes and for defining block ICA's length frame.

In fig. 5.5 it is possible to compare the system ICA's output and input.

Table 5.2: dB-SNR improvement for several noisy mixtures (for these cases the speech signal is corrupted with a noise signal and WN, PK, HF and Car means White Noise, Pink Noise, HF Channel Noise and Car Noise, respectively) and different block ICA frame lengths.

	WN	**PN**	**HF**	**Car**
135 ms	15.85	14.80	16.31	11.13
200 ms	27.04	21.34	22.55	19.19
400 ms	29.07	29.58	32.31	18.69

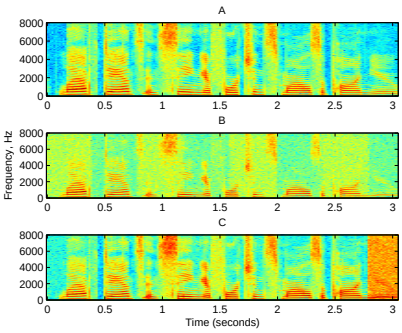


Figure 5.5: Spectrogram of a) clean speech, b) noisy speech corrupted with White Noise at 10dB SNR, c) the estimated speech signal after ICA.

Chapter 6

Conclusion

The core system was developed and its main current advantages are the blind extraction of information about the sources that can be used for adaptive denoising and ICA's automatic suitability, and for improving the ICA's output-SNR. The proposed method is also suitable for further post-processing based on the information extracted on the first stage.

The ICA's suitability is accordingly to the noise that is detected in the noisy mixture and follows table 6.1.

Table 6.1: ICA suitability per type of noise

	WN	**PN**	**HF**	**Car**	**Factory1**
KERNEL ICA	73%	19%	0%	86%	78%
FAST ICA	27%	53%	76%	14%	22%
JADE	0%	28%	24%	0%	0%

Future work aims at automatic identification of the points where the mixing process changes; system robustness for the occurrence of simultaneous noise types and ICA's modification for the cases when the sources/microphones are too close to each other.

Bibliography

[1] Xuedong Huang and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001. Foreword By-Raj Reddy.

[2] Eric M. Visser and Te-Won Lee. Blind source separation in mobile environments using a priori knowledge. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 3, pages iii–893–6vol.3, 17-21 May 2004.

[3] John R. Deller, John G. Proakis, and John H. Hansen. *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1993.

[4] Erik M. Visser and Te-Won Lee. Speech enhancement using blind source separation and two-channel energy based speaker detection. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, pages I–884–I–887vol.1, 6-10 April 2003.

[5] Yariv Ephraim, Hanoch Lev-Ari, and William J.J. Roberts. *A brief survey of Speech Enhancement*. Electronic Engineering Handbook, CRC PRESS, 2005.

[6] E. Visser, M. Otsuka, and Te-Won Lee. A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. *Speech Communication*, 41:393–407, 2003.

[7] J. Droppo, L. Deng, and A. Acero. Evaluation of the splice algorithm on the

aurora 2 database. In *Proc. (Eurospeech) '2001', Aalborg, Denmark*, pages 217–220, 2001.

[8] M. Lieb and A. Fischer. Experiments with the philips continuous asr system on the aurora noisy digits database. In *Proc. (Eurospeech) '2001', Aalborg, Denmark*, pages 625–628, 2001.

[9] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean-square error short-time spectral amplitude estimator. In *IEEE Transcriptions Acoustical Speech Signal Processing ASSP-32*, pages 1109–1121, 1984.

[10] H. Hermansky and N. Morgan. Rasta processing of speech. In *IEEE Transcriptions Speech Audio Processing ASSP-32*, pages 578–589, 1994.

[11] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J.Acoust.Soc.of America*, 55:1304–1312, 1974.

[12] M. Brandstein and H. Silverman. A practical methodology for speech source localization with microphone arrays. *Computational Speech Language*, 11 (2):91–126, 1997.

[13] D. Johnson and D. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice Hall, Englewood Cliffs, 2nd edition, 1993.

[14] M. Dahl and I. Claesson. Acoustic noise and echo cancelling with microphone array. In *IEEE Transcriptions Veh. Technological '1999'*, volume 48 (5), pages 1518–1526, 1999.

[15] D. Ward, R. Williamson, and R. Kennedy. Broadband microphone arrays for speech acquisition. *Acoustical Aust.*, 26 (1):17–20, 1998.

[16] S. Fisher and K. Simmer. Beamforming microphone arrays for speech acquisition in noisy environments. *Speech Communication*, 20 (3-4):215–227, 1996.

[17] H. Attias, J. Platt, A. Acero, and L. Deng. Speech denoising and dereverberation using probabilistic models, 2001.

[18] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. In *IEEE Transcriptions Speech Audio Processing 8*, pages 320–327, 2000.

[19] A. Bell and T. Seijnowski. An information-maximisation approach to blind separation and blind deconvolution. In *Neural Computational 7 (6)*, pages 1004–1034, 1995.

[20] Te-won Lee, A. Bell, and Lambert R. Blind separation of delayed and convolved sources. *Advances in Neural Information Processing Systems*, 9:758–764, 1997.

[21] Simon Haykin and Zhe Chen. The cocktail party problem, 2005.

[22] António Rui Ferreira Rebordao, M. K. Islam Molla, Keikichi Hirose, and Minematsu Nobuaki. A speech denoising system based on ica and voice activity detection. In *Proc. of Acoustical Society of Japan (ASJ 2008 Spring Meeting), Chiba, Japan*, 17-19 March 2008.

[23] António Rui Ferreira Rebordao, M. K. Islam Molla, Keikichi Hirose, and Minematsu Nobuaki. An adaptive speech denoising system based on ica and voice activity detection. In *Proc. of International Workshop on Nonlinear Circuits and Signal Processing (NCSP08)*.

[24] António Rui Ferreira Rebordao, M. K. Islam Molla, Keikichi Hirose, and Minematsu Nobuaki. Adaptive ica usage for signal enhancement. In *International Conference on Audio, Language and Image Processing 2008 (ICALIP 2008), Shangai, China*, 7-9 July 2008.

[25] Michael Yaw Appiah, Raimonda Makrickaite, Milda Gusaite, and Sasikanth Munagala. Robust voice activity detection and noise reduction mechanism using higher-order statistics. The University of Aalborg, 2005.

[26] James V. Stone. *Independent Component Analysis: A Tutorial Introduction*. MIT Press, Cambridge, MA, USA, 2004.

[27] A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4-5):411–430, 2000.

[28] Savyasachi Singh and Kumar Ritwik. Blind source speech separation. *EEL 6825 Pattern Recognition*, 2006.

[29] Aapo Hyvarinen. Survey on independent component analysis.

[30] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[31] Aapo Hyvarinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13 (4-5):411–430, 2000.

[32] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural Comput.*, 9(7):1483–1492, 1997.

[33] Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley-Interscience, May 2001.

[34] Jean-François Cardoso. High-order contrasts for independent component analysis. *Neural Comput.*, 11(1):157–192, 1999.

[35] E. Visser and Te-Won Lee. Speech enhancement using blind source separation and two-channel energy based speaker detection. In *Proc. (ICASSP) '2003'*, pages 884–887, 2003.

[36] E. Visser and Te-Won Lee. Blind source separation in mobile environments using a priori knowledge. In *Proc. (ICASSP) '2004'*, volume III, pages 893–896, 2004.

[37] X. Huang, A. Acero, and Hsiau-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, New Jersey, 2nd edition, 2001.

[38] Jean-Franois Cardoso. Blind signal separation: statistical principles. In *Proc. (IEEE) '1998'*, volume 9 (10), pages 2009–2025, October 1998.

[39] Te-won Lee and Yao Kaishen. Speech enhancement by perceptual filter with sequential noise parameter estimation. In *Proc. (ICASSP) '2004'*, volume I, pages 693–696, 2004.

[40] Savyasachi Singh and Kumar Ritwik. Blind source speech separation. *EEL 6825 Pattern Recognition*, 2006.

[41] Zheng Yongrui, Lin Qiuhua, Yin Fuliang, and Liang Hualou. Speech enhancement using ica with emd-based reference, 2006.

[42] Ephraim Yariv and Cohen Israel. Recent advancements in speech enhancement, 2004.

[43] Javier Ortega-Garcia and Joaquin Gonzalez-Rodriguez. Overview of speech enhancement techniques for automatic speaker recognition, 1999.

[44] Michael Casey. Separation of mixed audio sources by independent subspace analysis, 2001.

[45] N. Huang. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *journal Real Society London*, 454:903–995, 1998.

[46] Li-Ran Shen, Xue-Yao Li, Qing-Bo Yin, and Hui-Qiang Wang. Speech enhancement in short-wave channel based on ica in empirical mode decomposition domain, 2006.

[47] Steven Boll. Supression of acoustic noise in speech using spectral subtraction. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume ASSP-27(2), pages 113–120, 1979.

[48] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. (ICASSP) '1979'*, pages 208–211, 1979.

[49] F. Bach and M. Jordan. Kernel independent component analysis, 2001.

[50] Jounghoon Beh, R.H. Baran, and Hanseok Ko. Dual channel based speech enhancement using novelty filter for robust speech recognition in automobile environments. In *Consumer Electronics, 2006. ICCE '06. 2006 Digest of Technical Papers. International Conference on*, pages 243–244, 7-11 Jan. 2006.

[51] E. Bingham. A fast fixedpoint algorithm for independent component analysis of complex valued signals.

[52] Jean-Francois Cardoso. Blind signal separation: statistical principles. In *Proc. (IEEE) '1998'*, volume 9 (10), pages 2009–2025, October 1998.

[53] Michael A. Casey and Alex Westner. Separation of mixed audio sources by independent subspace analysis, 2000.

[54] Pierre Comon. Independent component analysis, a new concept? *Signal Process.*, 36(3):287–314, 1994.

[55] Simon Haykin and Zhe Chen. The cocktail party problem, 2005.

[56] A. Hyvrinen. Survey on independent component analysis, 1999.

[57] Taesu Kim, H. Attias, Soo-Young Lee, and Te-Won Lee. Robust time delay estimation in noisy reverberant environments with a probabilistic graphical model. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 5, pages v/525–v/528Vol.5, 18-23 March 2005.

[58] Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee. Speech feature extraction using independent component analysis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1631–1634vol.3, 5-9 June 2000.

[59] Te-Won Lee, B.-U. Koehler, and R. Orglmeister. Blind source separation of nonlinear mixing models. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 406–415, 24-26 Sept. 1997.

[60] Te-Won Lee, A. Ziehe, R. Orglmeister, and T. Sejnowski. Combining time-delayed decorrelation and ica: towards solving the cocktail party problem. In *Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 1249–1252vol.2, 12-15 May 1998.

[61] Steven R. Long Manli C. Wu Hsing H. Shih Quanan Zheng Nai-Chyuan Yen Chi Chao Tung Henry H. Liu Norden E. Huang, Zheng Shen. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Royal Society London Series*, A 454:903995, 1998.

[62] Hyung-Min Park, Ho-Young Jung, Soo-Young Lee, and Te-Won Lee. On subband-based blind separation for noisy speech recognition. In *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on*, volume 1, pages 204–209vol.1, 16-20 Nov. 1999.

[63] L. Parra and C. Spence. Separation of non-stationary natural signals, 2001.

[64] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. 2000.

[65] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods, 1999.

[66] Erik M. Visser, Manabu Otsuka, and Te-Won Lee. A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. *Speech Communication*, 41(2-3):393–407, 2003.