Applied Acoustics 113 (2016) 170–184

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Real-time blind source separation system with applications to distant speech recognition

Alberto E.A. Ferreira [a], Diogo Alarcão [b,*]

[a] Dept. of Physics, Técnico Lisboa, University of Lisbon, Av. Rovisco Pais, P-1049-001 Lisbon, Portugal
[b] CAPS – Técnico Lisboa, University of Lisbon, Av. Rovisco Pais 1, P-1049-001 Lisbon, Portugal

ARTICLE INFO

Article history:
Received 16 December 2015
Received in revised form 23 June 2016
Accepted 27 June 2016
Available online 4 July 2016

Keywords:
Blind source separation
Distant speech recognition
DUET

ABSTRACT

A real-time BSS system based on DUET was developed and implemented in order to assess its potential as the front-end for a DSR engine. The system uses only two closely-spaced standard omni-directional microphones and a computer soundcard and was developed for low reverberation environments with several human speakers and different noise sources.

A novel multi-source real-time audio streaming module was developed, with arbitrary statistics, movement tracking, continuity cues such as position and cross-correlation, a spurious peak classifier stage based on kurtosis, and spectral subtraction post-processing.

Two intrinsic error causes for the binaural attenuation and delay estimators were identified, due to FFT spectral leakage and to sibilants, which violate the taken for granted DUET assumptions. A comprehensive study on time windows was done and new window types proposed in order to minimize the DUET assumptions violations.

The implemented system correctly identifies the clusters in the binaural estimators' space for the case of a real room with two human speakers, up to distances of 2 m from the two microphones, although for distances greater than 1 m the separation quality quickly degrades.

1. Introduction

Speech is one of the most natural forms of human communication and that is why a great deal of research has been done to allow machines to recognize it.

As the word recognition rate of Automatic Speech Recognition (ASR) systems keeps getting better in scenarios where the speaker is close to a microphone and little noise is present [1], performing Distant Speech Recognition (DSR), where the distance between the user and the microphone(s) increases is not yet solved satisfactorily, especially in real-time implementations [2]. As the target speaker distance to the microphone grows, lower received speech power, added reverberation and noise sources from different origins, even from different speakers, are some of the factors that decrease the speech Signal-to-Noise-Ratio (SNR).

One way to solve the DSR problem is to add a pre-processing stage that would ideally isolate the target speaker from the remaining signals.

The task of separating multiple sound sources' signals one from another given one or more mixtures of the signals is ascribed to the field of Source Separation (SS), which is traditionally split into Blind Source Separation (BSS), Semi-Blind Source Separation (SBSS) and Informed Source Separation (ISS), in increasing order of the assumed information about the sound sources.

The Degenerate Unmixing Estimation Technique (DUET) [3], is a rather simple but efficient BSS method that allows the separation of any number of sources from only two mixtures registered at two microphones. Degenerate BSS, where one tries to separate more sources than there are input channels, places a great challenge because the mixing process is non-invertible.

Some approaches to BSS that are very efficient and robust were proposed in recent years, however, most are very computationally demanding and unable to run in real-time on today's commodity hardware and for the foreseeable future. Examples are the Nonnegative Matrix Factorization (NMF) [4], Model based Expectation Maximization Source Separation and Localization (MESSL) [5], Mouba [6] and Sawada [7], which use the same core principles as DUET, being more robust in reverberant environments, but relying on a much more complex model.

Another family of methods, based on Independent Component Analysis (ICA), which is traditionally used for SS, also presents several drawbacks as its separation capability responds linearly with the distance between microphones. Thus, these methods will

* Corresponding author.
E-mail address: diogo.alarcao@tecnico.ulisboa.pt (D. Alarcão).

http://dx.doi.org/10.1016/j.apacoust.2016.06.024
0003-682X/© 2016 Elsevier Ltd. All rights reserved.

only separate signals well if the microphones are considerably separated in space.

Techniques based on spectral subtraction can be additionally used in real-time [8,9] to clean the speech signals, however, supplementary dereverberation and source separation stages are required in a live environment.

This area of research has seen a great activity in recent years and some large funded projects have been carried out such as the European project DICIT, with a 5 M€ budget, and such as Samsung's Smart TV project [10–12], but only with a limited success outcome in what concerns their DSR capabilities.

2. Motivation

The motivation for this work is the creation of a front-end to a DSR system that would allow issuing voice commands to a media device (such as a TV) in a living room with more than one person present, together with various noise sources, amongst which the device itself, since it emits audio too, when on.

Besides being able to run in real-time (low latency) the system should be cheap, small, non-intrusive and reliable. Therefore, the system should be as simple as possible meaning that the speech signal enhancement could be implemented in real-time with only two closely-spaced microphones.

Another pursued goal was to design the system to handle very low SNR values as well, meaning that it is able to extract weak speech sound signals in relatively noisy environments.

3. System development

Given the available choice of methods, DUET was chosen due to its computational simplicity to allow a real-time implementation which was created in C++. An iterative spectral subtraction module was also implemented.

DUET [3], upon which the developed system was based, adopts the noisy anechoic mixing model for two microphones with the relative attenuation a_j and delay δ_j between right (x_2) and left (x_1) microphone signals for N sources:

$$\begin{aligned} x_1(t) &= \sum_{j=1}^{N} s_j(t) + n_1(t) \\ x_2(t) &= \sum_{j=1}^{N} a_j s_j(t-\delta_j) + n_2(t) \end{aligned} \tag{1}$$

where s_j is the signal of the audio source j and $n_m(t)$ represent independent Gaussian noise signals for each microphone m that for instance model electronic amplification noise.

This technique works for sources with different spatial signatures. For two microphones it means that any sources located at the same azimuth and distance from the microphone pair will not be separable.

3.1. W-disjoint orthogonality

DUET requires orthogonally W-disjoint source signals. Two functions are *W-disjoint orthogonal* if given a window function $w(t)$, the supports of the windowed Fourier transforms of the functions are disjoint. The windowed Fourier transform[1] of $s(t)$ is defined as

$$S^w(\tau,\omega) := \mathscr{F}^w[s](\tau,\omega) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} w(t-\tau)s(t)e^{-i\omega t}dt \tag{2}$$

where τ is the time offset of the analysis window w and ω the angular frequency.

The condition of W-disjoint orthogonality thus requires that:

$$\forall_{(\tau,\omega)} \forall_{j \neq k} \quad S_j^w(\tau,\omega) S_k^w(\tau,\omega) = 0 \tag{3}$$

This condition allows the creation of binary masks M_j, plain indicator functions with which a source j can be separated from the mixture, for instance, by multiplying the mask M_j with the windowed Fourier transform of the left input channel X_1^w:

$$M_j(\tau,\omega) := \begin{cases} 1, & S^w(\tau,\omega) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\forall_{(\tau,\omega)} \ \hat{S}_j^w(\tau,\omega) = M_j(\tau,\omega) X_1^w(\tau,\omega) \tag{5}$$

where $\hat{S}_j^w$ is the estimated sound signal of source j.

3.2. Local stationarity

The time shift property of the Fourier transform, true for $w(t)=1$ and approximately true for windows with finite support [13,14], states that for a signal s,

$$\begin{aligned} \mathscr{F}^w[s(t-\delta)](\tau,\omega) &= \exp(-i\omega\delta)\mathscr{F}^w[s(t)](\tau-\delta,\omega) \\ &\approx \exp(-i\omega\delta)\mathscr{F}^w[s(t)](\tau,\omega) \end{aligned} \tag{6}$$

where the time shift δ of the transformed function is assumed negligible. This means that the solutions in the histograms (see Section 3.6) are not exact and will vary according to the time shift.

DUET requires Eq. (6) to hold, even if w has finite support up to a maximum delay Δ/c between sensor signals, *id est*, $\forall_j \ |\delta_j|c \leqslant \Delta$ for the method to work, where c is the speed of sound. Δ also corresponds to the chosen distance between the two microphones.

The tests performed in this work determined that such hypothesis does not hold for non-stationary speech signals such as sibilants [15].

3.3. Separation principle

By the local stationarity and W-disjoint properties of the signals, for a single active source j at a certain angular frequency ω, X_1^w and X_2^w, the windowed Fourier transforms of the inputs form the following system:

$$\begin{aligned} \begin{bmatrix} X_1^w(\tau,\omega) \\ X_2^w(\tau,\omega) \end{bmatrix} &\approx \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-i\omega\delta_1} & \cdots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1^w(\tau,\omega) \\ \vdots \\ S_N^w(\tau,\omega) \end{bmatrix} \\ &\overset{Wdisj.}{=} \begin{bmatrix} 1 \\ a_j e^{-i\omega\delta_j} \end{bmatrix} S_j^w(\tau,\omega) \end{aligned} \tag{7}$$

By this observation the most important quantity in DUET is $F(\omega,\tau)$, where the property of Eq. (6) is used:

$$F(\tau,\omega) := \frac{X_2^w(\omega)}{X_1^w(\omega)} = a_j e^{-i\omega\delta_j} \frac{S_j^w(\tau-\delta_j,\omega)}{S_j^w(\tau,\omega)} \approx a_j e^{-i\omega\delta_j} \tag{8}$$

If both local stationarity and W-disjoint properties are respected, the approximation in Eq. (8) is valid and since ω is known, it is possible to extract both the delay and attenuation between the microphones for the active source j from the complex quantity F, forming the *amplitude-phase function* Λ [16]:

$$\Lambda = (a_j, \delta_j)_\omega = (\|F(\tau,\omega)\|, \ -\arg(F(\tau,\omega))/\omega) \tag{9}$$

Each (τ,ω) bin can then be assigned to a binary mask of one of the sources on the time-frequency (T-F) domain.

[1] If $w(t)=1$, the windowed Fourier transform $\mathscr{F}^w$ is represented by $\mathscr{F}$.