Speech Communication 83 (2016) 42–53

Contents lists available at ScienceDirect

Speech Communication

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

Near-field signal acquisition for smartglasses using two acoustic vector-sensors

Dovid Y. Levin[a,*], Emanuël A.P. Habets[b,1], Sharon Gannot[a]

[a] Bar-Ilan University, Faculty of Engineering, Building 1103, Ramat-Gan, 5290002, Israel
[b] International Audio Laboratories Erlangen, Am Wolfsmantel 33,Erlangen91058, Germany

ARTICLE INFO

Article history:
Received 21 February 2016
Accepted 12 July 2016
Available online 18 July 2016

PACS:
43.60.Fg
43.60.Mn
43.60.Hj

Keywords:
Beamforming
Acoustic vector-sensors
Smartglasses
Adaptive signal processing

ABSTRACT

Smartglasses, in addition to their visual-output capabilities, often contain acoustic sensors for receiving the user's voice. However, operation in noisy environments may lead to significant degradation of the received signal. To address this issue, we propose employing an acoustic sensor array which is mounted on the eyeglasses frames. The signals from the array are processed by an algorithm with the purpose of acquiring the desired near-field speech signal produced by the wearer while suppressing noise signals originating from the environment. The array is comprised of two acoustic vector-sensors (AVSs) which are located at the fore of the glasses' temples. Each AVS consists of four collocated subsensors: one pressure sensor (with an omnidirectional response) and three particle-velocity sensors (with dipole responses) oriented in mutually orthogonal directions. The array configuration is designed to boost the input power of the desired signal, and to ensure that the characteristics of the noise at the different channels are sufficiently diverse (lending towards more effective noise suppression). Since changes in the array's position correspond to the desired speaker's movement, the relative source-receiver position remains unchanged; hence, the need to track fluctuations of the steering vector is avoided. Conversely, the spatial statistics of the noise are subject to rapid and abrupt changes due to sudden movement and rotation of the user's head. Consequently, the algorithm must be capable of rapid adaptation toward such changes. We propose an algorithm which incorporates detection of the desired speech in the time-frequency domain, and employs this information to adaptively update estimates of the noise statistics. The speech detection plays a key role in ensuring the quality of the output signal. We conduct controlled measurements of the array in noisy scenarios. The proposed algorithm preforms favorably with respect to conventional algorithms.

1. Introduction

Recent years have witnessed an increased interest in *wearable computers* (Barfield, 2016; Randell, 2005). These devices consist of miniature computers worn by users which can perform certain tasks; the devices may incorporate various sensors and feature networking capabilities. For example, a *smartwatch* may be used to display email messages, aid in navigation, and monitor the user's heart rate (in addition to functioning as a timepiece).

One specific type of wearable computer which has garnered much attention is the *smartglasses* — a device which displays computer generated information supplementing the user's visual field. A number of companies have been conducting research and development towards smartglasses intended for consumer usage (Cass and Choi, 2015) (e.g., Google Glass (Ackerman, 2013) and Microsoft HoloLens). In addition to their visual-output capabilities, smartglasses may incorporate acoustic sensors. These sensors are used for hands-free mobile telephony applications, and for applications using a voice-control interface to convey commands and information to the device.

The performance of both of these applications suffers when operating in a noisy environment: in telephony, noise degrades the quality of the speech signal transmitted to the other party; similarly, the accuracy of automatic speech recognition (ASR) systems is reduced when the desired speech is corrupted by noise. A review of one prominent smartglasses prototype delineated these two issues as requiring improvement (Sung, 2014).

To deal with these issues, we propose a system for the acquisition of the desired near-field speech in a noisy environment. The system is based on an acoustic array embedded in eyeglasses

* Corresponding author.
E-mail addresses: david.levin@live.biu.ac.il (D.Y. Levin), emanuel.habets@audiolabs-erlangen.de (E.A.P. Habets), Sharon.Gannot@biu.ac.il (S. Gannot).

[1] A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

http://dx.doi.org/10.1016/j.specom.2016.07.002
0167-6393/

frames worn by the desired speaker. The multiple signals received by the array contain both desired speech as well as undesired components. These signals are processed by an adaptive beamforming algorithm to produce a single output signal with the aim of retaining the desired speech with little distortion while suppressing undesired components.

The scenario of a glasses mounted array presents some challenging features which are not encountered in typical speech processing. Glasses frames constitute a spatially compact platform, with little room to spread the sensors out. Typically, when sensors are closely spaced the statistical qualities of the noise at each sensor are highly correlated presenting difficulties in robust noise suppression (Bitzer and Simmer, 2001). Hence, special care must be taken in the design of the array.

The proposed array consists of two AVSs located, respectively, at the fore of the glasses' right and left temples. In contrast to conventional sensors that measure only the pressure component of a sound field (which is a scalar quantity), an AVS measures both the pressure and particle-velocity components. An AVS consists of four subsensors with different spatial responses: one omnidirectional sensor (corresponding to pressure) and three orthogonally oriented dipole sensors (corresponding to the components of the particle-velocity vector). Hence, the array contains a total of eight channels (four from each AVS). Since each subsensor possesses a markedly different spatial response, the statistical properties of the noise at the different subsensors are diverse. Consequently, robust beamforming is possible in spite of the limited spatial aperture. Another advantage afforded by the use of AVSs is that the dipole sensors amplify near-field signals more so than conventional omnidirectional sensors. Due to these sensors, the desired speech signal (which is in the near-field due to the proximity to the sensors) undergoes a relative gain and is amplified with respect to the noise. The *relative gain* is explained and quantified in Section 2. The interested reader is referred to the Appendix for further information on AVSs.

The configuration in which the array is mounted on the speaker's glasses differs from the typical scenario in which a microphone array is situated in the environment of the user. The glasses configuration possesses particular properties which lead to a number of benefits with respect to processing: (i) The close proximity of the desired source to the sensors leads to high signal-to-noise ratio (SNR) which is favorable. (ii) For similar reasons, the reverberation of the desired speech is negligible with respect to its direct component, rendering dereverberation a nonissue. (iii) Any change in the location of the desired source brings about a corresponding movement of the array which is mounted thereon. Consequently, the relative source-sensors configuration is essentially constant, precluding the need for tracking changes of the desired speaker's position.

Conversely, the glasses-mounted configuration presents a specific challenge. The relative positions of the undesired acoustic sources with respect to the sensor array are liable to change rapidly. For instance, when the user rotates his/her head the relative position of the array to external sound sources undergoes significant and abrupt changes. This necessitates that the signal processing stage be capable of swift adaptation.

The proposed algorithm is based on minimum variance distortionless response (MVDR) beamforming which is designed to minimize the residual noise variance under the constraint of maintaining a distortionless desired signal. This type of beamforming was proposed by Capon (1969) in the context of spatial spectrum analysis of seismic arrays. Frost (1972) employed this idea in the field of speech processing using a time-domain representation of the signals. Later, Gannot et al. (2001) recast the MVDR beamformer in the time-frequency domain. In the current work, we adopt the time-frequency formulation.

Fig. 1. The proposed sensor locations are indicated in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the proposed algorithm, the noise covariance matrix is adaptively estimated on an ongoing basis from the received signals. Since the received signals contain both desired and undesired components, the covariance matrix obtained from a naive implementation would contain significant contributions of energy from the desired speech. This is detrimental to the performance of the processing. To prevent desired speech from contaminating the noise covariance estimation, a speech detection component is employed. Time-frequency bins which are deemed likely to contain desired speech are not used for estimating the noise covariance.

To further reduce noise, the output of the MVDR stage undergoes post-processing by a single-channel Wiener filter (SWF). It has been shown (Simmer et al., 2001) that application of MVDR beamforming followed by a SWF is optimal in the sense of minimizing the mean square error (MSE) [since it is equivalent to the multichannel Wiener filter (MWF)].

The paper is structured as follows: Section 2 describes the motivation guiding our specific array design. In Section 3, we introduce the notation used to describe the scenario in which the array operates and then present the problem formulation. Section 4 presents the proposed algorithm and how its various component interrelate. Section 5 evaluates the performance of the proposed algorithm, and Section 6 concludes with a brief summary.

2. Motivation for array design

In this section, we discuss the considerations which lead to our choices for the placement of the sensors and the types of sensors used.

An AVS is located at the fore of each of the glasses' temples (see Fig. 1). The reason for selecting this location is that there is a direct "line of sight" path from the speaker's mouth to the sensors. For other locations on the frames, such as the temples' rear sections or the areas above the lenses, the direct path is obstructed by human anatomy or the physical structure of the glasses. The areas underneath the lenses were also considered as they *do* have an unobstructed line to the mouth; however, embedding a microphone array at this locale was deemed to render the resulting frame structure too cumbersome.

Choosing an AVS based array, rather than using conventional sensors, leads to several advantages. Firstly, the inherent directional properties of an AVS lend to the distinction between the desired source and sound arriving from other directions. In contrast, a linear arrangement of conventional omnidirectional sensors along a temple of the glasses frame would exhibit a degree of directional ambiguity – it is known that the response of such linear arrays maintains a conical symmetry (Van Trees, 2002).