Pattern Recognition Letters 33 (2012) 373–380

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Voice activity detection and speaker localization using audiovisual cues

Dante A. Blauth [a], Vicente P. Minotto [b], Claudio R. Jung [b,*], Bowon Lee [c], Ton Kalker [d,1]

[a] Applied Computing – UNISINOS, Av. Unisinos, 950, São Leopoldo 93022-000, RS, Brazil
[b] Institute of Informatics – UFRGS, Av. Bento Gonçalves, 9500, Porto Alegre 91501-970, RS, Brazil
[c] Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA
[d] Huawei Innovation Center US R&D, 2330 Central Expressway, Santa Clara, CA 95050, USA

ARTICLE INFO

Article history:
Available online 10 September 2011

Keywords:
User interfaces
Voice activity detection
Speaker localization
Multimodal analysis
Hidden Markov Models

ABSTRACT

This paper proposes a multimodal approach to distinguish silence from speech situations, and to identify the location of the active speaker in the latter case. In our approach, a video camera is used to track the faces of the participants, and a microphone array is used to estimate the Sound Source Location (SSL) using the Steered Response Power with the phase transform (SRP-PHAT) method. The audiovisual cues are combined, and two competing Hidden Markov Models (HMMs) are used to detect silence or the presence of a person speaking. If speech is detected, the corresponding HMM also provides the spatio-temporally coherent location of the speaker. Experimental results show that incorporating the HMM improves the results over the unimodal SRP-PHAT, and the inclusion of video cues provides even further improvements.

1. Introduction

Nowadays, keyboard and mouse are the most popular devices for Human Computer Interaction (HCI), adopted by the vast majority of personal computers. Despite being intuitive and easy to use, they may not be adequate in a variety of applications. For instance, the manual annotation of multimedia data (photos, videos, and music clips, to name a few) into a set of tags using keyboard and mouse is a tiresome task, and other ways of interaction (such as audiovisual data) seem to be more natural. In particular, the analysis of speech and facial features appear to be promising in the development of multimodal HCI systems (Jaimes and Sebe, 2007).

There are several challenges when exploring facial cues and speech data to develop HCI systems. Firstly, the face must be detected and tracked robustly in time, which is a complex task in the presence of partial occlusions, head tilts and turns. Secondly, the audio analysis (mainly speech detection) is highly corrupted by background noise (e.g. an air-conditioning system), so that it is necessary to detect when a person is speaking or not (this problem is usually referred to as voice activity detection – VAD). Finally, when more than one user is captured by the camera, it is important to determine which person is actually interacting with the computer (in the case of voice commands, that means finding the active speaker at a given time).

This paper presents a new approach to estimate the location of the active speaker based on audiovisual cues. We propose a Hidden Markov Model (HMM) that characterizes the expected spatio-temporal properties of a typical speaker considering the input captured by an array of microphones, and its extension to include visual cues. This HMM imposes spatio-temporal constraints on the location of the active speaker, improving the results of audio-only localization. Another HMM to model silence periods is also presented, so that VAD can also be achieved by comparing the speech and silence HMMs.

The remainder of this paper is organized as follows: Section 2 presents some related work, and the proposed approach is described in Section 3. Some experimental results are provided in Section 4, and the conclusions are drawn in Section 5.

2. Related work

There are several approaches for VAD and for SSL, using mostly audio cues, video cues or a combination of both (multimodal processing). In general, VAD and SSL are considered as two separate problems, and a brief revision of both problems is presented below.

Most of existing approaches for VAD are based on audio cues, either relying on characteristics of voice patterns in the frequency domain or pre-determined (or estimated) levels of background noise. Sohn et al. (1999) presented a statistical method using the complex Gaussian assumption in the frequency domain for both speech and noise, based on the likelihood ratio test (LRT). Additionally, they also proposed an effective hang-over scheme based on a HMM. Ramírez et al. (2005) proposed its extension, by employing

* Corresponding author.
E-mail addresses: danteab@gmail.com (D.A. Blauth), vpminotto@inf.ufrgs.br (V.P. Minotto), crjung@inf.ufrgs.br (C.R. Jung), bowon.lee@hp.com (B. Lee), ton.kalker@huawei.com (T. Kalker).
[1] Ton Kalker performed the work while at Hewlett-Packard Laboratories.

0167-8655/$ - see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.patrec.2011.09.002

multiple observations to include temporal smoothing. A further improvement was presented in (Ramírez et al., 2007), by using contextual multiple hypothesis testing.

Other authors have used different statistical models to characterize speech periods. For instance, a Laplacian distribution was used to model speech in the DCT domain in (Gazor and Zhang, 2003), which has shown to be a better model than a Gaussian. More recently, Lee and Muhkerjee (2010) proposed a statistical algorithm for VAD, aiming to detect higher level speech activities (e.g. sentences instead of syllables, words, phrases, etc.). Their approach uses two distinct features for VAD, energy and entropy in the DCT domain, which are modeled as chi-square and Gaussian distributions respectively.

It should be noticed that the approaches described in (Sohn et al., 1999; Gazor and Zhang, 2003) aim to detect very short-time silence periods suitable for tasks such as speech coding or automatic speech recognition, while our approach aims to detect higher level speech activities, as in (Lee and Muhkerjee, 2010; Ramírez et al., 2005, 2007).

Regarding SSL (in our case, the sound source is the active speaker), most approaches rely only on audio information. It is necessary to have more than one microphone, and to analyze the relationships among the signals captured by different microphones (Brandstein and Ward, 2001). For a two-microphone case, we can find the time difference of arrival (TDOA) using the generalized cross-correlation (GCC) method, which involves a frequency weighting function (Brandstein and Ward, 2001). One of the most popular frequency weighting for GCC is the phase transform (PHAT), which effectively whitens the microphone signals to equally emphasize all frequencies before computing the cross correlation. Even with noise and reverberation, the GCC-PHAT has been reported to work reasonably well in many practical situations (Brandstein and Ward, 2001).

For multiple microphone cases, we can find the source location by triangulation given a set of TDOA's from different microphones pairs (Brandstein et al., 1997). The TDOA-based method becomes unreliable when the individual TDOA estimates are inaccurate to begin with. Unfortunately, this is often the case in typical acoustic environments. Alternatively we can use the *Steered Response Power* (SRP) method (DiBiase, 2000), which can be considered as an extension of the GCC method to multiple microphones. The main idea of the SRP is to steer the microphone array to all possible candidate source locations to find one with the maximum power, typically using some frequency weighting (filtering in the time domain). In particular, the SRP method with the PHAT frequency weighting (SRP-PHAT) has been reported to be more robust with respect to acoustic corruptions, such as background noise and reverberation compared with the TDOA-based methods (Brandstein and Ward, 2001; DiBiase, 2000; Do et al., 2007).

With an array consisting M microphones, the SRP-PHAT of the sound source at a position $\boldsymbol{q}$ is (DiBiase, 2000)

$$P(\boldsymbol{q}) = \sum_{m=1}^{M}\sum_{l=1}^{M}\int \frac{X_m(\omega)X_l^*(\omega)}{|X_m(\omega)X_l^*(\omega)|} e^{j\omega(\Delta_m^{\boldsymbol{q}}-\Delta_l^{\boldsymbol{q}})} d\omega, \tag{1}$$

where $X_m(\omega)$ is the Fourier Transform of the signal at the mth microphone, $\Delta_m^{\boldsymbol{q}}$ is the time delay computed from position $\boldsymbol{q}$ to the mth microphone.

Since the SRP-PHAT requires significant amount of computation, i.e., double summation and one integration for each candidate source, an alternative expression was proposed in (Zhang et al., 2007):

$$P(\boldsymbol{q}) = \int \left| \sum_{m=1}^{M} \frac{X_m(\omega)}{|X_m(\omega)|} e^{j\omega\Delta_m^{\boldsymbol{q}}} \right|^2 d\omega. \tag{2}$$

SSL can also be done using statistical modeling of multichannel audio signals using e.g. multivariate complex Gaussian (Zhang et al., 2007), or Laplacian (Lee et al., 2008) distributions whose computational cost is typically much higher than the SRP-PHAT and thus not quite suitable for real-time implementations. Furthermore, it is important to note that the audio-based approaches described so far work on sound buffers independently, not exploring temporal coherence.

Multimodal approaches for SSL explore different sensors, mostly focused on audiovisual integration. Wang and Brandstein (1997) proposed a face tracking algorithm based on both sound and visual cues. Initial talker locations are estimated acoustically from microphone array data, based on the TDOA estimation followed by a triangulation procedure. The final location is obtained based on video cues (acquired with a single camera), using mostly motion and edge information. Their work was further extended in (Wang et al., 2000) by adding head pose estimation using multiple cameras and multi-channel speech enhancement techniques.

In (Vermaak et al., 2001; Perez et al., 2004), a particle filtering framework was employed for data fusion, since it deals better with non-Gaussian distributions (opposed to Kalman filtering). Both approaches explore only a pair of microphones and a single monocular camera, and obtain the sound localization by measuring the TDOA between signals arriving at the two microphones. In (Vermaak et al., 2001), an active contour tracking approach was used to explore visual data, using monochromatic information, while color and motion cues were used in (Perez et al., 2004).

Gatica-perez et al. (2007) presented a probabilistic approach to jointly track the location and speaking activity of multiple speakers in meeting room equipped with a circular microphone array (with 8 microphones) on a table and multiple cameras, that capture frontal and top views of the participants. They used the SRP-PHAT approach for SSL, and the visual observations were based on models of the shape and spatial structure of human heads. The fusion was performed with a Markov Chain Monte Carlo particle filter. Despite the good results presented in the paper, a point to be improved is the initialization of the targets. Also, the computational cost was not discussed, and more than one camera is required.

The group of Zhang et al. (2008) proposed a multimodal speaker identification approach that fuses audio and visual information at the feature level by using boosting to select features from a combined pool of both audio and visual features simultaneously, using a circular array with 6 microphones and a panoramic camera. The authors showed that their multimodal approach performed better than a unimodal sound source locator, but the motion cue may pose problems for stationary participants. Talantzis et al. (2009) proposed an approach that estimates independently the position of the active speaker in cluttered and reverberant environments using audio and video information, and combined both outputs. Their method was tested with a large microphone array (80 microphones located in different locations inside the acoustic enclosure and organized in different topologies), and a set of five synchronized and calibrated cameras. One clear drawback of their approach is the hardware requirement (tens of microphones and multiple calibrated cameras).

Although VAD and SSL have been treated as independent problems in the literature, this paper presents a new approach that explores spatio-temporal characteristics that arise in the SSL problem to solve the VAD problem. To further improve SSL results, we have also included visual information captured by a single camera. The proposed approach is described in the next section.

3. Our approach

The proposed approach explores the expected spatio-temporal consistency of audio signals through two competing HMMs (one for silence and the other for speech) to distinguish silence from speech. When voice activity is detected, visual information is