Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 622–642
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners

Yan-Chen Lu [a,*], Martin Cooke [b,c]

[a] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
[b] Ikerbasque (Basque Foundation for Science), 48011 Bilbao, Spain
[c] Language and Speech Laboratory, Faculty of Letters, University of the Basque Country, Spain

Available online 8 June 2010

Abstract

Localisation in azimuth and distance of sound sources such as speech is an important ability for both human and artificial listeners. While progress has been made, particularly for azimuth estimation, most work has been directed at the special case of static listeners and static sound sources. Although dynamic sound sources create their own localisation challenges such as motion blur, moving listeners have the potential to exploit additional cues not available in the static situation. An example is motion parallax, based on a sequence of azimuth estimates, which can be used to triangulate sound source location. The current study examines what types of listener (or sensor) motion are beneficial for localisation. Is any kind of motion useful, or do certain motion trajectories deliver robust estimates rapidly? Eight listener motion strategies and a no motion baseline were tested, including simple approaches such as random walks and motion limited to head rotations only, as well as more sophisticated strategies designed to maximise the amount of new information available at each time step or to minimise the overall estimate uncertainty. Sequential integration of estimates was achieved using a particle filtering framework. Evaluations, performed in a simulated acoustic environment with single sources under both anechoic and reverberant conditions, demonstrated that two strategies were particularly effective for localisation. The first was simply to move towards the most likely source location, which is beneficial in increasing signal-to-noise ratio, particularly in reverberant conditions. The other high performing approach was based on moving in the direction which led to the largest reduction in the uncertainty of the location estimate. Both strategies achieved estimation errors nearly an order of magnitude less than those obtainable with a static approach, demonstrating the power of motion-based cues to sound source localisation.

Keywords: Active hearing; Sound source localisation; Interaural time difference; Motion parallax; Particle filtering

1. Introduction

Listeners routinely solve problems of navigation, object identification and avoidance in challenging environments with an efficiency which disguises the true complexity of the task. Localisation of sounds in space is particularly critical. Anyone who has ridden a bicycle on a busy road will appreciate that the ability to rapidly and robustly locate sounds source of interest is of great everyday importance. In general, knowing *where* to listen can improve speech intelligibility in the presence of other sound sources (Kidd et al., 2005).

Most work on understanding and modelling human sound localisation has taken place in a setting where the sensors and sources of interest are assumed to be static. While there are a limited number of situations, such as recording studios or human–machine spoken interaction with headset microphones and headphones, where this assumption is approximately true, there are also many contexts where listeners and/or sound sources are mobile. Further, certain types of application, such as hearing aid processors or other forms of wearable audio (Sawhney

* Corresponding author.
E-mail address: y.c.lu@dcs.shef.ac.uk (Y.-C. Lu).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.06.001

and Schmandt, 2000; Lukowicz et al., 2004) presuppose mobile sensors. While these active scenarios require more sophisticated processing to handle issues such as auditory motion blur and the need to track dynamic sources, they also contain opportunities to exploit cues not available to static listeners. It is well-known that head movements can be used to resolve front-back ambiguities in localisation (Wallach, 1940; Thurlow et al., 1967; Mackensen, 2004) and other studies (e.g. Speigle and Loomis, 1993; Ashmead et al., 1995) suggest that body motion helps in distance estimation. There are many other ways in which motion might help in audition (Cooke et al., 2008). For example, moving towards a source can improve the ratio of direct to reverberant energy. Head rotation can locate the target source in the frontal plane where spatial resolution is at its finest. Movement away from hard surfaces can reduce the effect of reverberation. The head and body can attenuate intense competing sources by a significant amount. Further, a target sound object is easier to segregate from omni-directional reverberation by exploiting its relative response to body motion (Martinson and Schultz, 2006).

Engineering solutions to the localisation problem typically exploit microphone arrays containing $N > 2$ sensors. However, it is of interest to explore binaural/stereo approaches, not only because there is a wealth of existing behavioural data and models of binaural processing to draw upon, but also to test predictions about the value of particular motion strategies emerging from the model in listeners. For some applications involving wearable audio such as personal audio diaries, retaining the link to listener behaviour may be important in making sense of data collected. In any case, one useful way to think about mobile binaural audio is that it provides N binaural *asynchronous* sensors. Even for mobile sources, the additional data provided by sampling at distinct spatial locations can be exploited to improve localisation, as we show in this article.

One intriguing question is: what kind of motion is beneficial for sound source localisation? Does arbitrary motion help in triangulating sources, or are certain trajectories optimal? Answering this question will help in the design of motion plans for mobile robots equipped with audio sensors and may lead to predictions about human movement strategies in adverse conditions. The primary purpose of the work described in this paper is to evaluate the performance of distinct motion strategies, ranging from simple approaches limited solely to head movement, to more sophisticated motions based on information-theoretic criteria such as moving in the direction which maximises estimate entropy. Of interest are those strategies which lead to robust estimates of source location, for both static and moving sources, and which also converge rapidly to a good solution.

Full source localisation involves estimating source azimuth, elevation and distance relative to the listener. Of these, azimuth and distance are of most practical relevance to human listeners. Location in azimuth has been thoroughly investigated (Jeffress, 1948; Blauert, 1997) and a number of computational models exist which produce levels of performance similar to listeners (Lindemann, 1986; Bodden, 1993; Gaik, 1993). Distance has received far less attention (Zahorik et al., 2005). Here, we introduce procedures for estimating both azimuth and distance.

The approach taken in the current study is depicted in Fig. 1. Left and right ear signals are derived from room simulations (anechoic and with mild and moderate reverberation). These signals are processed by an auditory filterbank and successive cross-correlations performed to allow estimation of sound source location in both azimuth and distance through triangulation. Microphone motion enables triangulation for sound distance by sequentially integrating azimuth observations obtained from binaural input (Lu et al., 2007) or array input (Sasaki et al., 2006). A range of location estimates are maintained with a sequential Bayesian particle filtering framework, and a given motion strategy is applied. A specific motion is chosen and updated binaural signals are computed based on the new location. The process cycles in this manner during the "walk".

Section 2 describes the source tracking framework and the extraction of cues used as the basis for sound source localisation. The particle filtering architecture is introduced in Section 3 together with an extension for moving listeners. Section 4 details the different motion strategies, or *strategic walks*, evaluated in the current study. A subset of walks is based on *motion entropy*, which is proposed as a measure of the uncertainty associated with the sound source location in response to listener movement. The outcome of an extensive evaluation of different motion strategies is presented in Section 5.

2. Source localisation in distance and azimuth

2.1. Source tracking framework

The current study addresses localisation in azimuth and distance of a single source, which may be in motion, based on binaural inputs received by an artificial listener, also potentially in motion. Source location x_t at time t is defined in terms of distance and azimuth components, $x_{r,t}$ and $x_{\phi,t}$ specified in a listener-centred spherical coordinate system whose 2D transverse plane passes through the ears:

$$x_t = [x_{r,t}, x_{\phi,t}]. \tag{1}$$

An output variable of the state x_t is denoted y_t and is assumed to be described by the output equation:

$$y_t = g(x_t, v_t), \tag{2}$$

where v_t represents the combined effect of distortions due to reverberation and noise. The unknown and possibly non-linear function g links the state to the noisy measurement. The observation of the underlying state x_t, i.e. the current source's location in distance and azimuth, is derived from the left and right binaural signals at time t,