A Bayesian System for Noise Robust Binaural Speaker Counting for Humanoid Robots

Matthew Tata, Austin Kothig, and Francesco Rea, *Member, IEEE*

***Abstract*— Humans make use of auditory cues to interact with others in complex acoustic environments. Such ability remains a challenge for most of binaural robots. Multiple distinct acoustic events usually mix together and binaural robots are requested to disambiguate and interpret the environmental scene with two sensors. More importantly, robots that interact with humans should be able to interpret correctly relevant insights such as how many speakers are present in the environment. In this paper, we propose a Bayesian method of selective processing of acoustic data that detects typical amplitude envelope dynamics of human speech to infer the number of speakers in the environment. Further, we measure how much this method outperforms traditionally methods based on amplitude detection only.**

I. Introduction

Understanding the auditory scene is a critical ability for social interactive robotics. Social robots able to selectively attend to salient events have a considerable advantage in operating in complex and cluttered acoustic environments. Such robots should be able to relocate attentional focus on task-relevant targets and specifically on the human social partner. Converging evidence shows that in both daily activities and in learning processes, the ability to focus on the social partner is a strategic advantage and this is especially advantageous for robotic platforms. Although models of auditory attention have been extensively studied by neuroscientists and psychologists for decades, such models have only been partially transferred to robotics platforms, primarily because unmixing the complex free-field acoustic signal remains a largely unsolved problem. The human brain solves the unmixing problem using only the inputs from two ears. Starting with this observation, this paper presents a neurobiologically inspired computational model of spatial selective orientating implemented on the iCub humanoid robot [1]. We exploited the fact that human speech has predictable temporal dynamics that differ from the dynamics of many distractors. Specifically, we developed a method to find and localize speech-specific amplitude modulations. Natural speech across languages is characterized by a 4-7 Hz envelope modulation related to the syllable rate, and results from human cognitive neuroscience have revealed brain mechanisms that seem to use these envelope dynamics to unmix complex acoustic scenes [2, 3].

M. Tata is with the Canadian Centre of Behavioural Neuroscience, University of Lethbridge, AL T1K 3M4 Canada (corresponding author to; e-mail: matthew.tata@uleth.ca).

A. Kothig, is with the University of Lethbridge, AL T1K 3M4 Canada (e-mail: kothiga@uleth.ca).

F.Rea is with the Robotics Brain and Cognitive Science, Istituto Italiano di Tecnologia, Genova, GE 16152 Italy, (e-mail: Francesco.rea@iit.it).

II. Computational Implementation

In this paper, we describe a biologically inspired computational model for localising speakers by a binaural humanoid robot. The fundamental ability of this approach is to parse a complex auditory scene to determine first the number of speakers, and second, their probable location. Our approach extends a Bayesian active hearing algorithm [4] that uses instantaneous egocentric evidence to update an allocentric posterior map. We extended the algorithm to test whether use of the 5hz envelope dynamics of speech can help to reject distractor sound sources. We present a computational framework that implements such inferring process on the iCub humanoid robot, and discuss neurobiological correlates of this approach.

A. Gammatone Filterbank and Sound Localizer

The input stage of the human auditory pathway was approximated by a Gammatone filter bank, which has a uniform distribution in equivalent rectangular bandwidth (ERB) frequency domain. Specifically a set of rectangular bandpass filters were implemented, where f_c is the center frequency of the filters in hz. The human auditory system makes use of the spatial dimension to unmix the auditory scene, and spatial information is derived in part from interaural phase differences due to different path lengths between unique sources and the two ears. This interaural time difference cue was extracted from the scene with a microphone array with two sensors with known displacement *d*. The spatial unmixing of the low-level auditory system was approximated with a bank of swept delay-and-sum beam formers. The method computes an approximate spatial distribution of sound energy by shifting acoustic samples of one channel relative to the other, followed by a summation of both the channels to compute a set of 43 azimuthally directed beams within each frequency band. The angle of each beam is governed by $\theta = Re(\sin^{-1}(cT_b)/d)$ in which T_b is the time delay between the shifted samples and d is the distance between the microphones. The output matrix stores this spectrospatially decomposed map of sound energy.

This output is then collapsed along the time dimension by two different approaches: in the Amplitude Only condition, the RMS amplitude of each band x beam signal is computed. In the Envelope condition, the 5hz envelope modulations due to speech are extracted from each band x beam signal by computing the absolute value of the Hilbert transform, then band-pass filtering the envelope and collapsing across time using RMS. In both conditions, the resulting bands x beams image is converted to a probabilistic map of sound sources by

normalizing each frequency band to sum to one. A linear interpolation between beam angles at each frequency allows for a degree-normal egocentric spatial representation of the acoustic scene. We next perform a co-registration of multiple egocentric acoustic maps generated by head rotations into an allocentric reference (AM^{allo}) sharing the same robot frame-of-reference regardless of the robot head orientation. To reduce front-back confusion and spatial aliasing associated with beamforming, these multiple maps are combined with a recursive Bayesian approach. The creation of the acoustic Bayesian map (ABM) is defined by the product of all the allocentric acoustic maps (AM^{allo}) and approximates the output of the inferior colliculus of the mammalian auditory pathway. Thus the Amplitude Only ABM describes the spectrospatial scene unmixed on the basis of signal amplitude, whereas the Envelope ABM represents the spectrospatial scene unmixed on the basis of 5hz envelope dynamics. To arrive at a single posterior distribution of sound sources across the azimuthal plane, we averaged across frequency bands, yeilding a distribution of belief that a sound source occupied a particular azimuthal angle. Each peak in this distribution can be considered a sound source and a candidate for target selection.

III. Experiment and Results

Our goal was to develop a system that reliably reports the presence and location of a human voice regardless of competing noise sources. We thus tested whether the Amplitude Only and Envelope approaches were differentially robust to increases in the set-size of noise sources when only a single target sound source was present in the scene. We varied the set size of simultaneous sound sources in the scene from one to six to assess how the computational model parsed increasingly complex auditory scenes. In particular, set-size conditions contained all combinations of one artificially generated 5hz amplitude modulated tone and between none and six unique pink noise sources. This resulted in 140 trials for all the seven noise source conditions and the total number of trials in the experiment is 980.

A. Audio capture and Data Set

We tested the localization accuracy by reproducing auditory targets and distractors in the free field in the auditory virtual-reality lab at the University of Lethbridge. This space features sound-proof walls and up to 14 studio monitors arranged 1.5m away from the center of the room, linearly spaced by 30 degrees on a semicircle arc in the front field from 0 to 180 degrees. We demonstrated the solution with a humanoid robot iCub head. We placed the iCub head at the center of the room facing the center of the semicircle. Two APEX150 cardioid microphones attached on the head with an interaural distance of 0.145m. During sound presentations, the head rotated through the azimuthal plane, between ±40 degrees of the allocentric midline in 3 seconds. Each audio evidence frame was 32768 samples with a 50% temporal overlap, and each frame was associated with a reading of the neck yaw encoder. A target tone complex was created by generating set of pure tones at every second centre frequency in the ERB-space gammatone filterbank. Distractors were uniquely generated (i.e. uncorrelated) noise sources with 1/f spectral distribution (pink noise). Each trial was 15 seconds and all audio playback signals were balanced to have an RMS amplitude of 0.045.

C. Results

We considered two performance measures: resolving a single 5-hz modulated tone complex by counting the number of candidate peaks in the azimuthal belief and localizing that signal by finding the absolute error between the highest peak and the true arrival angle.

Figure 1. Comparing amplitude only and envelope methods in counting and localizing candidate targets, with the number of true targets held at 1.

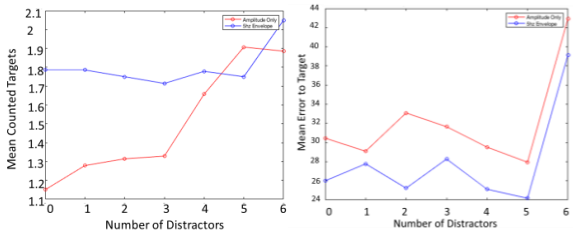

Figure 1 shows counting and error measurements as a function of increasing noise source set size. The critical observation in counting is not the absolute accuracy but rather whether or not there was an affect of set size. Whereas the counting performance relying on amplitude only was modulated with the number of distractors, the Envelope approach is largely unaffected by increasing distractors (left panel). A repeated-measures ANOVA with set-size and envelope approach supported this significant interaction ($F_{6,1668} = 4.9$; $p<0.001$). This advantage did not come at a cost of localization, as the error performance of the Envelope condition was significantly better than that of the Amplitude condition (right panel) (Tukey HSD $p<0.01$).

IV. Conclusion

Resolving sound sources from a mixture is a challenging task. Using a biologically inspired our results show that a biologically inspired Bayesian approach can successful resolve a target from noise sources based on amplitude envelope dynamics. This result validates the idea that characteristic dynamics of human speech might be used to successfully parse the auditory scene.

References

[1] G. Metta, G. Sandini, D. Vernon, L. Natale, F.Nori "The iCub humanoid robot: an open platform for research in embodiment cognition," in *Proceedings of the 8th workshop on performance metrics for intelligent systems*, ACM, 2008, pp. 50–56.

[2] Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., & Poeppel, D. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". In *Neuron*, 77(5), pp. 980-991.

[3] Hambrook, D. A., & Tata, M. S. (2019). The effects of distractor set-size on neural tracking of attended speech. In *Brain and language*, 190, pp. 1-9.

[4] Hambrook, D. A., Ilievski, M., Mosadeghzad, M., & Tata, M. (2017). A Bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue. in *PloS one*, 12(10),