In: Proceedings of the 24th International Conference on Artificial Neural Networks (ICANN 2014),
LNCS 8681, pp. 619-626, Springer Heidelberg. Hamburg, DE, September 15-19, 2014.

Improving Humanoid Robot Speech Recognition with Sound Source Localisation

Jorge Davila-Chacon[1], Johannes Twiefel[1], Jindong Liu[2], and Stefan Wermter[1]

[1] University of Hamburg, Department of Informatics, Knowledge Technology Group
Vogt-Kölln-Straße 30, D-22527 Hamburg, Germany
[2] Imperial College London, Department of Computing
Huxley Building, South Kensington Campus, London, SW7 2AZ, UK
http://www.informatik.uni-hamburg.de/WTM/

Abstract. In this paper we propose an embodied approach to automatic speech recognition, where a humanoid robot adjusts its orientation to the angle that increases the signal-to-noise ratio of speech. In other words, the robot turns its face to 'hear' the speaker better, similar to what people with auditory deficiencies do. The robot tracks a speaker with a binaural sound source localisation system (SSL) that uses spiking neural networks to model relevant areas in the mammalian auditory pathway for SSL. The accuracy of speech recognition is doubled when the robot orients towards the speaker in an optimal angle and listens only through one ear instead of averaging the input from both ears.

Keywords: Human-robot interaction, robot speech recognition, binaural sound source localisation.

1 Introduction

Perception is a complex cognitive task that allows us to represent our environment and find meaning in it. In the case of auditory perception, our brain is capable of extracting information from diverse cues encoded in sound. Low-level sound cues help us to localise sound sources in space and track their motion, which in turn, allows us to segregate sound sources from noisy backgrounds and to understand natural language. Recent work in automatic speech recognition (ASR) use robotic platforms to replicate such processing pipeline. Some robotic approaches use arrays of several microphones to locate speech sources in space. Afterwards they use this knowledge to separate the speech signals from a noisy background [1,9]. However, they require prior knowledge about the presence of sound sources and their number. Other robotic approaches make use of binaural platforms [5,8]. In these systems sound source localisation (SSL) only makes use of spatial cues in sound's low frequencies, and speech segregation can only be done when speech comes from the set of trained angles.

In this paper we explore the possibility of improving robotic ASR with a neural SSL system. Our SSL approach is inspired by the neural processing of sound in the mammalian auditory pathway, which makes use of sound cues in low

and high frequency components, and it is not constrained to specific angles [14]. For speech segregation we take a behavioural approach, where the robot orients optimally to the speaker.

1.1 Neural Correlates of Acoustic Localisation

Sound waves approaching a human are firstly affected by the absorption and diffraction from the torso, head and pinnae. The first stage for the neural encoding of information in sound waves occurs in the cochlea, where the organ of Corti transduces mechanical vibrations from the basilar membrane into neural spikes. These spikes are then delivered through the auditory nerve to the cochlear nucleus, a relay station that forwards information to the medial superior olive (MSO) and the lateral superior olive (LSO). The MSO and LSO are of our particular interest, as they are in charge of extracting interaural time differences (ITD) and interaural level differences (ILD).

The MSO encodes more efficiently ITDs from low-frequencies in sound [14]. Such delay mechanisms can be achieved by different axon lengths -or different thicknesses of myelin-sheaths on the axon- of excitatory neurons from the ipsilateral and contralateral cochlear nucleus. The LSO encodes more efficiently ILDs from high-frequencies in sound [14]. In the case of the LSO, the mechanism underlying the extraction of level differences is less clear. However, it is known that LSO neurons receive excitatory input from the ipsilateral ear and inhibitory input from the contralateral ear, and they show a characteristic change in their spiking rate for sound sources located at specific angles on the azimuthal plane. Finally, the output from the MSO and the LSO is integrated in the inferior colliculus (IC), where a more coherent spatial representation is formed across the frequency spectrum [14]. This combination of both cues can be seen as a multimodal integration case, where ITDs and ILDs represent the different modalities to be merged in order to improve the SSL performance of a natural system.

2 Experimental Setup and Basis Methodologies

2.1 Virtual Reality Setup

Our experiments were carried in a virtual reality (VR) setup designed for testing multimodal integration architectures. This experimental setup allows to control the temporal and spatial presentation of images and sounds to robotic and human subjects. When experiments are run, the subject is located at the radial center of a projection screen shaped as half cylinder. As seen in figure 1, behind the screen there are 13 speakers evenly distributed on the azimuth plane at angles $\theta_{lspk} \in \{0°, 15°, \ldots, 180°\}$. A detailed description of this setup –and the principles behind its design– can be found in [2].

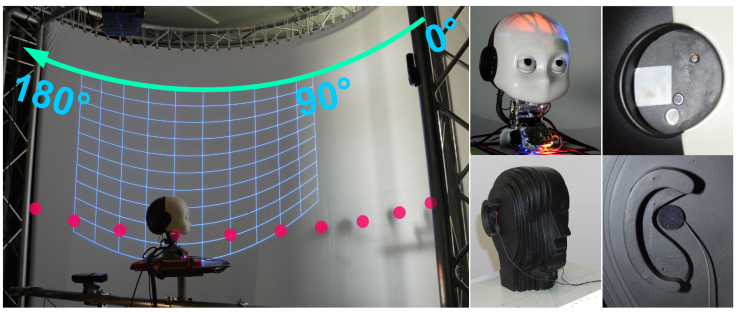

Fig. 1. On the left is shown the audio-visual VR experimental setup. The grid shows the curvature of the projection screen surrounding the iCub humanoid head and the dots represent the location of sound sources behind the screen. On the right can be seen the humanoid heads used during our experiments. The robot ears consist of microphones perpendicular to the sagittal plane surrounded by pinnae.

2.2 Humanoid Robotic Platforms

The humanoid platforms used in our experiments are the iCub robotic head [3] and the Soundman wooden head[3] modified by our group to rotate on the azimuth plane. The iCub is a humanoid robot designed for research in cognitive developmental robotics. The Soundman head is a commercial product designed for the production of binaural recordings that maximize the perception of spatial effects. Our intention is to find out if the resonance of the iCub head, from the skull and interior components, significantly reduces the performance of ASR. Both platforms offer the possibility of extracting spatial cues from binaural sound, as such cues are produced by the geometric and material properties of the humanoid heads. A lateral view of both platforms can be seen in figure 1.

2.3 Biomimetic Sound Source Localisation

As a first step in our SSL architecture, a filter bank modelling the cochlear frequency decomposition [15] is used to reproduce the phase-locking mechanism of the organ of Corti. Signals reaching the ears are decomposed in several frequency components $f \in \{1, 2, \ldots, F\}$. In healthy young people, f's are logarithmically separated and respond to frequencies between ∼20 Hz and ∼20000 Hz [14]. As we are mainly interested in the localisation of speech signals, we constrain the set of f's to frequencies between 200 Hz and 4000 Hz. The MSO is modelled as a Jeffress coincidence detector, where ITDs are represented spatially by different neurons that fire at specific time delays [12]. The LSO model represent ILDs spatially as $\log(I_f/C_f)$, where I_f and C_f are sound pressure levels at the ipsilateral and contralateral microphones for a given f [12].

The following layer in the architecture models the IC, where the output from the MSO and the LSO is integrated. The MSO and LSO connection weights to

[3] `http://www.soundman.de/en/dummy-head/`

the IC are estimated using Bayesian inference [12]. A more detailed description of the method can be found in [6]. The output of the IC layer can reflect the non-linear effects of reverberation and the intense levels of ego-noise produced by the head's cooling system (∼60 dB).

In order to classify the IC output more robustly for controlling the motion of a robotic head, we add an additional layer with a multilayer perceptron (MLP) to the architecture [7]. This layer is necessary for improving the classification performance and does not model a particular region in the auditory pathway. Figure 2 shows the SSL system correctly classifying an incoming sound from the left side. The experiments reported in this paper are performed with the robot heads static. However, the performance of our neural SSL architecture is documented in [7] and compared with other statistical approaches.

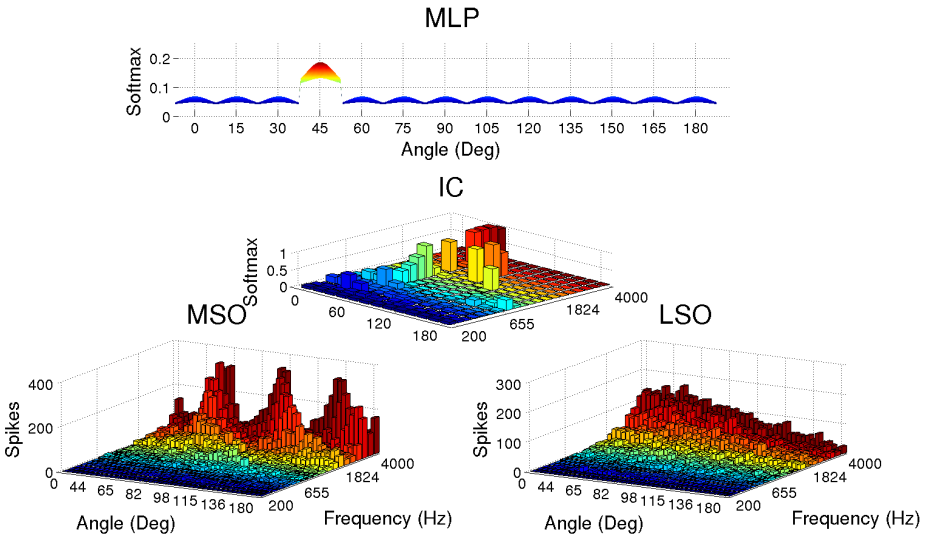

Fig. 2. Output of different layers in the SSL system for white noise presented at 45° on the azimuth. Even though many of the IC frequency components disagree on the sound source angle, the MLP is able to cope with these non-linearities and correctly classifies the IC output. Detailed results for static SSL using the iCub are given in [7]

2.4 Architecture for Automatic Speech Recognition

We selected the Google ASR engine [13] based on deep neural networks, as it was more robust against the ego-noise produced by the iCub than other popular ASR engines we tested. The speech corpus used in our experiments is the TIMIT core-test-set (CTS) [10], as it includes all the existing phoneme components in the English language. The CTS is formed by 192 sentences spoken by 24 different speakers: 16 male and 8 female pronouncing 8 sentences each. We map the output of Google ASR to the best matching sentence from the CTS. However,

this mapping does not impede generalization from our experimental results as we are only interested in measuring the performance of ASR with and without the support from an SSL system. Whenever a sound file is sent to the Google ASR engine, a list with the 10 most plausible sentences (G10) is returned. The post-processing consists of transforming the G10 and the CTS from *grapheme* to *phoneme* representation [4] and then computing the Levenshtein distance [11] between each of the generated phoneme sequences from the G10 and the CTS. Finally, the sentence in the CTS with the smallest distance to any of the sentences in the G10 is considered the winning result and it is considered a correct recognition when it is equal to the ground truth. We refer to this domain-specific ASR architecture as the DASR system.

In general, the Levenshtein distance $\mathcal{L}(\mathbf{A}, \mathbf{B})$ refers to the minimum number of *deletions*, *insertions* and *substitutions* required to convert string $\mathbf{A} = a_1 \dots a_m$ into string $\mathbf{B} = b_1 \dots b_n$. Formally, the distance $\mathcal{L}(\mathbf{A}, \mathbf{B}) = \mathbf{D}(m, n)$, where

$$\mathbf{D}(i,j) = \begin{cases} i & \text{for } \; 0 \le i \le m \; \text{ and } \; j = 0, \\ j & \text{for } \; 0 \le j \le n \; \text{ and } \; i = 0, \\ \min \begin{cases} \mathbf{D}(i-1,j)+1 \\ \mathbf{D}(i,j-1)+1 \\ \mathbf{D}(i-1,j-1)+\kappa \end{cases} & \text{for } \; 1 \le i \le m \; \text{ and } \; 1 \le j \le n. \end{cases}$$

Where $\kappa = 0$ if $a_i = b_j$ and $\kappa = 1$ if $a_i \neq b_j$.

3 Experimental Results and Discussion

In the following experiments we measure the performance of the DASR system when the speech is presented around the robotic heads from the loudspeakers at angles θ_{lspk} at $\sim$1.6 m from the robot. Let θ_{neck} be the angle faced by the robot at any given time, and δ_{diff} be the angular difference between θ_{lspk} and θ_{neck}. We hypothesise that there is an angle -or set of angles- δ_{best} for which the signal-to-noise ratio (SNR) is highest and hence, for which the DASR system performs better. For this purpose we measure the performance of the DASR system after reproducing 10 times the CTS utterances from the loudspeaker at angle θ_{lspk}. The performance is measured as the average success rate at the sentence-level for the entire CTS corpus over the 10 trials. We refer to success rate as to what *sentence accuracy* is in the ASR domain, the ratio of correct recognitions over the total number of recognitions. It is also desirable to visualize the results of this binary evaluation in a continuous domain. Such transformation is made by measuring the average Levenshtein distance between the output of the DASR system and the ground truth sentences.

It is important to remember that the sounds recorded through the robotic heads contain 2 channels, i.e. the audio waves from the left and right microphones. As the DASR system requires monaural files as input, there are 3 possible reduction procedures: Using the sound wave from the left channel only (LCh), using the sound wave from the right channel only (RCh) or averaging

the sound waves from both channels (LRCh). The average success rates of the 3 reduction procedures on the recordings obtained with both heads are shown in figure 3 and the average Levenshtein distances in figure 4. It is clear that the performance curves obtained from the recordings of both robotic heads follow the same patterns. Notice that the performance of the DASR system improves with the Soundman head on the most favourable angles δ_{best}. However, the difference is not large enough to conclude that the resonance of the iCub head reduces the performance of the DASR system.

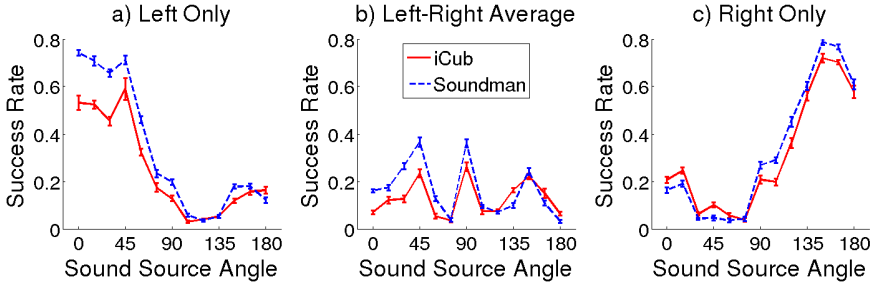

Fig. 3. Average success rates of the DASR system. Results obtained with both robotic heads for the frontal 180° on the azimuth plane.

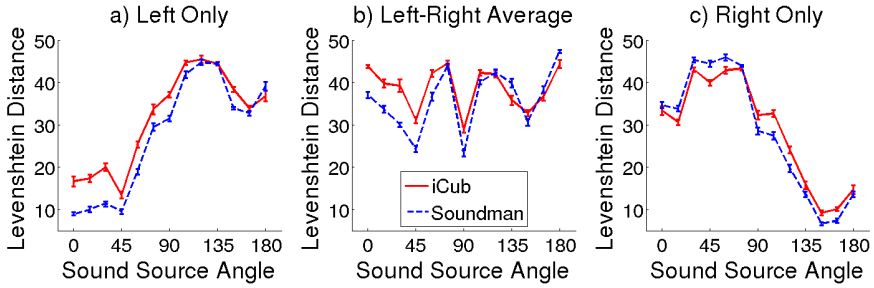

Fig. 4. Average Levenshtein distances between the DASR output and the ground truth. Results obtained with the both robotic heads for the frontal 180° on the azimuth plane.

Even though the volume of each loudspeaker was measured to output the same intensity level (∼75 dB), the smoothness of the performance curves is affected by the difference in fidelity from each of the loudspeakers. Nevertheless, the graphs clearly show a set of angles δ_{best} where the DASR system considerably improves its performance. For all reduction procedures with both robotic heads performance is best near $\delta_{best} \in \{45°, 150°\}$, where the robotic heads reduce minimally the SNR of incoming speech.

In the LRCh reduction, most sound source angles θ_{lspk} produce recordings where one channel has higher SNR than the other. Therefore, when both signals

are averaged the speech SNR will be reduced. The exceptions are sound sources at 90°, 45° and 150°. This can be explained by considering the moderate SNR both channels have in the case of 90°, and by considering the high and low SNR in the ipsilateral and contralateral signals in the case of 45° and 150°. It is also important to notice the magnitude of this effect, as the highest success rates from the LCh and RCh reductions are twice larger than the highest success rates from the LRCh reduction. This difference can be related to the strong shadowing from the geometry and material of the humanoid heads. The same effect can be seen in the LCh and RCh reductions alone. It is commonly assumed that speech SNR will increase when the sound source is in the front or parallel to the interaural axis, and when the input to an ASR system comes only from the channel closest to the sound source. However, the angles found to be optimal for our DASR system are counter intuitive, and the difference between the lowest and highest values in the LCh and RCh reductions is unexpectedly large.

The periodical shape in the LCh and RCh plots can be understood by considering the effect of the round shape of the heads and the position of the microphones. The pinnae are placed slightly behind the coronal plane. Therefore, the distance travelled by the sound waves from the sound source to the contralateral ear is maximal at approximately 45° and 135° instead of 0° and 180°. This explains the increase in performance before 135° for LCh and after 45° for RCh. On the other hand, the decrease in performance before 45° for LCh and after 135° for RCh can be produced by the shadowing of the pinnae and reverberation from the metal structure on the sides of the the VR setup.

4 Conclusion and Future Work

It became clear from the experimental results that robotic ASR systems can improve considerably their performance when supported by a parallel SSL system. Any humanoid robotic platform can be tested -or learn autonomously- to find the optimal angles for increasing the SNR of incoming signals. Such process can teach the robot to face a speaker in the optimal direction, in the same way people with auditory deficiencies do.

A natural extension of this work is to make the robot focus its attention in a single source of information from a possible multitude of concurrent stimuli. It is in this extended scenario where the input from other sensory modalities comes into play. Vision can be used to disambiguate the location of an addressing speaker between a crowd by observing the orientation of the torso, gaze and lips movement of each individual detected. Afterwards, this information can be used to perform auditory grouping in time and frequency domains in order to perform speech segregation in noisy environments [16]. This is the scope of current research by the authors towards multimodal speech recognition.

Acknowledgements. This work was supported by the DFG German Research Foundation (grant #1247) - International Research Training Group CINACS (Cross-modal Interaction in Natural and Artificial Cognitive Systems).

References

1. Asano, F., Goto, M., Itou, K., Asoh, H.: Real-time sound source localization and separation system and its application to automatic speech recognition. In: INTERSPEECH. pp. 1013–1016 (2001)
2. Bauer, J., Davila-Chacon, J., Strahl, E., Wermter, S.: Smoke and mirrors—virtual realities for sensor fusion experiments in biomimetic robotics. In: Intl. Conf. on Multisensor Fusion and Integration, MFI. pp. 114–119. IEEE (2012)
3. Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., Saltarén, R.: Design of the robot-cub (icub) head. In: Intl. Conf. on Robotics and Automation, ICRA. pp. 94–100. IEEE (2006)
4. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50(5), 434–451 (2008)
5. Cong-qing, L., Fang, W., Shi-jie, D., Li-xin, S., He, H., Li-ying, S.: A novel method of binaural sound localization based on dominant frequency separation. In: Intl. Cong. on Image and Signal Processing, CISP. pp. 1–4. IEEE (2009)
6. Davila-Chacon, J., Heinrich, S., Liu, J., Wermter, S.: Biomimetic binaural sound source localisation with ego-noise cancellation. In: Intl. Conf. on Artificial Neural Networks and Machine Learning, ICANN. pp. 239–246. Springer (2012)
7. Davila-Chacon, J., Magg, S., Liu, J., Wermter, S.: Neural and statistical processing of spatial cues for sound source localisation. In: Intl. Joint Conf. on Neural Networks, IJCNN. IEEE (2013)
8. Deleforge, A., Horaud, R.: The cocktail party robot: Sound source separation and localisation with an active binaural head. In: Proceedings of the international conference on Human-Robot Interaction. pp. 431–438. ACM/IEEE (2012)
9. Fréchette, M., Létourneau, D., Valin, J., Michaud, F.: Integration of sound source localization and separation to improve dialogue management on a robot. In: Intl. Conf. on Intelligent Robots and Systems, IROS. pp. 2358–2363. IEEE (2012)
10. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. NASA STI/Recon Technical Report N 93, 27403 (1993)
11. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. In: Soviet Physics Doklady. vol. 10, pp. 707–710 (1966)
12. Liu, J., Perez-Gonzalez, D., Rees, A., Erwin, H., Wermter, S.: A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation. Neurocomputing 74(1-3), 129–139 (2010)
13. Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M., Strope, B.: "your word is my command": Google search by voice: A case study. In: Advances in Speech Recognition, pp. 61–90. Springer (2010)
14. Schnupp, J., Nelken, I., King, A.: Auditory neuroscience: Making sense of sound. The MIT Press (2011)
15. Slaney, M.: An efficient implementation of the patterson-holdsworth auditory filter bank. Tech. rep., Apple Computer, Perception Group (1993)
16. Zion-Golumbic, E., Schroeder, C.E.: Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77(5), 980–991 (2013)