2012 2nd International Conference on Computer Design and Engineering (ICCDE 2012)
IPCSIT vol. 49 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V49.14

Simultaneous Recognition of Speech Commands by a Robot using a Small Microphone Array

Jose J. Lopez [1+], Maximo Cobos [2], Amparo Marti [1] and Emanuel Aguilera [1]

[1] Institute of Telecommunications and Multimedia App. Universitat Politècnica de València
[2] Departamento de Informática. Universitat de València

Abstract. This paper presents a speech processing system integrated into a mobile robot that enables the recognition of simultaneous speech commands. The proposed audio processing system employs a tetrahedral microphone array with a small inter-microphone distance that allows processing the speech signals produced by competing speakers independently, resulting in improved recognition accuracy. As opposed to other solutions based on large arrays, this tetrahedral structure can be easily integrated into a robot. To show the capabilities of this technology, both the microphone array and a real-time speech processing system have been integrated into a mobile robot platform developed by the authors. Several experiments have been carried out under different acoustic environments to evaluate the recognition accuracy of simultaneous speech commands using a vocabulary of 15 instructions. The results validate our proposed solution as a starting point for future improvements in the field of artificial audition systems for robots.

Keywords: Sensors and Signal Processing, Speech Recognition, Microphone Arrays, Speech Commands

1. Introduction

Automatic speech recognition (ASR) systems have been greatly improved in the last years. The advances in signal processing algorithms and the increased computational power of computers have been critical for this improvement. These features have also been applied to robotics, allowing autonomous robots to have advanced audition systems that provide them with a sophisticated human-machine interface. However, the classical difficulties of ASR, such as noise, interference and room effects (echoes and reverberation), are still problematic in modern speech applications. Array signal processing techniques can be used to minimize these problems and to reduce their negative effect on the speech recognition task. In this context, when ASR systems are installed in robots, these problems appear very commonly and it becomes necessary to implement and integrate these processing methods in real-time [1].

Another common problem appears when several people are speaking at the same time. This situation is widely known as the *cocktail-party problem* and it is especially important in the case of robots, since they must discriminate among subjects when they receive commands from more than one person simultaneously. The human auditory system is very good at focusing its attention to a single speech source in a mixture of several conversations. Several experiments have shown that this ability relies greatly on binaural audition, where the combination of the two ear signals in the brain results in an improved understanding of the target speech [2]. Therefore, incorporating this human ability into robots would be a desired and useful feature.

The recognition of two (or more) simultaneous speech sources is a challenging task. Today, most ASR systems have been designed to recognize only one speaker. Thus, if we want to make use of current ASR systems when this problem appears, the only solution is to separate competing speech signals by means of a sound source separation (SSS) algorithm. This problem has been recently tackled in [3], where a successful solution was provided by using an algorithm based on geometrical source separation (GSS), [4].

[+] Corresponding author. *E-mail address*: jjlopez@dcom.upv.es.

In this paper we propose to use a microphone array with a very small inter-microphone distance. As opposed to other reported solutions such as those based on beamforming, small arrays can be easily integrated into robot prototypes. Although small arrays have been extensively studied [5] in the last years, to the best of our knowledge, its application over robotic platforms has not previously discussed. In this context, several algorithms recently published by the authors have been used in this paper with the aim of providing a robot with a simultaneous speech recognition system.

2. System Overview

Before describing the complete signal processing system, it is interesting to present the robot used in our experiments. This robot has been completely developed and assembled by our research group and consists of 4 platforms, one above the other, containing its different electronic and mechanical subsystems, Fig. 1a. The robot's computer is based on a mini-ITX PC board with a compatible VIA x86 chip working over Windows XP. In the top there is the audition system of the robot based on a tetrahedral microphone, Fig 1b.

With this array, it is possible to estimate the Direction-Of-Arrival (DOA) of sound sources in the 3-D space and to separate the signals from each sound source by means of time-frequency processing techniques. It is also worth to mention that the array has been placed on top of the robot to avoid acoustic shadowing effects and to obtain a better visibility of the environment that surrounds the robot, including the individuals that interact with it by means of speech commands.

Figure 2 depicts the audition system of the robot, which includes the audio capturing system and several signal processing stages. The audio input is composed of the signals from the tetrahedral microphone array. This microphone array is used to capture different mixture signals corresponding to simultaneous speech commands.

The first processing stage is the DOA estimation algorithm. This stage receives the signals from the array and processes them in order to give an estimate of the directions corresponding to the different speakers. When using only two microphones, DOA estimation is usually performed via binaural localization cues. When a source is not located directly in front of the array, sound arrives slightly earlier in time at the microphone that is physically closer to the source, and with somewhat greater energy. This fact produces the interaural time difference (ITD) and the interaural intensity difference (IID) between the two sensors. DOA estimation methods based on. The DUET separation technique [6], which is also based on IID and ITD, can be used for estimating with high accuracy the TDOA of several sources in the time-frequency (TF) domain assuming that only one source is active in each TF point. The precise source localization technique used in this robot is a refinement of commented techniques that was developed by the authors in a previous work [7], where all the details are explained in detail.

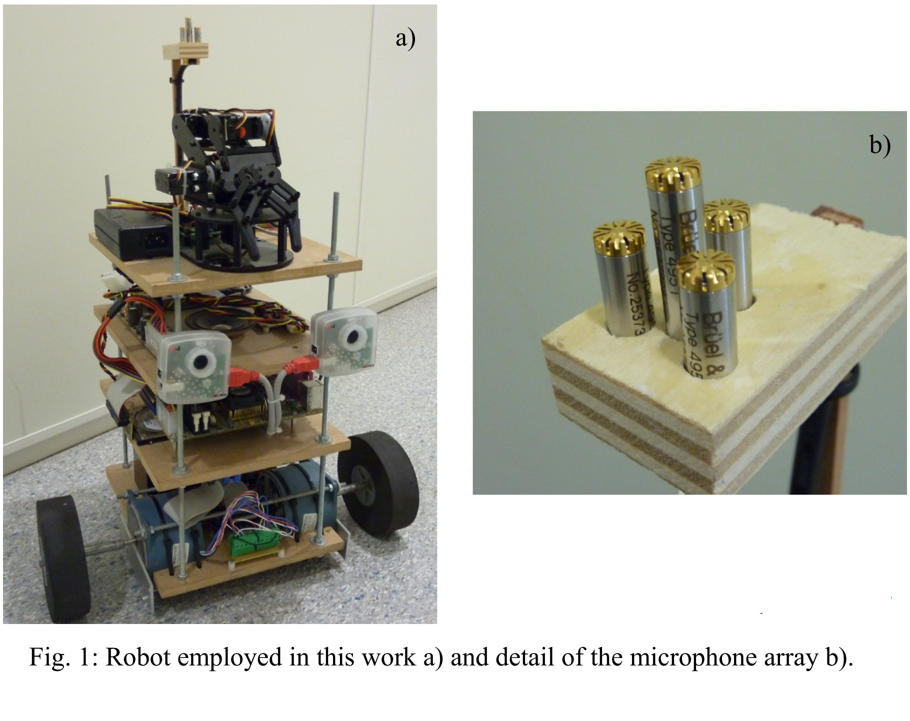

Fig. 1: Robot employed in this work a) and detail of the microphone array b).

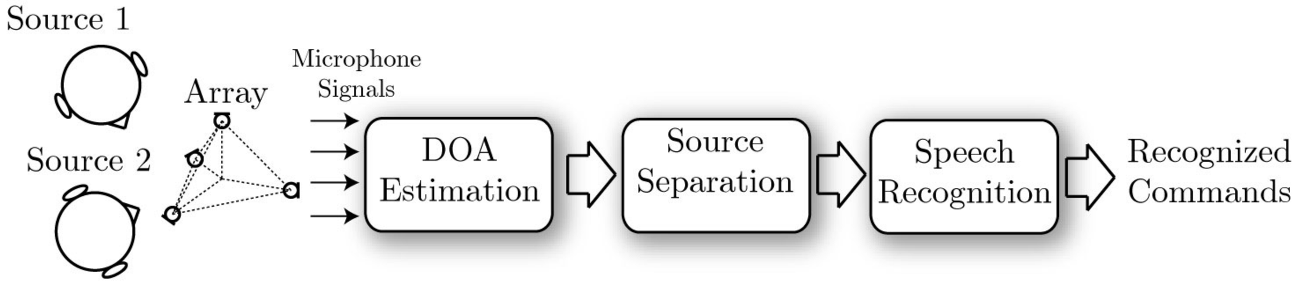

Fig. 2: Block diagram of the signal processing system.

Once the different individuals are localized, source separation is applied over the same microphone signals. The source separation algorithm is based on TF masking. Inspired by image segmentation techniques [8] separation is achieved by using a maximum interclass variance criterion between the angular distribution of the sources. With this criterion, it is possible to obtain a set of thresholds that divide the azimuth plane into angular sections corresponding to different speakers. Multilevel thresholding can be exploited to achieve fast separation in reverberant scenarios by identifying different angular areas wherein the speakers are located with a strong likelihood. All the details of the algorithm were published by the authors in [9].

The extracted speech sources are later used in the speech recognition stage. In our system, we decided to employ the HTK recognition toolkit, one of the most widely used toolkits by the robot research community [10]. The classification process in HTK is based on hidden Markov models (HMMs). Before performing the HMM processing, the audio data is parameterized into sequences of feature vectors using Mel Frequency Cepstral Coefficients (MFCCs). For each frame period (10 ms) it generates 39 coefficients (13 from the C0, 13 from the delta coefficients and 13 more from the acceleration coefficients).

Finally, after applying the speech recognition algorithm based on a limited grammar (as explained in Section 3) the recognized speech commands are managed by the robot according to a defined criterion.

3. Experiments

In order to test the capabilities of our robot to recognize simultaneous speech commands, a series of experiments were carried out in our laboratory.

3.1. The Command Set

A set of 15 speech commands were defined. These commands were short speech fragments of 2-3 words, as for example: *come here, turn around, go outside, raise your arm*, etc. Note that the actual commands were in Spanish. A command database was built by recording 20 different people (13 males, 7 females), which were only used as a recognition test set (Voxforge database [11] corpus was used to train the recognizer). These commands were recorded in optimal acoustic conditions (close microphone recording to avoid room effects). Each person repeated each speech command two times, obtaining a set of 600 commands. The commands were initially recorded with a sampling frequency of 48 kHz, however, all the signals were afterwards downsampled to 16 kHz to maintain the sampling frequency of the Voxforge corpus.

3.2. Experiment Set-Up

A series of four experiments was conducted. First, the recognition accuracy using the clean recorded commands previously described was evaluated as a reference for upper bound performance. For the rest of experiments the set-up shown in Figure 3 was used. Two loudspeakers were placed inside our recording studio with azimuth angles -30º and 20º with respect to the reference robot axis. Then, the impulse responses from each loudspeaker to each microphone were measured, which encode all the acoustic path effects, including attenuation, delay and reflections. Using these responses, it is possible to simulate the signals that the microphone array would acquire if a given speech command was reproduced over each of the loudspeakers. A set of 4000 mixtures made of randomly selected pairs of commands from different speakers was simulated in reverberant and anechoic conditions. Both the left and right microphone signals were obtained for each mixture.

Two different acoustic situations were studied. First, a reverberant case was simulated by using the recorded impulse responses. Then, an anechoic environment was considered with the aim of studying how the absence of room reflections improves the recognition accuracy. The anechoic case was simulated by properly delaying each command signal.

Once all the mixture signals were generated, the source separation algorithms described in Section 2 were applied with the aim of extracting the mixed speech commands, obtaining two separated signals from each of the 4000 mixtures in the two acoustic environments (reverberant and anechoic). These signals were the input of the speech recognition stage. Moreover, the recognition accuracy using the original clean recorded signals without mixing was also considered as an upper bound of the performance of the system.

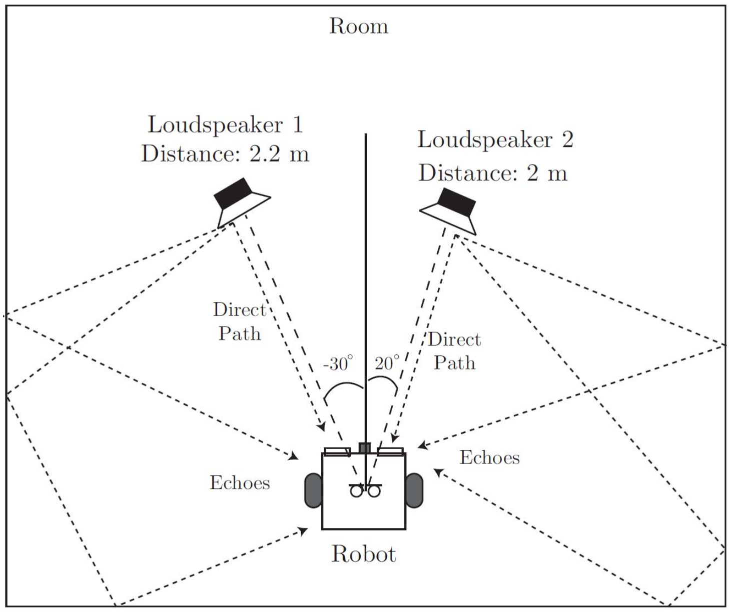

Fig. 3: Experiment set-up.

3.3. Results

Table 1 shows the recognition accuracy rounded to the closest integer. It can be observed that the recognition rate for the case of isolated speech commands is the highest, which demonstrates that the speech recognizer has been properly trained. The table also shows that the speech recognition rate is very poor if the source separation stage is omitted, being also a bit lower than the probability of recognizing any of the two simultaneous commands.

After applying the source separation algorithm for simultaneous commands in real environment with acoustic echoes, it is observed that the recognition rate is significantly improved (65%). For the simulated anechoic environment, the recognition rate improves to 76%. These results are promising and confirm the validity of the proposed system. However, the results can be considered as preliminary, since the speech recognition system has not been still improved for taking into account the properties of the speech recognition system. This opens new research lines, such as adapting the speech recognition toolkit both with separated speech and reverberant speech commands.

Table 1. Performance in Terms of Percentage of Correct Frame

Experiment	Recognition Accuracy
Isolated commands	98%
Reverberant without separation	18%
Reverberant with separation	65%
Anechoic with separation	76%

4. Conclusions

In this paper, we have presented a complete audition system for a mobile robot capable of recognizing speech commands from two simultaneous speakers. One of the major contributions of this work resides in the use of a small microphone array and time-frequency processing algorithms for DOA source localization and separation in difficult acoustic environments. The recognition accuracy with a complete robot audition system has been evaluated by means of several experiments. To this end, combinations extracted from a set of 15 short speech commands were considered to test the recognition rate of simultaneous instructions recorded from 20 different speakers. The experiments were designed to compare the recognition accuracy obtained in an ideal acoustic situation with the one reached with and without applying our separation approach.

The results showed that, without source separation, the recognition accuracy is extremely poor. However, the accuracy in the recognition stage can be substantially improved by applying the proposed method. In this context, a 75% recognition rate was obtained in the case of an echo-free room. For rooms with reverberation, a recognition rate of 65% was obtained. These results are promising, since accurate speech recognition in reverberant rooms is a difficult task even for non-mixed signals. However, some improvements are still needed for making more robust this processing system. In this context, further work will consider the properties of separated and reverberant speech to adapt the recognizer toolkit and increase its accuracy.

5. Acknowledgements

The Spanish Ministry of Science and Innovation supported this work under the project TEC-2009-14414-C03-01.

6. References

[1] .K. Nakadai, H. G. Okuno, and H. Kitano, “Real-time sound source localization and separation for robot audition,” in Proceedings of IEEE International Conference on Spoken Language Processing, 2002, pp. 193–196.

[2] A. W. Bronkhorst, “The cocktail party phenomenon: A review of research on speech intellibility in multiple-talker conditions,” Acustica, vol. 86, pp. 117–128, 2000.

[3] A. P. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, “Evaluating real-time audio localization algorithms for artificial audition in robotics,” in IROS, 2009, pp. 2033–2038.

[4] L. C. Parra and V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 6, pp. 352–362, 2002.

[5] M. Cobos, J. J. Lopez, and S. Spors, “Effects of room reverberation in source localization using small microphone arrays,” in 4th International Symposium on Communications, Control and Signal Processing (ISCCSP 2010), Limassol, Cyprus, March 2010

[6] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830–1847, July 2004.

[7] M. Cobos, J.J. López, D. Martinez, “Two-Microphones Multiple Speaker Localization Based on a Laplacian Mixture Model”, Digital Signal Processing (Elsevier), 21(1), pp 66-76, 2011

[8] N. Otsu, “A threshold selection method from graylevel histogram,” IEEE Transactions on System Man Cybernetics, vol. SMC-9, no. 1, pp. 62–66, 1979.

[9] M. Cobos and J. J. Lopez, “Two-microphone separation of multiple speakers based on interclass variance maximization,” Journal of the Acoustical Society of America, vol. 127, pp. 1661–1673, 2010.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.2) (Cambridge University Engineering Department, Cambridge, UK, 2002), pp. 14–23.

[11] “Voxforge homepage”, http://www.voxforge.org/.