Identification and Localization of One or Two Concurrent Speakers in a Binaural Robotic Context

Karim Youssef*, Katsutoshi Itoyama and Kazuyoshi Yoshii
Graduate School of Informatics
Kyoto University
Kyoto, Japan

***Abstract*—This paper presents a method of identification and azimuth estimation for one or two concurrent speakers in simultaneous utterances. This method is applicable to human-machine interaction and robot audition. Identification and localization have been rarely mutually addressed and related works rely on time-frequency exploitation strategies to extract and treat each source's contribution to the received signal. The presented method relies on a training made with one speaker at a time, but it can exploit a speech segment to identify and localize two speakers. A cochlear filtering-based binaural front-end allows to extract equivalent rectangular bandwidth frequency cepstral coefficients (ERBFCC) and interaural level difference (ILD) features. Artificial neural networks (ANNs) exploit ERBFCCs to provide identity information, and a histogram-based exploitation of ILDs provides azimuth angle information. The method was evaluated in contexts including overlapping segments in the presence of noises and sound reflections and its efficiency was demonstrated. Even with fully overlapping utterances, we reached an 83% identification rate of both speakers, an 82% estimation accuracy of both azimuths and an 68% correct mutual identity and azimuth estimation rate. At least one speaker was correctly identified and localized in more than 99% of the tests for utterances lasting near 5s.**

***Index Terms*—Speaker identification, localization, binaural inputs, human-machine interaction, robot audition.**

I. Introduction

Speech is an important component of human-machine interaction and robot audition applications [1], [2]. Indeed, it carries more information that a machine can reach and exploit than just the pronounced words. Such information are the speaker identity, position, prosody and the environment acoustic conditions for example. A machine or robot in sound-based interaction with humans is constrained by the environment acoustic properties, implying the presence of noise and sound reflections. In the presence of multiple persons, an additional constraint can arise when two or more speakers utter simultaneously. The robot has to identify and localize the speakers in order to decide which one or ones to interact with, which can be followed by adjusting its behavior accordingly. Among the different paradigms of sound acquisition, binaurality is particularly interesting and challenging. Indeed, it exploits signals not only acquired with only two microphones, which is a relatively low number for certain applications, but also acquired similarly with the human sound sensing. Thus, it allows to reduce material loads and to implement or propose models of human auditory processing stages.

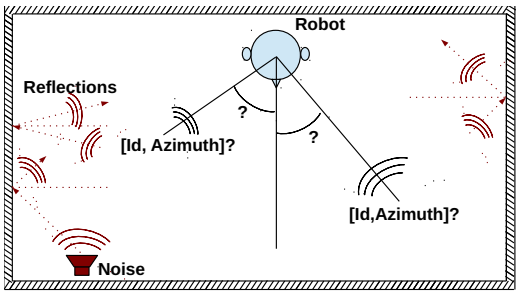

Fig. 1. Objects: to know the identities and azimuth angles of two speakers based on their voices, in the presence of sound reflections and noises.

In this paper, we mutually address the identification and localization of speakers in the binaural paradigm. We regard identification in this work as the task of estimating speaker identities among a set of speakers known by the system, and regardless of the words being pronounced. Similarly, localization in this paper addresses the estimation of speaker azimuth angles among a set of possible angles known by the system. The proposed method operates in scenarios where two speakers utter speech at the same time (see Figure 1), which is challenging for systems operating with only two input signals. With two speakers, overlapping speech segments complicate the task as they originate from two different speakers and positions and are received together. Previous work thus focused on separating the contribution of each speaker from the resulting signal to extract related identity and position information. We address the problem from a different point of view where we associate the most likely identity and position to each short speech time frame. Results are then integrated over an overall duration with a number of included frames to obtain the speakers' identities and positions.

The binaural paradigm was exploited in [3], where localization and tracking of an unknown number of speakers used differences of sound arrival times to the ears. Also relying on binaural sound acquisition, we add the identification module and extract different acoustic features, through a common binaural front-end to both modules. Equivalent rectangular

K. Youssef is an International Research Fellow of the Japan Society for the Promotion of Science.

978-1-4799-8697-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SMC.2015.82

bandwidth frequency cepstral coefficients (ERBFCC) and interaural level difference (ILD) acoustic features are extracted for identification and localization respectively. Both features rely on gammatone filterbanks emulating the cochlear filtering. Identification uses artificial neural networks (ANNs) to exploit the computed ERBFCCs and output speaker identities. Speaker azimuths are estimated through frequency and azimuth-dependent histograms of ILDs, established in a training phase. The proposed system was evaluated with data extracted in conditions implying the presence of sound reflections and noises. Different performance measures were computed, taking into account the utterance overlap and duration.

The paper is organized as follows. The next section overviews previous work that addressed similar tasks. Section III presents the system's consecutive steps and operation. Section IV presents the evaluation data, performed tests and results. Finally, Section V concludes the paper.

II. Related work

Sound signal processing was used in several applications of human-robot interaction. Multiple paradigms were conceived for sound acquisition and treatment, and proved the usefulness of sound as a means of communication. In [1], a quizmater robot relying on a microphone array and dealing with simultaneous speech utterances, was able to identify multiple speakers and recognize their respective speech segments. Also in [4], the robot audition system dealt with simultaneous speech utterances. It performed recognition, localization and separation, and estimated the number of speakerss. Moreover, in [2], interaction between a robot and multiple persons was tackled. The robot localized speakers, selected an interaction partner and acted accordingly. Binaural robot audition was addressed in [3], where the robot was able to localize and track multiple speakers with no *a priori* information about their number.

Whether used in human-machine interaction or other applications, speaker identification and localization were largely tackled in previous work, mostly separately and with a single active speaker. The previously proposed methods consist in acoustic feature extraction from the acquired signals, and exploitation to reach the requested information. Some of the previous studies proposed models for differend stages of the human auditory processing, from the cochlea to the haircells and even to upper stages. Additionally, some works tackled the problems of sound reflections and noises and proposed methods to suppress their negative effects. In the following, we separately exhibit previous work regarding identification and localization.

Sound source localization, applicable to human speakers localization, has been largely tackled in the binaural context. The presence of a source at a certain azimuth angle causes the emitted sound to be sensed at one ear before the other and more loudly. Thus, the interaural level and time differences (ILD and ITD) reflect the azimuth angle. Previous studies used them [5], [6], with machine learning exploitations techniques [6], or straightforward relations between the cues and the azimuth through mathematical equations [7]. Interaural acoustic features can either be extracted independently of the frequency, or on different frequency intervals. The second option allows to exploit these cues with relation to their frequency-dependent efficiencies. Indeed, the Duplex theory [8] suggests that ITDs are used at lower frequencies, while ILDs are exploited at higher frequencies.

As for the speaker identification, previous studies extracted features like mel-frequency cepstral coefficients (MFCCs) [9], [10], lineat predictive coding (LPC) [9] and i-vectors [11]. Feature exploitation relied on techniques like gaussian mixture models (GMMs) [12], support vector machines (SVMs) [13] and deep neural networks (DNNs) [14]. It is to note that most of the previous speaker identification studies operated in a single-microphone context. Works operating with microphone arrays reach a final identification through combination of the different identification results or exploitation of the different signals to reach a single better signal and then performing feature extraction and exploitation on it.

Localization and identification have been addressed together in [15]. This work presented a system performing localization followed by speech activity detection and identification. The estimated direction allowed to select the closest ear and to extract identification features from the corresonding signal. Speech segregation, and localization were tackled in [16], [17], while identification was not directly addressed. Nevertheless, segregation can provide information to be based on for identification cue extraction, using missing data techniques for example.

III. Proposed method

The proposed system is designed for sound-based human-machine interaction and robot audition applications. A robot dealing with a human or more is in a constant need of knowing the identity and position of its partners, in order to keep track on their activity and detect outliers. The system extracts different acoustic features for localization and identification of speakers, but relying on the same front-end, which is beneficial in a computational point of view. As specified earlier, the object is to know the identities and azimuth angles of two speakers uttering at the same time, relying on two signals only.

Previous systems operating with multiple sources, aiming for source separation or localization for example, rely on the sparseness assumption. Exploiting the spectral contents of the acquired signals and assigning time-frequency units to specific sources was used for this task. The system we present is conceived to be trained with speech signals corresponding to a single speaker, and to be able to deliver results if two speakers are concurrently active. Adopting a small frame duration and a large number of cochlear filtering channels increases the resolution of the system and thus its ability to extract the most active speaker at each time frame and frequency bin. Each time frame is assigned to the most likely speaker and azimuth, without decomposition of the frame into multiple frequency bins, but using the most reliable bins. Speaker identification relies on ERBFCCs exploited by ANNs, and localization uses

ILD histogram-based likelihood estimations to estimate the most likely speaker and position respectively.

A. Feature extraction

Features are extracted on time frames of approximately 15ms each, extracted with a frame overlap of approximately 7.5ms. But before feature extraction, an energy-based voice activity detection (VAD) allows to eliminate the signal portions with no speech and thus no identity or azimuth information. A thresholding can be used for this VAD to discard silence portions. The energy of each frame is computed based on its original temporal waveform, independently of the frequency, and compared to the energy threshold E_{th}:

$$E_{th} = E_{min} + C(E_{max} - E_{min}) \quad (1)$$

where E_{min} and E_{max} are respectively the maximal and minimal frame energies across the analyzed segment. C is a coefficient used to adjust the threshold for more or less severe frame selection (currently, $C = 0.02$). This process is applied to the left and right ear signals independently but only time frames that have commonly left and right energies higher than the respective thresholds are used.

Both identification and localization acoustic features are extracted based on the outputs of left and right-ear gammatone filterbanks [18], [19] of N_f filters each, exploiting the inputs of the left and right ears (currently, $N_f = 100$). Let $E^l_{t,f}$ and $E^r_{t,f}$ respectively the output energies of the left-ear and right-ear t^{th} frames and f^{th} gammatone filters. ILDs and ERBFCCs are computed as follows.

1) ILD: The ILD reflects the difference in loudness perceived by the two ears, due to the presence of the sound source to the side of one ear, and the presence of the head as an obstacle to the sound wave propagation. This causes the creation of a shadow zone containing the contralateral ear and in which the energy is reduced. This head shadow effect [20] depends on the size and shape of the head and on the source azimuth. For the f^{th} gammatone filter and t^{th} frame, the ILD $\delta_{t,f}$ can be computed as:

$$\delta_{t,f} = 20\log_{10}\frac{E^l_{t,f}}{E^r_{t,f}}. \quad (2)$$

2) ERBFCC: These coefficients are conceived in a similar way with the widely used MFCCs. The mel-scale triangular filterbanks used in MFCC computation are replaced by ERB-scale gammatone filterbanks, which provide a better model for the cochlear filtering. Moreover, the same filterbanks used for ILD computation are used with ERBFCCs, which is advantageous from a computational point of view. The k^{th} ERBFCC coefficient at the left ear, $\gamma^l_{t,k}$, is obtained through discrete cosine transform:

$$\gamma^l_{t,k} = w_k \sum_{f=1}^{N_f} \log_{10}(E^l_{t,f})\cos\Big(\frac{\pi(2f-1)(k-1)}{2N_f}\Big), \quad (3)$$

$$w_k = \begin{cases} \frac{1}{\sqrt{N_f}}, & k = 1 \\ \sqrt{\frac{2}{N_f}}, & k > 1 \end{cases}.$$

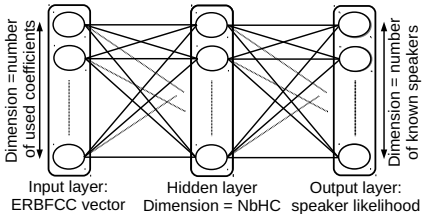

Fig. 2. ANN-based identification feature exploitation.

ERBFCCs are computed at the right ear in a similar way. They are assembled into left and right ear vectors before their exploitation by ANNs:

$$\Gamma^l_t = [\gamma^l_{t,2}, .., \gamma^l_{t,K}]^T, \quad (4)$$

and

$$\Gamma^r_t = [\gamma^r_{t,2}, .., \gamma^r_{t,K}]^T, \quad (5)$$

where K (currently, $K = 40$) is the index of the last cepstral coefficient taken into account at both ears, and a vector v^T is the transpose of a vector v.

Note that gammatone filterbanks have also been used in [21], [22] for speaker identification through the gammatone frequency cepstral coefficients.

B. Feature exploitation

The computed features are exploited in a way to obtain classifications over sets of possible speakers and azimuth angles. These classifications are obtained based on likelihood esttimations performed by ANNs and histograms respectively.

1) Identification ANNs: One feed-forward multi-layer perceptron with sigmoid activation and one hidden layer is associated to each ear. As shown in Figure 2, each ANN takes as input the corresponding ear's ERBFCC vectors. The number of hidden cells is $NbHC = 40$ and the number of output cells is equal to the number of speakers known by the system. The taining associates a constant positive value α, with $0 < \alpha < 1$, to the output cell number s if the current training input vector corresponds to the speaker number s, and the value $-\alpha$ to the other output cells. The testing provides a set of values at the output cells for each testing input vector. The identification result for each frame is thus the speaker associated to the cell with the highest activity.

Left and right-ear ANNs provide their outputs separately. We combine the two outputs corresponding to each time frame in a way to consider only the frames with the same result as valid, and discard the frames with different results. The object of this operation is to eliminate part of the monaural wrong outputs through a binaural exploitation which can improve the performances of the system.

2) ILD histograms: ILDs are first computed through all the training data according to Equation 2, for the gammatone filters associated to the gammatone channels where ILDs are the most reliable, in the ensemble f_{rel}. The ILD channel selection and extraction of f_{rel} will be shown in the next paragraph. After computation, the interval between the overall minimal

and maximal ILD values is decomposed into N_b equal bins. Data are decomposed into groups where each group collects the ILDs computed at each azimuth angle and reliable channel from all the speakers. Thus, the number of groups is equal to the product of the number of possible azimuths and the number of reliable channels. The histogram $h_{az,f}, f \in f_{rel}$, corresponding to each group is then formed and normalized to provide ILD probabilities $p(\delta_{t,f}|h_{az,f})$. After histograms are formed, the testing goes as follows. For each time frame t, the ILDs $\delta_{t,f}$ are computed for the frequencies $f \in f_{rel}$. Each ILD value is submitted to the corresponding histogram of each azimuth direction for likelihood estimation. According to each candidate azimuth angle, the log-likelihood can be integrated over all the frequency channels taken into consideration:

$$L_{az,t} = \sum_{f \in f_{rel}} \log_{10} p(\delta_{t,f}|h_{az,f}), \tag{6}$$

and assuming that all azimuth angles are equally probable, and for each test's specific features at a time, the estimated azimuth $\hat{az}$ can be obtained through:

$$\hat{az}_t = \arg\max_{az} L_{az,t}. \tag{7}$$

3) ILD channel selection: ILDs computed at different gammatone channels do not have the same effectiveness in providing azimuth information. For example, the Duplex theory [8] suggests that ILDs are useful starting from a certain frequency, which depends on the head size. Moreover, and due to the presence of sound reflections and noises, certain frequency bands might be more or less affected and thus less or more able to provide efficient ILDs. A selection of the most reliable ILDs is thus needed. To this purpose, we use the monodimensional Wilks' Lambda [23], [24]. ILDs computed at a certain channel and azimuth angle for a number of time frames can be assembled into a frequency and azimuth-specific group. For each channel, it is of interest to have azimuth groups that are easily distinguishable, which means that ILDs provide good azimuth information. Lambda compares the intra-group dispersion to the overall dispersion. We can calculate this measure for each gammatone channel's ILDs as:

$$\lambda = \frac{\frac{1}{N}\sum_{g=1}^{G} n_g v_g}{V}, \tag{8}$$

where N is the total number of examples across groups. n_g and v_g are respectively the number of examples and their variance in the group of the azimuth g. V is the variance of the whole set of values computed over all the G groups. The value of λ does not reflect a certain degree of system performance. But the smaller it is, the better the feature component quality is because groups become tighter with less chance of confusion between them. Lambdas λ_f^{ILD} are computed for the ILDs at all the channels. These ILDs are computed on a setup database including several speakers and different azimuth positions. The reliable channel ensemble f_{rel} is then defined as follows:

$$f_{rel} = \{f|\lambda_f^{ILD} < 0.3\}, \tag{9}$$

comprising the channels corresponding to the most informative ILDs as set by this thresholding process.

C. System operation

The system operation takes place in scenarios like the following. Two speakers address speech to the robot, and sound reflections and noises interfere to the speech signals. Located at different azimuth angles, the speakers' signals are exploited by the system to estimate their respective azimuths and identities. The object is not to separate the speech signals, i.e. assign a set of frequency bins to each speaker at each time unit, but to know the speakers identities and azimuth angles. The system provides for each time frame a single identity and a single azimuth output based on the most likely speaker and azimuth. The consecutive frame results are then exploited to provide the final results over the entire mutual utterances duration. The two identities and azimuths being the most estimated over the whole test duration are then taken as test results. Tests are made in a text-independent context as the speech segments used for testing are not included in the training datasets.

IV. Evaluation data and results

In this section, we present the performed evaluations and their results. Evaluation data are extracted from recordings made in an environment with sound reflections and noises. We present the database, and then the tests and results.

A. Recorded database

Speech from 10 male speakers was extracted from the TSP Speech Database [25], and emitted through a loudspeaker to be received by the SIG2 humanoid robot's binaural microphones. Indeed, SIG2 has a human-like head with human-like ears and microphones placed inside them. The speaker-receiver distance was set to 1.5m, and 25 azimuth angles were considered, ranging between -60° and 60° with a step of 5°. Sound signals were sampled at 48kHz. The robot was placed as shown in Figure 3. The environment had a length of 14.3m, a width of 6.3m and a height of 6.9m. Three of the walls and the ceiling were of brick and concrete, one of the walls was covered by curtains and the floor was covered by carpets. Noise in the environment was mainly caused by a fridge. See Figure 3 for more details about the environment geometry. Such a database is useful for applications like speaker identification, speech recognition, and speaker localization in binaural or monaural contexts, it is intended to be made publicly available.

B. Tests and results

In the presented database, the utterances of each speaker were recorded independently of other speakers. The performed tests were made with signals obtained by adding two utterances at a time. Test utterances were generated randomly, each consisted of two different speakers located at two different azimuths with at least a 30° difference between them. In the performed tests, as previously seen, the system made classifications over sets of speakers and azimuth angles. The performances could thus be evaluated through the rates of the number of tests with correct classifications to the total numbers of tests. The following measures were computed to evaluate the effectiveness:

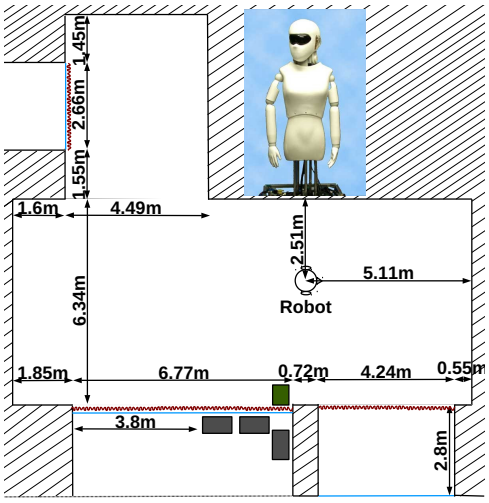

Fig. 3. Environment geometry and SIG2 robot head position and orientation. Blue: glass, red:curtains, green: sound-causing refrigerator, grey: air conditioner radiators.

- The rate of correct identification of both speakers, denoted in the results figures by "2 spk".
- The rate of correct azimuth estimates of both speakers, denoted by "2 az".
- The rate of correct identification and localization of both speakers, denoted by "2 spk + 2 az".
- The rate of correct identification of at least one of the two speakers, denoted by "1 spk".
- The rate of correct estimation of one of at least one of the two azimuth angles, denoted by "1 az".
- The rate of correct estimation of at least one identity and one azimuth at a time, denoted by "1 spk + 1 az".

The tests addressed the effect of the utterance overlap and duration on the performances. Each measure was computed based on a total number of random tests that was between 450 and 500. We detail the evaluations next.

1) Effect of utterance overlap percentage: In these tests, the signals of the two speakers were considered with different overlap percentages. The duration of each speaker's utterance is 2 seconds. The signals were considered in such a way that for an overlap of $\alpha\%$, the second speaker's utterance began when $(100-\alpha)\%$ of the first speaker's utterance was finished. Thus, the overlap percentage was computed in function of the utterance duration, and not directly in function of the speech activity inside it. Indeed, the utterances included silence portions which might more or less have reduced the actual speech activity duration inside them. Silence portions were removed by an energy-based VAD as specified in III-A, but on the resulting signal. Three percentages were considered: 30%, 50% and full overlap. In the last case, both utterances started and ended together.

Results are reported in Figure 4. The figure shows that at least one speaker is accurately identified and localized in all conditions. Indeed, the success rate of identifying or localizing at least one speaker was at least near 90%, even with full

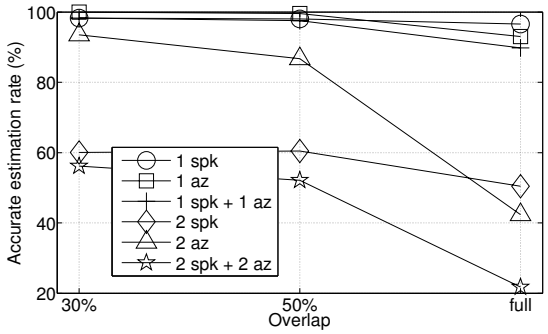

Fig. 4. Performance evaluation in function of the overlap percentage. The plotted measures are as explained in IV-B. Each speaker's utterance was of 2 seconds.

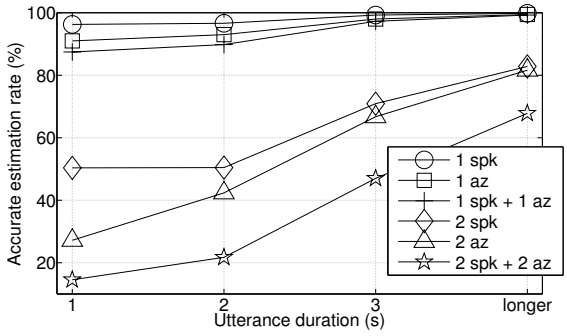

Fig. 5. Performance evaluation in function of the utterance duration. The plotted measures are as explained in IV-B. The utterances of both speakers were totally overlapping.

overlap. The performance measures significantly decreased when the overlap percentage increased. For less overlap, the localization performances were higher than the identification, while identification rates became higher than localization rates with full overlap. This shows a higher sensitivity of the localization module to the overlap. For low overlap, the system's ability to localize and identify both speakers at the same time depended more on the identification as the azimuths were estimated with a much higher accuracy. For high overlap, the two-speaker identification and localization performances depended on both modules with none of them being able to achieve perfect classification rates.

2) Effect of utterance duration: In this test, the two utterances of both speakers started and ended at the same time with full overlap. This is thus the hardest testing context as shown in the previous evaluation. We studied the effect of the utterance duration on the performances of the system. Four durations were considered: 1s, 2s, 3s and longer. In the last case, the smallest utterance duration betwen the two being uttered was considered, the average test duration was near 5s.

Figure 5 shows the results. They show that increasing the utterance duration improved the performances of the system. Indeed, the results were integrated over a larger number of

frames and could take benefit of more available information. The system showed high performances in the localization or identification of at least one speaker even for relatively short utterances. For two speakers at the same time, we can see again that the localization had lower classification rates than the identification, having fully overlapping utterances. Nevertheless, the performances of both modules increased with increasing utterance duration, which also increased the ability to simultaneously localize and identify both speakers. In approximately 68% of the tests, the system was able to perform these tasks when the speakers uttered for the tests with more than 3 seconds. Reducing the overlap percentage with such durations is able to improve the performances as shown in the previous paragraph.

In all the cases, a large number of the azimuth estimation errors were caused by confusions with the azimuth angles neighbouring the actual ones, as the resolution of the database in terms of azimuth was of 5°. If the application can bear to estimate the azimuth within a 5° range, the computed azimuth estimation rates and thus overall performances would highly increase. Finally, note that a functioning mode exploiting ILD vectors with ANNs was also tested. It was not as beneficial as the proposed histogram exploitation for two concurrent speakers, despite being efficient in scenarios with one speaker.

V. Conclusions

A method for mutual localization and identification of one or two concurrent speakers in human-machine interaction and robot audition contexts has been presented. The system adopts supervised learning that is based on data provided by a single speaker, but is able to deliver identities and azimuth angles of two speakers in a speech segment if present. Evaluations addressed the effect of the test duration and utterance overlap on the performances, in echoic and noisy conditions. They showed that the system's outputs become more accurate when the utterance duration increases, and when there is less overlap between utterances of two speakers. In all the cases, the system is able to identify or localize at least one of the two speakers with high accuracy. Current and future work is addressing the presence of more than two speakers at the same time, and speaker movement. Strategies to improve the robustness of the approach are being implemented, such as an integration of framewise results that allows to limit the numbers of candidate speakers and azimuths at each time frame based on previous frames results and their coherence.

References

[1] I. Nishimuta, K. Yoshii, K. Itoyama, and H. G. Okuno, "Development of a Robot Quizmaster with Auditory Functions for Speech-based Multiparty Interaction," *IEEE/SICE International Symposium on System Integration*, 2014.

[2] T. Tasaki, T. Ogata, and H. G. Okuno, "The interaction between a robot and multiple people based on spatially mapping of friendliness and motion parameters," *Advanced Robotics*, vol. 28, no. 1, 2013.

[3] U.-H. Kim and H. G. Okuno, "Improved binaural sound localization and tracking for unknown time-varying number of speakers," *Advanced Robotics*, vol. 27, no. 15, 2013.

[4] K. Nakamura, K. Nakadai, and H. G. Okuno, "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," *Advanced Robotics*, vol. 27, no. 12, 2013.

[5] J. Woodruff and D. Wang, "Binaural Localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, 2012.

[6] K. Youssef, S. Argentieri, and J.-L. Zarader, "A learning-based approach to robust binaural sound localization," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013.

[7] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ild and itd," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 1, 2010.

[8] L. Rayleigh, "On our perception of sound direction," *Philosophical magazine*, vol. 13, no. 74, pp. 214–232, 1907.

[9] S. Farah and A. Shamim, "Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization," *IEEE International Conference on Computer, Control and Communication*, 2013.

[10] K. Youssef, S. Argentieri, and J.-L. Zarader, "From Monaural to Binaural Speaker Recognition for Humanoid Robots," in *IEEE-RAS International Conference on Humanoid Robots*, pp. 580 – 586, Dec. 2010.

[11] M. McLaren and D. van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, 2012.

[12] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Transactions on Audio, Speech and Language processing*, vol. 20, no. 4, 2012.

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, 2011.

[14] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small foortprint text-dependent speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[15] T. May, S. van de Par, and A. Kohlrausch, "A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 7, 2012.

[16] J. Woodruff and D. Wang, "Binaural Detection, Localization, and Segregation in Reverberant Environments Based on Joint Pitch and Azimuth Cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, 2012.

[17] N. Roman, D. Wang, and G. G. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, october 2003.

[18] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex Sounds and Auditory Images," in *International Symposium on Hearing, Auditory physiology and perception*, pp. 429–446, 1992.

[19] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," tech. rep., Apple Computer, 1993.

[20] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner, "A probabilistic model for binaural sound localization," *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, vol. 36, October 2006.

[21] X. Zhao, Y. Shao, and D. Wang, "CASA-Based Robust Speaker Identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, 2012.

[22] X. Zhao, Y. Wang, and D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, 2014.

[23] A. El Ouardighi, A. El Akadi, and A. Aboutajdine, "Feature selection on supervised classification using wilk's lambda statistic," *International Symposium on Computational Intelligence and Intelligent Informatics*, March 2007.

[24] L. Lebart, M. Piron, and A. Morineau, *Statistique exploratoire multidimensionnelle, visualisation et inférence en fouille de données*. Dunod, 2008.

[25] P. Kabal, "Tsp speech database," tech. rep., Department of Electrical & Computer Engineering, McGill University, 2002.