Paper:

Simultaneous Identification and Localization of Still and Mobile Speakers Based on Binaural Robot Audition

Karim Youssef, Katsutoshi Itoyama, and Kazuyoshi Yoshii

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
E-mail: karim.youssef987@gmail.com
[Received July 31, 2016; accepted December 19, 2016]

This paper jointly addresses the tasks of speaker identification and localization with binaural signals. The proposed system operates in noisy and echoic environments and involves limited computations. It demonstrates that a simultaneous identification and localization operation can benefit from a common signal processing front end for feature extraction. Moreover, a joint exploitation of the identity and position estimation outputs allows the outputs to limit each other's errors. Equivalent rectangular bandwidth frequency cepstral coefficients (ERBFCC) and interaural level differences (ILD) are extracted. These acoustic features are respectively used for speaker identity and azimuth estimation through artificial neural networks (ANNs). The system was evaluated in simulated and real environments, with still and mobile speakers. Results demonstrate its ability to produce accurate estimations in the presence of noises and reflections. Moreover, the advantage of the binaural context over the monaural context for speaker identification is shown.

Keywords: robot audition, binaural acoustic features, cepstral features, azimuth estimation, speaker identification

1. Introduction

Auditory scene analysis (ASA) is essential in human communication. Our human auditory capabilities allow us to identify speech and speakers, localize and separate different sound sources or speakers, and attend to a single conversation in the presence of noise. Computational auditory scene analysis (CASA) studies the general framework of sound processing and understanding [1]. It undertakes tasks like sound processing and de-noising (speech, music, and environmental sounds), sound source localization, separation and identification. A growing field of CASA research is binaural computer audition [2–5]. Relying on signals acquired inside a human-like head and ears, it attempts to create a computational reproduction of the human auditory system stages. Human audition is, however, not yet fully understood and binaural artificial audition allows for advances in this field through proposing and testing hypotheses. Aside from its biological inspiration, the binaural context allows for a reduction in hardware, using two microphones for tasks that usually require larger microphone arrays. This context notably targets robot audition, which is important for robot operations involving human interaction [1, 6, 7]. A hearing robot can localize speakers, obtain characteristics of the acoustic scene, and recognize speaker identities and utterances, for example. These tasks are essential for the operation of a socially-interactive robot, which reproduces human auditory capabilities without necessarily relying on the exact computational models of the auditory processing stages.

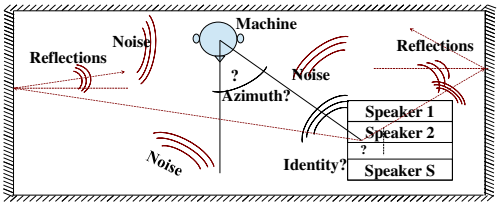

Fig. 1. Context and objects of the proposed system: identification and localization in the presence of noises and sound reflections, using binaural inputs.

In this paper, simultaneous speaker identification and localization are addressed in the binaural context as described in **Fig. 1**. The speaker identity estimation thus uses sound only in this context. Similarly, localization relies on acoustic features to estimate the speaker's position, and specifically the azimuth angle, in the current work. Both tasks provide necessary information for a human-robot interaction system. Performing them at the same time allows for the use of a common front end for feature extraction and for mutual exploitation of the outputs, reducing the number of computational steps and improving performance. The system operates in a closed-set fashion, dealing with a limited number of known candidate speakers, in contexts like aiding elderly persons or providing assistance in workplaces.

The acoustic feature extraction in this work relies on a cochlear filtering-based framework. Gammatone filterbanks with filters regularly spaced on the equivalent rectangular bandwidth (ERB) scale are used for their

proved efficiency. They replace the commonly used triangular filters in the mel-frequency cepstral coefficients (MFCC) computation and the outperforming ERB frequency cepstral coefficients (ERBFCC) features are extracted as identity features. The same filterbank outputs are used for the azimuth-related interaural level difference (ILD) feature computation. The features are exploited by artificial neural networks (ANNs) which are more advantageous in this context than generative models, like Gaussian mixture models (GMMs). They imply relatively low complexity as one ANN can be assigned to all of the speakers and positions for identification or localization. This ANN framework also implies low training data requirements and provides reliable generalization capabilities. A combination of the outputs of both the identification and localization modules is proposed, improving their performance. The system was evaluated in simulated and real environments ranging between noisy, noiseless, echoic and anechoic, and in different testing contexts.

The paper is organized as follows. Section 2 gives an overview of previous studies on identification and localization and related tasks. Section 3 details the proposed system. Section 4 presents the evaluation data, tests, results and discussions. Finally, Section 5 concludes the paper and introduces future directions.

2. Related Work

Audition is used in various applications of human-robot interaction for performing tasks including scene analysis, speaker identification, speech recognition, speaker localization and tracking [7–10]. For example, a robot was able to manage a quiz game with multiple players playing at the same time based on their speech utterances in [6]. Further, relying on sound, in [7] an interaction partner was selected among many, and the system adapted its behavior accordingly. In [11], audition was integrated alongside vision and motion to track multiple speakers. In [12], a robot performed sound identification, relying on microphone-array-based sound acquisition. Paradigms of sound acquisition include microphone arrays [6, 12], and binaurality with the humanoid SIG2 [8], for example. Whatever the context and paradigm, speaker identification and speaker localization – or localization of sound sources in general – have seldom been addressed simultaneously. Regarding these two tasks, our review of previous work is divided into two separate parts. Techniques used to counter the negative effects of sound reflections and noises are then reviewed.

2.1. Speaker Identification

Sound-based speaker identification requires discriminative acoustic feature extraction and exploitation. Among acoustic features, MFCCs and linear predictive coding (LPC) coefficients are widely used [13–15]. More recently, other features, such as i-vectors have been adopted [16–18]. Exploitation has been addressed with several techniques, like GMMs [15, 19] and support vector machines (SVMs) [13, 20]. Other approaches rely on universal background modeling (UBM) [21], usually used with GMMs, probabilistic linear discriminant analysis (PLDA) [22, 23] and deep neural networks (DNN) [24, 25].

Most of the previous speaker identification systems operated in a single-microphone context. Nevertheless, multi-microphone systems have also been adopted when addressing this task [26–29]. In this situation, one possibility is to use the present signals for separate speaker identifications and combine them later into a final decision. Another possibility is to process the present signals to produce a signal that is better suited for single-signal feature extraction and exploitation. Aside from the multiple and single-microphone contexts, binaural audition has emerged but very few studies on it have been done thus far. In [15], binaural speaker identification relied on MFCCs and GMMs, exploiting both ear signals at the same time. It was proved to perform better than single-microphone-based systems with similar computational steps.

2.2. Speaker Localization

Sound-based localization has been addressed in different paradigms, like microphone arrays [30, 31]. Unlike speaker identification however, numerous previous studies have approached this task in the binaural context. Binaural localization studies have largely focused on estimation of the azimuth angle and relied on interaural differences used in human audition. Indeed, when a speaker, or a more general sound source, is located at a given non-zero azimuth according to the receiver (see **Fig. 1**), the corresponding received sound is louder in one of the ears, which also senses the sound before the other ear. This creates interaural time, level and phase differences (ILD, ITD and IPD), which are widely computed and exploited. Several different techniques have been adopted for computing these features. In some cases, monodimensional features have been computed [2, 32, 33], using only the genuine signals. Alternatively, multidimensional features can be extracted as sets of frequency-dependent components, which require a signal frequency-dependent decomposition. This can be done through FFTs [2, 34–36] or filterbanks, notably of gammatone filters [4, 5, 37, 38].

Different techniques have been used for the exploitation of the extracted features. One option is to link the azimuth angle to the difference of distances between the source and the ears and thus to a difference of sound arrival times at the ears. This explicitly links the ITD to the azimuth [32, 35, 39]. Nevertheless, noises and sound reflections corrupt the signals and thus the acoustic features, limiting the applicability of these straightforward approaches. Alternatives rely on machine learning methods [2, 3, 5, 34]. It is also worth noting that different models have been proposed to reproduce the neural transduction stages of the human auditory system [3, 40–42].

Among studies that address identification and localization at the same time, a binaural scene analyzer was pro-

posed in [43]. The system consecutively performed localization, speech detection and speaker identification. Identification was monaural as only the cues extracted from the ear on the side of the source were used. In related work, speech segregation and localization were tackled in [44, 45]. In these studies, speaker identification was not directly addressed, but the outputs can be further exploited for identification based on missing data classification, which was used in [43].

2.3. Processing of Additive Sound Effects

Sound reflections and noises alter the acquired sounds of interest and may reduce the intelligibility of the speech or position. The environmental acoustics are usually characterized by the reverberation time RT60. This is the frequency-dependent time for the sound level in the environment to decrease by 60 dB after the present sound sources stop emitting. The impact of noise in the environment is characterized by the signal-to-noise ratio (SNR). These problems have been tackled in previous studies with strategies like spectral subtraction and dereverberation [46, 47].

The precedence effect allows humans to overcome sound reflections up to a certain extent [48–50]. In some studies that aim to reproduce the precedence effect, signal onsets are expected to be the least affected by reverberations. A selection of only highly energetic portions permits exploitation of less altered information [37]. Other approaches measure the coherence between signals acquired in different channels, as more "coherent" portions are considered to be more useful. This was applied in the binaural context for a localization purpose in [32, 40].

This concludes the overview of previous work addressing speaker identification, localization and performance enhancement under acoustically constraining conditions. In the next section, a binaural method for simultaneous speaker localization and identification is proposed.

3. Proposed Method

The proposed method is designed in a binaural context. For both localization and identification, the system uses the same front end and requires acoustic feature extraction and exploitation for each. Signals from both the left and right channels are submitted to a feature extraction stage that delivers identification features for each channel and localization features relying on both channels. This stage processes sound waves by filterbanks imitating the cochlear filtering, with a specific number of filters and frequency band extension. Extracted features are submitted to ANNs that exploit information in the features to reveal the speaker identity and speaker-robot azimuth angle. Each ANN has a hidden layer with a specific number of hidden cells. The number of input cells depends on the dimension of the feature vector used. Two identification ANNs are used, one for each channel, and a single ANN is used for localization. Finally, a combination of identity and azimuth outputs is proposed as a means to reduce the errors. The consecutive blocks composing the system are summarized in **Fig. 2**, and their components and settings criteria are detailed in the following.

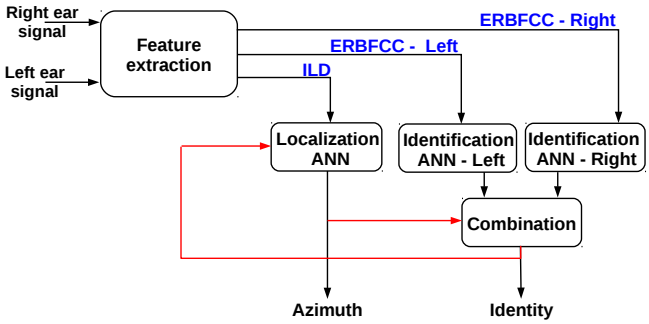

Fig. 2. Overall system architecture.

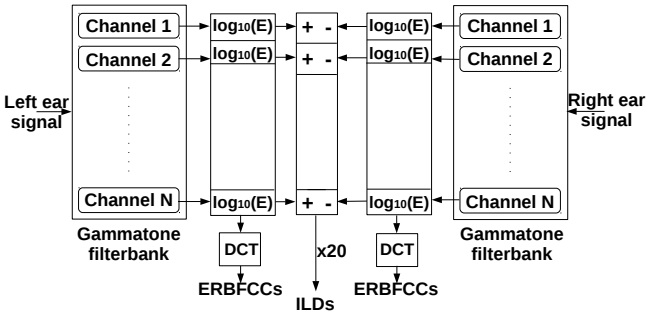

Fig. 3. Architecture of the feature extraction stage.

3.1. Acoustic Feature Extraction

The feature extraction stage is shown in **Fig. 3**. A common binaural front end reduces the system's complexity and computational loads. Voice activity detection (VAD) is performed by thresholding the energies of time frames and discarding silent portions with low energy. A gammatone filterbank of N_f filters regularly spaced on the ERB scale is employed in each ear. Indeed, gammatone filterbanks [51, 52] are widely used for their efficiency in mimicking the functioning of the cochleae [3, 4, 53]. The outputs of these filters are then employed to compute the features. Speaker identification uses ERBFCC features and speaker localization relies on ILDs.

3.1.1. ERBFCC

As mentioned in Section 2.1, MFCCs, requiring mel-scale triangular filterbanks, are widely used for speaker identification. In order to have a common binaural front end for localization and identification and to benefit from using ERB-scale gammatone filters, an acoustic feature is extracted, that relies on them. This feature is named ERBFCC, as classical mel-scale triangular filterbanks are replaced by gammatone filterbanks. Let $E^l_{t,f}$ be the output energy of the left ear's t-th frame and f-th gammatone filter. The k-th ERBFCC coefficient at the left ear, $\gamma^l_{t,k}$, is

computed through the discrete cosine transform:

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

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

ERBFCCs are computed at the right ear in a similar way. Note that gammatone filterbanks have also been used in [54, 55] for speaker identification through the GFCC feature.

3.1.2. ILD

Widely used for azimuth estimation, alongside the IPD or ITD [35, 36, 45], the ILD is adopted in the presented system. The difference in the levels of the signals reaching the two ears is caused mainly by the head shadow effect [39]. This effect depends on the source's azimuth and the head shape and is negligible at low frequencies. Contrary to level differences, the Duplex theory [56] adds that time differences are exploited at lower frequencies in humans. ITDs are not extracted in this work since here ILDs are considered to be reliable enough and in order to not increase the computational loads. ILDs are advantageously extracted through the same front end used for identification cue extraction. For the left and right f-th gammatone filters output energies $E^l_{t,f}$ and $E^r_{t,f}$, in this work the ILD $\delta_{t,f}$ is computed as:

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

As depicted in **Fig. 3**, the binaural front end computes log output energies at the left and right ear gammatone filters. They are then submitted to subtraction and the discrete cosine transform to compute the ILDs and ERBFCCs respectively. For each time frame t, an ILD vector Δ_t and two left and right ERBFCC vectors Γ^l_t and Γ^r_t are available, as

$$\Delta_t = [\delta_{t,f_i}, \delta_{t,f_{i+1}}, \ldots, \delta_{t,f_l}]^T, \quad (3)$$

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

and

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

where f_i and f_l are respectively the first and last gammatone filter indices between which ILDs are taken into account. K is the index of the last cepstral coefficient taken into account at both ears, and the vector v^T is the transpose of a vector v.

3.2. Acoustic Feature Exploitation

Feature vectors are available for each time frame. The system adopts supervised learning, which allows it to perform classification for speaker identification, and prediction for speaker azimuth estimation. ANNs are used for both tasks for their practicality as compared to generative methods, as one ANN can be used for each task. Their architectures and usage are depicted in the following. Before their exploitation by the ANNs, feature vectors are submitted to a smoothing operation as shown next.

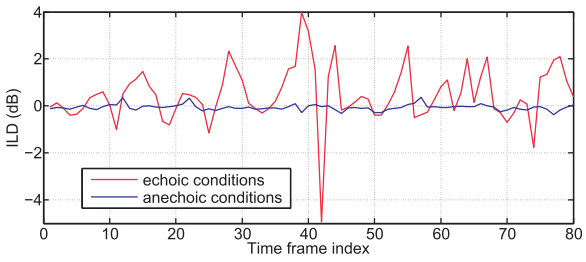

Fig. 4. Consecutive frames ILDs computed at the 20th gammatone filter of a 40-filters filterbank covering up to 22050 Hz. The source is at an azimuth of 0°. The plots correspond to anechoic conditions, and echoic conditions with RT60 = 200 ms at 1 kHz.

3.2.1. Feature Smoothing

Feature smoothing is employed in the present system to reduce the negative effects of sound reflections and noises on the extracted acoustic features. In the presence of reflections, features are distorted when computed using the original signals [57]. **Fig. 4** contains two plots of the same feature computed as a function of time in anechoic and echoic conditions. It illustrates feature fluctuations in the presence of reflections. In this system, low-coherence or onset-following signal portions are not discarded as was proposed in previous work. The system keeps all the frame-wise computed features and reduces their fluctuations through a smoothing process. A similar smoothing operation in the training and testing phases is expected to improve the performance of the system. This operation is not done to provide features that are too stable across time, since training features should keep a certain dynamic aspect allowing the system to have generalization capabilities. Given a series of consecutive vectors, for the vector v_t of the t-th frame, a new vector v^S_t is obtained through triangular weighting:

$$v^S_t = \frac{1}{(n_v+1)^2} \sum_{l=t-n_v}^{l=t+n_v} (-|l-t|+n_v+1)v_l,$$

where n_v is the number of vectors taken into consideration before and after the time index t.

3.2.2. Artificial Neural Networks

Feed-forward multi-layer perceptrons exploit the acoustic features. They are trained with the full gradient back-propagation algorithm with regular cross-validation.

a. Localization ANN. As shown in **Fig. 5**, this ANN exploits the vectors Δ and is trained to output the corresponding azimuth angles. Thus, it has only one output, which is the value of the azimuth angle.

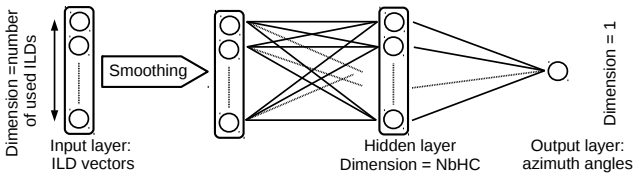

Fig. 5. ANN-based localization feature exploitation.

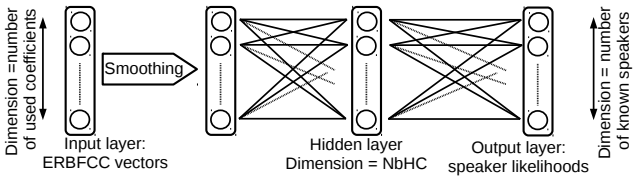

Fig. 6. ANN-based identification feature exploitation.

b. Identification ANNs and Left-Right Output Combination. One ANN is associated with each ear. There are as many output cells as speakers that may be in interaction with the machine (see **Fig. 6**). During training, for an input vector extracted from speaker s, the output vector takes a value α, $0 < \alpha < 1$ at cell s and all the other output components take the value $-\alpha$. During testing, the speaker corresponding to each frame input vector is identified through the most active output cell. For each time frame, each ear ANN provides its independent identity output. The monaural results are combined in a way to keep only the frames having the same output at both ears as valid and to discard the rest. This strategy aims to benefit from the binaural context and to remove a part of the erroneous monaural results, improving the performance.

3.3. Localization and Identification Combination

By operating with different features, the identification and localization modules may not have the same weaknesses in different acoustic conditions. Indeed, depending on the system's training, current conditions and the data, one of the two modules may operate better than the other. Nevertheless, this system presents the advantage of performing both tasks simultaneously, collecting frames and extracting features and outputs. It then combines the outputs of both modules, and can thus take advantage of the stability of one to enhance the other. Such a combination is proposed through the feeds as presented by the arrows linking the combination and localization blocks in **Fig. 2**. The proposed algorithm is based on the idea that when either module provides stable outputs for a certain time window, a single speaker/position is active. In this case, both modules should provide coherent and stable outputs for the window in consideration. The operation is shown in **Algorithm 1**.

As other reliable algorithms can be envisaged, this algorithm is a step towards more beneficial mutual exploitation of identity and azimuth information. It refines the outputs of the system under the assumption of the presence of a single speaker. This speaker is either in a fixed position or moving to a certain azimuth at a changing pace.

Algorithm 1: Identification and localization outputs combination algorithm.

Let n_{obs} denote the number of frames to observe, $\theta_{\max}$ an admissible azimuth interval, and τ a constant with $0 < \tau < 1$;
Consider only the frames kept by the left and right-ear identification outputs combination;
for *each frame,* **do**
  consider a window containing the n_{val} valid frames that were not discarded from the original n_{obs} frames preceding (or following) it. In this window;
  if *the interval between the highest and lowest estimated azimuths is smaller than* $\theta_{\max}$ ***or*** *the number of occurrences of the estimated speaker identities mode is higher than* $\tau \times n_{val}$, **then**
    output the window's average azimuth estimate and speaker identity estimates mode;
  else
    either more than one speaker/position are concurrently active or the features are highly corrupted; the window outputs are discarded;
  end
end

3.4. Overall System Architecture and Parameters

The system architecture has now been detailed. In brief, sound waves are received at the left and right-ear microphones, and signals are processed for cochlear filtering-based feature extraction as detailed in Section 3.1. The resulting acoustic features are then exploited through a smoothing process followed by separate modules based on ANNs, providing the identity and azimuth outputs. The outputs of the system are in the form [identity, azimuth], in which the confidence/stability of one can help to improve the stability of the other. A combined output can be used to improve performance as described in Section 3.3.

For the rest of the paper, and unless specified otherwise, the system is parameterized as follows. Features are extracted in 23-ms frames. Consecutive frames have a 50% overlap. The ANNs have 30 hidden cells each, and the training is done for 5000 iterations. The total number of training utterances is (50 to 100 frames)×(number of directions)×(number of speakers). For the smoothing operation, n_v is between 10 and 20. This setting guarantees a fast and light training of the system with small amounts of data. It is practical in contexts in which collecting training data is costly in terms of equipments and setup.

4. Evaluation

In this section, the proposed system's overall performance is evaluated. A study is also conducted to assess the ability of the extracted feature components to reflect the identity or position information. Evaluations are per-

formed in simulated and real environments that present challenging conditions with sound reflections and noises.

4.1. System Operation and Evaluation

The system operation takes place as follows. One speaker from a predefined set of speakers utters speech, and the machine estimates his/her azimuth and identity. Both the speaker and the machine are in an environment where sound reflections and noises may exist, and the speaker is either in a static position or moving. Speaker identification is evaluated with the true identification rate I_R:

$$I_R = \frac{\text{number of tests correctly identified}}{\text{total number of tests}} \times 100. \quad (6)$$

Note that a test can be based on the feature vectors of a single time frame, or on longer durations as in the proposed output combination algorithm in Section 3.3. Regarding the uttered speech, speaker identification can occur in two ways:

- Text-dependent: identification is based on specific words pronounced by the speaker.
- Text-independent: identification is performed regardless of what the speaker says.

In the presented evaluations, some of the tests are performed with frames not presented to the ANNs during training, but extracted randomly from the same speech utterances used for training. In the sense that these frames belong to utterances partially known by the system, the corresponding tests are considered text-dependent here although this testing context does not coincide with conventional text-dependent contexts. The remaining tests are performed with frames from speech segments not exploited for training and although this does not comply with the term's conventional use, they are regarded as text-independent. These tests consider all the available data not exploited for training, and their numbers change for different speakers and directions. Regarding the speakers to be identified by the system, speaker identification can be made in two contexts:

- On a closed set of speakers: only a known identity can be the output.
- On an open set of speakers: either a specific identity or "stranger" or "unknown" can be the output.

The proposed system is evaluated on a closed set of speakers as the foreseen applications, in assistance contexts for example, operate with limited numbers of speakers.

Tests consider all the speakers and directions available in the evaluation datasets. The localization performance over a given test set is reflected by the average frame-wise difference between the real azimuth values and corresponding estimations. This measure is called the location estimation error (L_E).

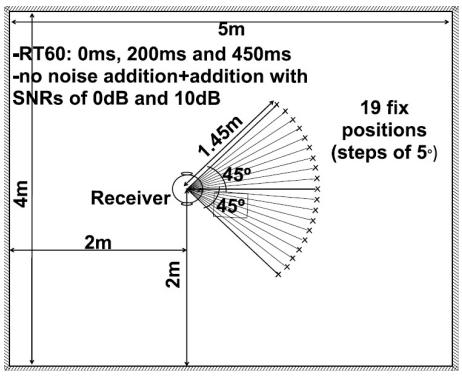

Fig. 7. Simulated room: binaural receiver position, all source positions, and acoustic conditions.

Before presenting the performance evaluation results, the established evaluation datasets are presented. A method to analyze the effectiveness of extracted acoustic features is also proposed. This can help to identify and discard the least useful features, reducing the system's complexity.

4.2. Established Datasets

Two datasets of binaural sound acquisition corresponding to several speakers uttering from different positions were established, in simulation and in real environments. These datasets cover simple as well as tough acoustic conditions judged annoying in ordinary listening tests, which include noise and reverberations at different levels.

4.2.1. Simulation Database

A 4 m $\times$ 5 m $\times$ 2.75 m shoebox room environment and binaural sound acquisition were simulated using Roomsim [58]. The binaural room impulse responses of the room, receiver and source configuration were generated with the image method [59]. They were based on head-related transfer functions (HRTFs) provided in the CIPIC HRTF database [60]. The original speech segments were taken from good quality radiophonic recordings. The speech was generated as if uttered by ten male speakers, each placed at the 19 azimuth angles ranging between $-45°$ and $45°$ with a step size of $5°$. The speaker-machine distance was 1.45 m (see **Fig. 7**). The database covered anechoic and echoic environments at 200 ms and 450 ms RT60s at 1 kHz. A white noise signal was added to the two ears signals at the same level, which was set according to the desired SNR in the left ear. As a result, the two ears signals might not have been at the same SNR as they might not have had the same original signal level before noise addition. To be more precise, the SNRs in the following results were measured at the left ear. This was done for SNRs of 0 dB and 10 dB, in addition to signals with no noise added.

4.2.2. Recorded Database

Recordings were made in an echoic environment with walls and a ceiling of brick and concrete, curtains on one

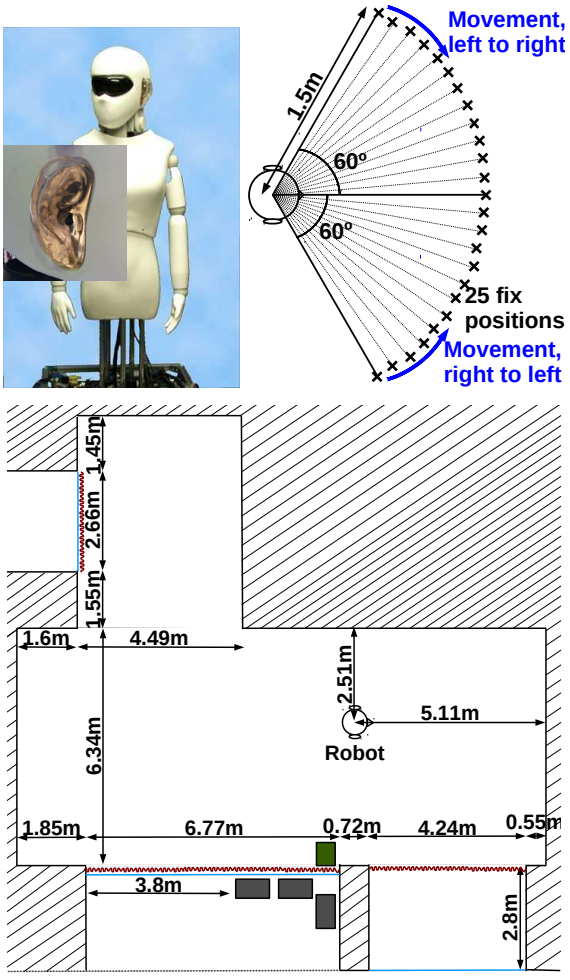

Fig. 8. Upper left: SIG2 humanoid robot and its right ear. Upper right: speaker-robot fixed positions and movement directions. Lower: robot head position and orientation in the real room.

of the walls and carpets on the floor. The reverberation time was estimated to be near 800 ms, and the ambient noise was mainly generated by a refrigerator and was clearly audible to a human listener in the environment. The SIG2 robot binaural head was placed and oriented as shown in **Fig. 8**. Speech utterances were emitted through a loudspeaker placed at the same height as the SIG2 ears, and oriented towards it. The utterances were extracted from the TSP Speech Database [61], which provides 60 English sentences per speaker with an average duration of 2.3 s per sentence over all the speakers. In brief, the recordings can be characterized as follows:

- 10 speakers.
- The distance was constant at approximately 1.5 m.
- Fixed positions: 25 azimuth angles, between $-60°$ and $60°$ with a step of $5°$; specific speech (two sentences) for each speaker and position.
- Movements: two segments per speaker, from left to right and from right to left from the perspective of the robot; specific speech (ten concatenated sentences) for each speaker but identical for both segments.

The sound signals were sampled at 48 kHz. The source-receiver (speaker-robot) position was set and measured using a NaturalPoint OptiTrack Motion Capture system.[1] This allowed us to determine ground truth information about the source-receiver positions. For the fixed positions, the manual actual positioning of the source and receiver included some error when compared to the desired theoretical positioning. However, the actual positions were adjusted so as to maintain small errors. The average errors were approximately $0.11°$ in azimuth and 0.5 cm in distance. For movement recordings, the speed was approximately 3 azimuth degrees per second.

Such a database is needed for computer audition contexts. It can be used for binaural and monaural applications like speaker identification, speech recognition, and speaker localization, and is intended to be made publicly available.

4.3. Acoustic Feature Analysis and Selection

The computed feature vector components do not have the same relevance regarding the required information. For example, the Duplex theory [56] states that ILDs give better information about the azimuth at higher frequencies than at lower frequencies. For a still speaker, features from different frames do not maintain the same value, although reflecting the same position, for several reasons (computation, frame duration, spectral content, noises and reflections, etc.). Feature components whose values vary the least when the speaker is in a fixed position but do vary for different positions are well suited for the system operation. Discarding the least suited components reduces the complexity and computational loads of the system. To this end, the monodimensional Wilks' Lambda [62, 63] is used. This tool was used in a study comparing different methods of extracting binaural robot audition cues in [64]. It compares the intra-group dispersion to the overall dispersion for each component. In the given example, the ILD values computed at each azimuth angle and gammatone channel are assembled in a corresponding group. This measure can be calculated for the f-th gammatone channel's groups of ILDs as:

$$\lambda_f = \frac{\frac{1}{N}\sum_{g=1}^{G} n_g v_{g,f}}{V_f}, \quad \ldots\ldots\ldots\ldots \quad (7)$$

where N is the total number of examples across groups. n_g and $v_{g,f}$ are respectively the number of examples and their variance in the group of azimuth g at channel f. V_f is the variance of the whole set of values computed over all the f-th channel's 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 discrimination ability is. Feature group clusters become tighter and better separated from each other when it decreases.

1. http://www.naturalpoint.com/optitrack [Accessed February 5, 2017]

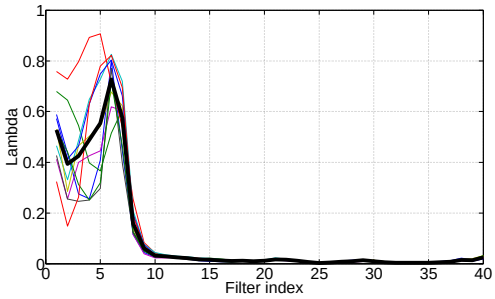

Fig. 9. Simulations: λ values for ILDs as a function of the gammatone filter index. Each colored plot corresponds to an individual speaker, while the bold black curve plots the average measures across speakers; anechoic environment, no noises added to the signals.

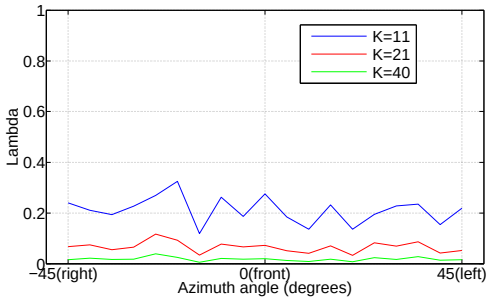

Fig. 10. Simulations: λ values for ERBFCC vectors as a function of the azimuth angle and number of considered coefficients; anechoic environment, no noise added to the signals.

Similarly, in the case of multidimensional features, λ can be calculated through the intra-group covariance matrices m_g and the overall covariance matrix M for all data from all the groups:

$$\lambda = \frac{\det\left(\frac{1}{N}\sum_{g=1}^{G} n_g m_g\right)}{\det(M)}. \quad \ldots\ldots\ldots \quad (8)$$

4.3.1. ILD Component Analysis

Using simulation data and Eq. (7), λ values were computed for ILDs corresponding to 40 gammatone filters composing the left and right-ear filterbanks. **Fig. 9** plots the measures for all the speakers as functions of the filter index in colored curves. The average value over all speakers is plotted in the bold black curve. The figure reveals two main aspects:

- λ values were relatively high at low frequencies. ILDs are thus less reliable at lower frequencies than at higher frequencies, which is in accordance with the Duplex theory.
- The values varied slightly across speakers. This indirectly shows that ILDs do not only depend on the speaker position but also on the speaker and/or speech. Certain speakers can thus be more accurately localized than others. Moreover, the λ differences across speakers were more obvious at the lower frequencies. This also indicates the non-reliability of their corresponding ILDs.

For the system operation, the ILDs corresponding to the 20 gammatone filters indexed from 16 to 35 are used.

4.3.2. Cepstral Feature Analysis

Cepstral features, such as MFCCs, are usually considered in groups, i.e, in sets of coefficients composing the cepstral feature vector. The current ERBFCC analysis aimed to find the most adequate K. K is the index of the last cepstral coefficient taken into account at both ears. **Fig. 10** plots the λ values computed using the simulation data and Eq. (8) for $K = \{11, 21, 40\}$ at different azimuth angles. The cepstral coefficients that were used were extracted from the left ear signals. The figure shows a decrease of λ and thus an improvement in speaker discrimination ability as the number of coefficients increases. For the system operation, K is thus set to 40.

4.4. Performance Evaluations

The evaluation results described in Section 4.1 are now presented. The simulation data made it possible to study the system performance for still speakers as a function of the SNR and RT60, while the recorded data were acquired in ambient acoustic conditions with no intentional intervention in the SNR or the RT60. This data allowed for the study of the performance with still and mobile speakers.

4.4.1. Performance with Simulation Data

Tests with data extracted from the simulation database address identification and localization separately without the combined output for the moment. The object here is instead to study their accuracy as a function of the text-dependence, acoustic conditions, and context of channel acquisition. Note that these tests each last for the duration of only one frame. This duration forms the basis for longer duration tests and allows for an efficient study of the localization and identification accuracies before exploiting longer durations.

a. Speaker Identification. Identification rates are reported in **Fig. 11** for single frame tests. The tests were performed in the proposed binaural context and in a monaural context based on the left ear signals. They addressed text-dependent and text-independent scenarios as explained in Section 4.1. The I_R increased for increasing SNR, it reached values near 95% and 85% in the binaural text-dependent and text-independent contexts, respectively.

The binaural context outperformed the monaural context in all acoustic conditions, showing the benefit of the proposed binaural exploitation. Moreover, the results were better in the text-dependent context than the

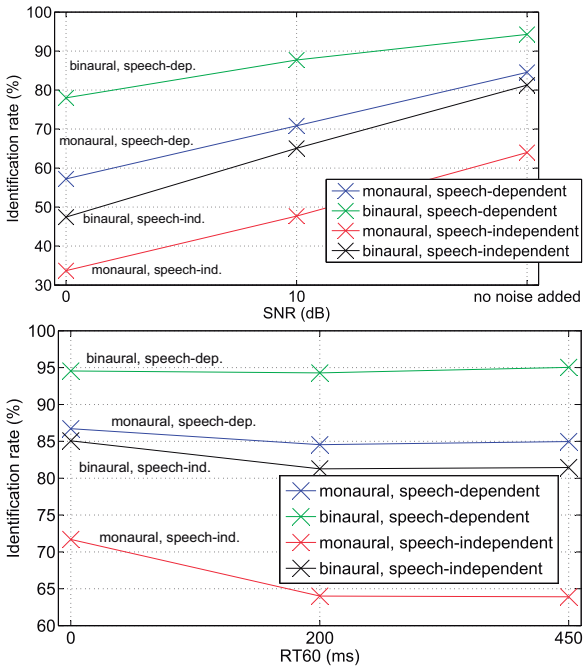

Fig. 11. Simulations: frame-based speaker identification rates. Upper: RT60 = 200 ms, different SNRs. Lower: no noises added, different RT60s.

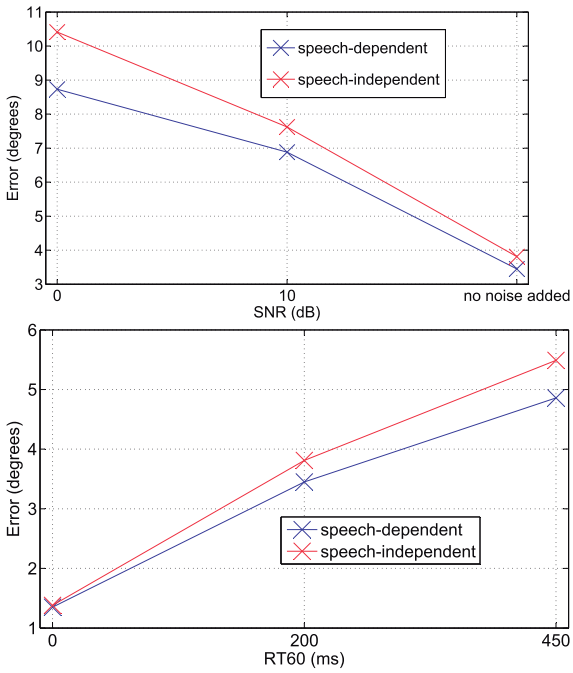

Fig. 12. Simulations: frame-based speaker localization mean errors. Upper: RT60 = 200 ms, different SNRs. Lower: no noises added, different RT60s.

text-independent context. This was expected since text-dependent features have more similarities with the training features than text-independent ones do. It can also be stated that the system was much more sensitive to noises than it was to reflections. Noises originated from a different source than any of the speaker and altered the speaker-specific information contained in the signals, whereas, despite their negative effects on the speech and position intelligibility, reflections corresponded to the given speaker and no to any other. The system thus learned reflections as corresponding to speakers, and they did not significantly increase the possibility of confusion between speakers.

b. Speaker Localization. L_E values are reported in **Fig. 12**. Contrary to speaker identification, for which the performance was similar for different RT60s, localization was sensitive to both noises and reflections. The L_E increased with the decrease of the SNR and the increase of RT60. Nevertheless, in moderate acoustic conditions, it stayed low, even reaching 1.3° for anechoic speech with no noise added. As previously specified, the L_E is the average frame-wise azimuth error. The azimuth error standard deviation was lower than the L_E in all tests, but not more than 3° lower. For example, its value was near 2.5° at a 200 ms RT60 and with no noise added. The error standard deviation also increased with decreasing SNR and increasing RT60, reaching 6° with a 200 ms RT60 and a 0 dB SNR, for example. Localization was also less sensitive to text-dependence or text-independence than identification was. Errors in the speech-dependent context were still slightly lower. Indeed, ILDs are theoretically only position-dependent and have only a slight dependence on speaker and/or speech characteristics as previously seen.

Table 1. Real data: performance measures with independent and combined outputs contexts.

Operation	Regular	Combination
I_R [%]	79.4	85.6
L_E [deg.]	5.3	3.9

4.4.2. Performance with Real Data

The recorded database was used to evaluate the proposed system, not only with a machine and speakers at a fixed position, but also with movements. Independent and combined localization and identification outputs were considered. As for the simulation data, VAD was applied prior to feature extraction as shown in Subsection 3.1. In the following, the test results are presented.

a. Still Speakers. Tests in scenarios without movements were made in a text-independent context. The tests considered both independent and combined identification and localization outputs in order to study the effectiveness of the output combination. For the combination-based functioning, the parameters n_{obs}, τ and $\theta_{\max}$ were respectively set to 50, 0.51 and 20. The results are reported in **Table 1**. They show the improvement to the system performance as a result of the proposed output combination. An I_R increase of more than 6% and an L_E reduction of nearly of nearly 1.4° were observed. The parameter setup of the system is in relation to the assumptions about the scenario. For the assumption of a single speaker in a fixed position, larger values of n_{obs} can be tolerated. In the current test, with a single unmoving speaker, an evaluation was made with n_{obs} extended to cover the entire utterance duration. A single azimuth output and a single identity output are thus provided for each utterance. This

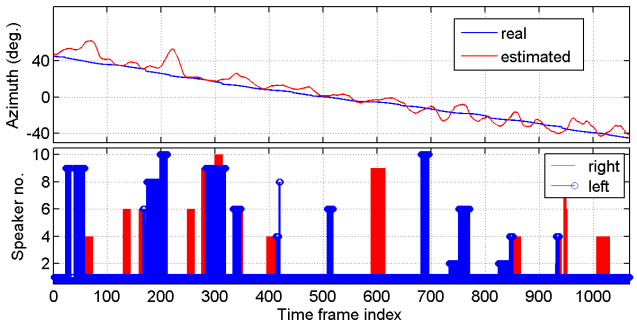

Fig. 13. Real data: moving speaker localization and identification as functions of time. Upper: localization. Lower: monaural identifications; speaker 1 is the real speaker.

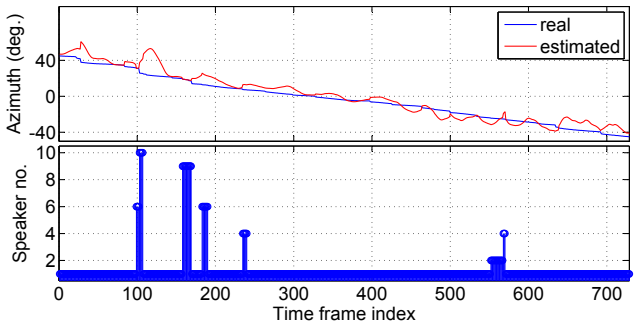

Fig. 14. Real data: moving speaker localization and identification as functions of time; frame selection based on binaural speaker identification. Upper: localization. Lower: binaural identification; speaker 1 is the real speaker.

increased the I_R to 99.6% and decreased the L_E to 3.1°. For a moving speaker, the average azimuth estimate over a large number of frames cannot be used to efficiently track the movement. Similarly, the mode of speaker identity outputs disregards the possibility of another speaker making an utterance during the test, even if both utterances do not overlap. An assumption of a speaker moving fast or of a set of alternating utterances for different speakers leads to a decrease in n_{obs}.

b. Moving Speakers. The system trained with data collected from a still machine and still speakers was tested with a moving speaker and still machine. This text-independent test addressed the ability of the system to both track a moving speaker and to generalize to positions not included in the training set. The results presented here considered speaker 1 moving between the azimuth angles 45° and −45°. Independent localization and monaural identification results are reported in **Fig. 13**. **Fig. 14** plots the results taking into consideration only the frames kept by the monaural identification combination module. The figures show the results with independent outputs for azimuth and identity. The frame-based azimuth estimation error had a value of nearly 5.8° in both cases. The monaural I_R took values of 74% (left ear) and 77.3% (right ear), and the binaural I_R was 92.6%. As expected, binaural exploitation of the monaural results significantly improves the identification rate. Applying the algorithm proposed in Section 3.3 to the same test segment, and using the

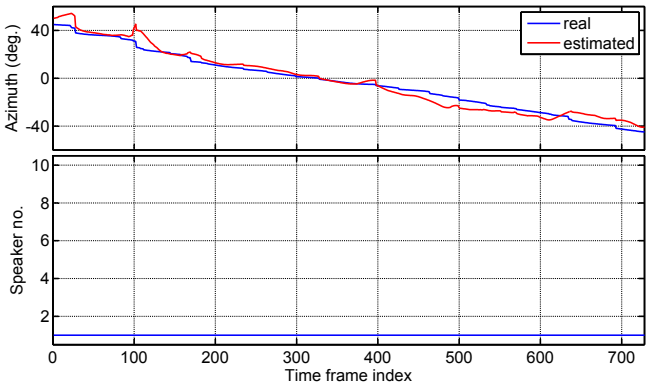

Fig. 15. Real data: moving speaker localization and identification as functions of time, after output combination. Upper: localization. Lower: binaural identification; speaker 1 is the real speaker.

same trained ANN, led to the results shown in **Fig. 15**. Here, the L_E was reduced from 5.8° to 4° and the IR increased to 100%. This shows that the proposed system is also advantageous in a moving speaker scenario. The proposed combination reduced the fluctuations of one output based on the stability of the other and improved the coherence of the outputs. This is done through exploitation of the results over durations longer than a single time frame, which also improves the output stability. This algorithm was beneficially applied in the situation presented here and can be beneficial in windows with more output fluctuations.

4.5. Discussions

The proposed system's architecture, modules, evaluation strategies and results have been presented. The system aims to achieve the best possible performance with limited complexity, training durations and dataset sizes. Evaluations started with the acoustic features themselves before exploitation by ANNs. A study of the feature statistics across speaker identities and positions was made for selection of the most reliable features. Identification and localization modules were then exploited and evaluated, in separate and combined output modes. The results demonstrated the accuracy of the system as well as the following:

- Evaluations in simulated and real environments lead to the same conclusions, but differences in the I_R and L_E can be observed. These differences are not linked to whether the data is simulated or real, but to the environments themselves. The currently established simulated environment creates symmetry around the machine's ears placement. This symmetry is geometric since the distances of the machine's ears to the two closest walls are the same, and the walls have the same acoustic nature. The real data do not present these symmetries, as the machine is not positioned in the middle of the room, and there are curtains to the side of one ear and a wall of concrete to the side of the other. This dissymmetry might affect the performance, but the system was able to efficiently learn in any position, and

produce reliable performance.

- Feature fluctuations are accentuated by the presence of reflections and noises and cause the system outputs to fluctuate. This pattern of outputs brought feature smoothing to the fore as an option to stabilize, and thus improve, the outputs. With moving speakers, moreover, the extracted acoustic features have a dynamic aspect in relation to the movement. In this context, the smoothing window size should be adjusted in accordance with the hypothetical maximal movement speed. Feature velocities and accelerations can also be exploited to further improve performance.

- An alternative to the proposed architecture is to use a single ANN taking the entire set of acoustic features as the input and outputting both the speaker identity and position. This approach allows the ANN to make use of possible position information present in the identification cues and *vice versa*. Such an approach has been implemented and evaluated. Localization performance was generally weaker than for the proposed system architecture. This also caused losses in terms of increasing complexity and computational loads. For example, a neural network with 30 hidden cells trained with 5000 iterations provided a 77.9% I_R and an 8.08° L_E in a text-dependent context and using recorded data. For a neural network with 100 hidden cells, the performance measures were respectively at 84.61% and 7.53° with 5000 training iterations and 87.49% and 7.01° with 15000 iterations.

- Another alternative is to assign one ANN to each speaker and speaker position. Identification and localization results can then be integrated over the positions and speakers. Despite the potential improvement in performance, however, this approach drastically increases the complexity and training duration of the system. The number of ANNs to train becomes equal to the number of speakers multiplied by the number of training directions in such a case.

- Previous speaker identification studies largely relied on single-microphone signals. Although cepstral features and ANNs were used, the ERBFCC feature was presented and binaural signals were exploited. As shown in Section 4.4, a performance comparison demonstrated the advantage of the proposed binaural technique. Note that in a different work, the advantage of the ERBFCC over the widely used MFCC was also witnessed.

- As mentioned in Section 2, a method addressing both localization and identification was proposed in [43]. It was designed in such a way that localization was the first block, followed by speech detection and then speaker identification. Evaluations in noisy and reverberant conditions verified its ability to deliver reliable outputs. Here, the problem is addressed with a different strategy, in which both localization and identification modules run in parallel. They are linked at the input and output levels. A common signal exploitation module is used to extract features and to save computation times. A combination of the identity and azimuth outputs is proposed, guaranteeing the coherence of the outputs and improving the performance.

5. Conclusions

A system was designed for simultaneous identification and azimuth estimation of speakers in fixed or moving positions. The system uses binaural sound inputs and can operate in acoustically constraining environments with noises and sound reflections. It requires limited training data and operates in a way that reduces feature extraction computational steps. This aspect is practical in contexts in which training data collection is difficult and costly. The system extracts the ERBFCC and ILD features and exploits them with ANNs.

A study of the statistical characteristics of the features addressed their saliency for the required information. This allowed for feature selection prior to the overall evaluation. Evaluation data was extracted from real recordings and simulation datasets. The data presented different acoustic conditions and speech uttered by several speakers from different positions. The system demonstrated its accuracy on the performed tasks, with high performance in moderate acoustic conditions and even for very short speech segments. The binaural context outperformed the monaural context for speaker identification. The simultaneous output of identity and azimuth information made it possible to combine beneficial information from the identification and localization modules, increasing the performance of both. This system can furthermore be used to efficiently track a moving speaker. It is also applicable in realistic scenarios with more than one speaker making non-overlapping utterances, creating a sequence of single speaker segments. Our future work will concentrate on evaluating the system in mismatched training and testing acoustic conditions. Our current work addresses the modification of the system to cover identification, localization and tracking of more than one speaker with overlapping speech segments. It also extends the multiple-speaker recognition and localization works presented in [65]. In future work, other kinds of tools and features, such as DNNs and filterbank features, are to be investigated. This is to be done in an architecture maintaining acceptable computation and data collection loads, which is the aim of the currently proposed architecture as discussed earlier.

Acknowledgements

At the time this work was done, K. Youssef was an International Research Fellow of the Japan Society for the Promotion of Science. We would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped us to improve the quality and the presentation of this paper.

References:

[1] H. G. Okuno and K. Nakadai, "Computational Auditory Scene Analysis and its Application to Robot Audition," Hands-Free Speech Communication and Microphone Arrays 2008 (HSCMA), 2008.

[2] E. Berglund and J. Sitte, "Sound Source Localization Through Active Audition," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2005.

[3] T. May, S. van de Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End,"

IEEE Trans. on Audio, Speech and Language Processing, Vol.19, No.1, 2011.

[4] J. Woodruff and D. Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments," IEEE Trans. on Audio, Speech and Language Processing, Vol.20, No.5, 2012.

[5] K. Youssef, S. Argentieri, and J.-L. Zarader, "A Learning-Based Approach to Robust Binaural Sound Localization," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2013.

[6] 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 Int. Symposium on System Integration, 2014.

[7] 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.

[8] 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.

[9] 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.

[10] Y. Sasaki, S. Masunaga, S. Thompson, S. Kagami, and H. Mizoguchi, "Sound Localization and Separation for Mobile Robot Tele-Operation by Tri-Concentric Microphone Array," J. of Robotics and Mechatronics, Vol.19, No.3, 2010.

[11] K. Nakadai, K.-i. Hidai, H. G. Okuno, H. Mizoguchi, and H. Kitano, "Real-time Auditory and Visual Multiple-speaker Tracking For Human-robot Interaction," J. of Robotics and Mechatronics, Vol.14, No.5, 2002.

[12] Y. Sasaki, M. Kaneyoshi, S. Kagami, H. Mizoguchi, and T. Enomoto, "Pitch-Cluster-Map Based Daily Sound Recognition for Mobile Robot Audition," J. of Robotics and Mechatronics, Vol.22, No.3, 2010.

[13] S. S. Karajekar, "Four Weightings and a Fusion / a Cepstral-SVM System for Speaker Recognition," IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.

[14] S. Farah and A. Shamim, "Speaker Recognition System using Mel-Frequency Cepstrum Coefficients, Linear Prediction Coding and Vector Quantization," IEEE Int. Conf. on Computer, Control and Communication, 2013.

[15] K. Youssef, S. Argentieri, and J.-L. Zarader, "From Monaural to Binaural Speaker Recognition for Humanoid Robots," IEEE-RAS Int. Conf. on Humanoid Robots, pp. 580-586, Dec. 2010.

[16] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector Based Speaker Recognition on Short Utterances," Interspeech, 2011.

[17] M. McLaren and D. van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources," IEEE Trans. on Audio, Speech and Language Processing, Vol.20, No.3, 2012.

[18] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech," IEEE Odyssey, 2010.

[19] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information," IEEE Trans. on Audio, Speech and Language processing, Vol.20, No.4, 2012.

[20] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," IEEE Trans. on Audio, Speech and Language Processing, Vol.19, No.4, 2011.

[21] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, Vol.10, No.1-3, January 2006.

[22] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2013.

[23] E. Khoury, L. El Shafey, M. Ferras, and S. Marcel, "Hierarchical Speaker Clustering Methods for the NIST i-vector Challenge," Odyssey: The Speaker and Language Recognition Workshop, 2014.

[24] T. Yamada, L. Wang, and A. Kai, "Improvement of Distant-Talking Speaker Identification using Bottleneck Features of DNN," Interspeech, 2012.

[25] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2014.

[26] M. Ji, S. Kim, H. Kim, K. C. Kwak, and Y. J. Cho, "Reliable Speaker Identification Using Multiple Microphones in Ubiquitous Robot Companion Environment," IEEE Int. Conf. on Robot and Human Interactive Communication, 2007.

[27] N. Zulu and D. Mashao, "Evaluating Microphone Arrays for a Speaker Identification Task," Fifteenth Annual Symposium of the Pattern Recognition Association of South Africa, 2004.

[28] S. Squartini, E. Principi, R. Rotili, and F. Piazza, "Environmental Robust Speech and Speaker Recognition through Multi-Channel Histogram Equalization," Neurocomputing, Vol.78, 2012.

[29] Q. Jin, T. Schultz, and A. Waibel, "Far-Field Speaker Recognition," IEEE Trans. on Audio, Speech and Language Processing, Vol.15, 2007.

[30] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. Takano, "Real-Time 2 Dimensional Sound Source Localization by 128-Channel Huge Microphone Array," IEEE Int. Workshop on Robot and Human Interactive Communication, 2004.

[31] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2006.

[32] R. Liu and Y. Wang, "Azimuthal Source Localization Using Interaural Coherence in a Robotic Dog: Modeling and application," Robotica, Cambridge University Press, Vol.28, pp. 1013-1020, 2010.

[33] H. Finger, S.-C. Ruvolo, Paul aznd Liu, and J. R. Movellan, "Approaches and Databases for Online Calibration of Binaural Sound Localization for Robotic Heads," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2010.

[34] P. Smaragdis and P. Boufounos, "Position and Trajectory Learning for Microphone Arrays," IEEE Trans. on Audio, Speech and Language Processing, Vol.15, No.1, 2007.

[35] M. Raspaud, H. Viste, and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD," IEEE Trans. on Audio, Speech and Language Processing, Vol.18, No.1, 2010.

[36] J. Nix and V. Hohmann, "Sound Source Localization in Real Sound Fields based on Empirical Statistics of Interaural Parameters," J. of the Acoustical Society of America, Vol.119, No.1, 2006.

[37] M. Heckmann, T. Rodemann, F. Joublin, C. Goerick, and B. Schölling, "Auditory Inspired Binaural Robust Sound Source Localization in Echoic and Noisy Environments," Int. Conf. on Intelligent Robots and Systems, 2006.

[38] K. Youssef, S. Argentieri, and J.-L. Zarader, "A Binaural Sound Source Localization Method Using Auditive Cues and Vision," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2012.

[39] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner, "A Probabilistic Model for Binaural Sound Localization," IEEE Trans. on Systems, Man and Cybernetics – Part B: Cybernetics, Vol.36, No.5, October 2006.

[40] C. Faller and J. Merimaa, "Source Localization in Complex Listening Situations: Selection of Binaural Cues based on Interaural Coherence," J. of the Acoustical Society of America, Vol.116, No.5, November 2004.

[41] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory Model Based Direction Estimation of Concurrent Speakers from Binaural Signals," Speech Communication, Vol.53, 2011.

[42] L. Bernstein, S. van de Par, and C. Trahiotis, "The Normalized Interaural Correlation: Accounting for NoSPi Thresholds Obtained with Gaussian and "Low-Noise" Masking Noise," J. of the Acoustical Society of America, Vol.106, No.2, August 1999.

[43] 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 Trans. on Audio, Speech and Language Processing, Vol.20, No.7, 2012.

[44] J. Woodruff and D. Wang, "Binaural Detection, Localization, and Segregation in Reverberant Environments Based on Joint Pitch and Azimuth Cues," IEEE Trans. on Audio, Speech and Language Processing, Vol.21, No.4, 2012.

[45] N. Roman, D. Wang, and G. G. Brown, "Speech Segregation based on Sound Localization," J. of the Acoustical Society of America, Vol.114, No.4, October 2003.

[46] Y.-L. Wan, K. T.-Q. Zhang, Z.-C. Wang, and J. Jin, "Robust Speech Recognition based on Multi-Band Spectral Subtraction," IEEE Int. Congress on Image and Signal Processing, 2013.

[47] B. W. Gillespie, H. Malvar, and D. Florencio, "Speech Dereverberation via Maximum-Kurtosis Subband Adaptive Filtering," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2001.

[48] J. Blauert, "Spatial Hearing. The Psychophysics of Human Sound Localization," chapter Progress and Trends since 1982, The MIT Press, 1996.

[49] B. C. J. Moore, "Springer Handbook of Acoustics," chapter Psychoacoustics, Springer, 2007.

[50] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," J. of the Acoustical Society of America, Vol.106, No.4, 1999.

[51] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex Sounds and Auditory Images," Int. Symposium on Hearing, Auditory physiology and perception, pp. 429-446, 1992.

[52] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Technical report, Apple Computer, 1993.
[53] T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, and B. Schölling, "Real-time Sound Localization With a Binaural Head-system Using a Biologically-inspired Cue-triple Mapping," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, October 2006.
[54] X. Zhao, Y. Shao, and D. Wang, "CASA-Based Robust Speaker Identification," IEEE Trans. on Audio, Speech, and Language Processing, Vol.20, No.5, 2012.
[55] X. Zhao, Y. Wang, and D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions," IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol.22, No.4, 2014.
[56] L. Rayleigh, "On our Perception of Sound Direction," Philosophical magazine, Vol.13, No.74, pp. 214-232, 1907.
[57] S. Devore and B. Delgutte, "Effects of Reverberation on the Directional Sensitivity of Auditory Neurons across the Tonotopic Axis: Influences of ITD and ILD," The J. of Neuroscience, Vol.30, No.23, 2010.
[58] D. R. Campbell, K. Palomäki, and G. Brown, "A MATLAB Simulation of "Shoebox" Room Acoustics for use in Research and Teaching," Computer Information Systems, Vol.9, No.3, 2005.
[59] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," J. of the Acoustical Society of America, Vol.65, No.4, 1979.
[60] V. Algazi, R. Duda, R. Morrisson, and D. Thompson, "The CIPIC HRTF Database," Proc. of the 2001 IEEE Workshop on Applications of Signal Processing to audio and Acoustics, pp. 99-102, 2001.
[61] P. Kabal, "TSP Speech Database," Technical report, Department of Electrical & Computer Engineering, McGill University, 2002.
[62] A. El Ouardighi, A. El Akadi, and A. Aboutajdine, "Feature Selection on Supervised Classification Using Wilk's Lambda Statistic," Int. Symposium on Computational Intelligence and Intelligent Informatics, March 2007.
[63] L. Lebart, M. Piron, and A. Morineau, "Statistique exploratoire multidimensionnelle, visualisation et inférence en fouille de données," Dunod, 2008.
[64] K. Youssef, S. Argentieri, and J.-L. Zarader, "Towards a systematic study of binaural cues," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2012.
[65] K. Youssef, K. Itoyama, and K. Yoshii, "Identification and Localization of One or Two Concurrent Speakers in a Binaural Robotic Context," IEEE SMC, 2015.

Name:
Karim Youssef

Affiliation:
Lecturer, University of Balamand

Address:
Balamand, Al Kurah, Lebanon

Brief Biographical History:
2013 Received Ph.D. degree from Pierre and Marie Curie University
2014- JSPS International Research Fellow, Graduate School of Informatics, Kyoto University
2015- Lecturer, University of Balamand

Main Works:
• "A Learning-Based Approach to Robust Binaural Sound Localization," 2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2013.
• "A Binaural Sound Source Localization Method Using Auditive Cues and Vision," 2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2012.

Name:
Katsutoshi Itoyama

Affiliation:
Assistant Professor, Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 417, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:
2011- Received Ph.D. degree from Graduate School of Informatics, Kyoto University
2011- Assistant Professor, Graduate School of Informatics, Kyoto University

Main Works:
• "Query-by-Example Music Information Retrieval by Score-Informed Source Separation and Remixing Technologies," EURASIP J. on Advances in Signal Processing, Vol.2010, No.1 pp. 1-14, January 17, 2011.

Membership in Academic Societies:
• The Institute of Electrical and Electronics Engineers (IEEE)
• The Acoustical Society of Japan (ASJ)
• Information Processing Society of Japan (IPSJ)

Name:
Kazuyoshi Yoshii

Affiliation:
Senior Lecturer, Speech and Audio Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Address:
Room 412, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Brief Biographical History:
2008- Received Ph.D. degree from Graduate School of Informatics, Kyoto University
2008- Research Scientist, Information Technology Research Institute (ITRI), National Institute of Advanced Industrial Science and Technology (AIST)
2013- Senior Researcher, AIST
2014- Senior Lecturer, Graduate School of Informatics, Kyoto University

Main Works:
• "A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation," IEEE Trans. on Audio, Speech, and Language Processing, Vol.20, No.3, pp. 717-730, 2012.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE)
• Information Processing Society of Japan (IPSJ)
• The Institute of Electronics, Information, and Communication Engineers (IEICE)