2012 IEEE RO-MAN: The 21st IEEE International Symposium on
Robot and Human Interactive Communication.
September 9-13, 2012. Paris, France.

An Active Audition Framework for Auditory-driven HRI: Application to Interactive Robot Dancing

João Lobato Oliveira[1,2,4], Gökhan Ince[3], Keisuke Nakamura[3], Kazuhiro Nakadai[3], Hiroshi G. Okuno[4], Luis Paulo Reis[1,5], and Fabien Gouyon[2]

Abstract— **In this paper we propose a general active audition framework for auditory-driven Human-Robot Interaction (HRI). The proposed framework simultaneously processes speech and music on-the-fly, integrates perceptual models for robot audition, and supports verbal and non-verbal interactive communication by means of (pro)active behaviors. To ensure a reliable interaction, on top of the framework a behavior decision mechanism based on active audition policies the robot's actions according to the reliability of the acoustic signals for auditory processing. To validate the framework's application to general auditory-driven HRI, we propose the implementation of an interactive robot dancing system. This system integrates three preprocessing robot audition modules: sound source localization, sound source separation, and ego noise suppression; two modules for auditory perception: live audio beat tracking and automatic speech recognition; and multi-modal behaviors for verbal and non-verbal interaction: music-driven dancing and speech-driven dialoguing. To fully assess the system, we set up experimental and interactive real-world scenarios with highly dynamic acoustic conditions, and defined a set of evaluation criteria. The experimental tests revealed accurate and robust beat tracking and speech recognition, and convincing dance beat-synchrony. The interactive sessions confirmed the fundamental role of the behavior decision mechanism for actively maintaining a robust and natural human-robot interaction.**

I. Introduction

Socially intelligent robots must be able to autonomously interact with humans in natural environments by means of active perception and interactive communication. Active perception depends on a dynamic coupling between perception and behavior, where the robot sensing coordinates its actions while its actions should adapt to improve its perception of the environment [1]. Interactive communication can be verbal, through spoken and natural language, or non-verbal, through embodied expressive behaviors. In combination, both these forms of communication when coordinated by active perception enable robots to exchange information with their human partners while engaging in turn-taking interactions and robustly responding to the shared environment.

This work was partially supported by SFRH/BD/43704/2008 PhD scholarship endorsed by the Portuguese Government through FCT.

[1] Artificial Intelligence and Computer Science Laboratory (LIACC) – FEUP, Porto, Portugal. (joao.lobato.oliveira@fe.up.pt)

[2] Institute for Systems and Computer Engineering of Science and Technology (INESC TEC), Porto, Portugal.

[3] Honda Research Institute Japan Co., Ltd., Saitama, Japan.

[4] Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan.

[5] University of Minho, School of Engineering - DSI, Guimarães, Portugal.

When driven by audition, the creation of such socially intelligent robots capable of interacting with humans in real-world scenarios implies the integration of *i)* Computational Auditory Scene Analysis (CASA) algorithms, able to robustly localize, identify and continuously process live acoustic signals [2]; *ii)* the design of autonomous (pro)active behaviors, able to react or deliberate according to the robot's auditory perception of the environment; and *iii)* the conduction of a robust and natural interaction by policing the robot's behaviors according to the principles of active audition [1]. Hence, in this paper we propose a general active audition framework for auditory-driven HRI, which *i)* simultaneously processes speech and music on-the-fly, *ii)* integrates perceptual models for robot audition, *iii)* supports verbal and non-verbal interactive communication, and *iv)* integrates a behavior decision mechanism based on active audition, which coordinates the robot's actions according to the reliability of the acoustic signals for auditory processing.

To validate the framework's application to general auditory-driven HRI, we propose the implementation of an interactive robot dancing system. The implemented system integrates three preprocessing robot audition modules: Sound Source Localization (SSL), Sound Source Separation (SSS), and Ego Noise Suppression (ENS); with two parallel modules for auditory perception: live Audio Beat Tracking (ABT), and Automatic Speech Recognition (ASR). In addition, the system integrates multi-modal behaviors for verbal and non-verbal interaction, respectively through speech-driven dialoguing and music-driven dancing; and proactive behaviors to ensure a reliable auditory processing. To fully assess the system, we set up experimental and interactive real-world scenarios with highly dynamic acoustic conditions, and defined a set of evaluation criteria. The overall results confirm the application of the proposed framework to general auditory-driven HRI.

II. Related Work

Auditory-driven interactive robotic systems range from musical [3] and dancing robots [4] to conversational agents [5]. These are applied in different social contexts such as entertainment, pedagogical, and therapeutic scenarios. On research with dance-interactive robots, Kozima *et al.* investigated the role of rhythmic engagement and the effects of "interactional synchrony" in human-robot interactions applied to education and child care [4]. By exploring the role of imitation on embodied non-verbal communication, Tanaka *et al.* used QRIO, an infant-size entertainment humanoid

from Sony, for developing a dancing robot that could also interact with children in educational settings [6].

Despite the interactive concepts of these approaches both disregarded the problems inherent to robot audition, such as the effects of noise, the consideration of multiple sound sources, or the simultaneous processing of different acoustic modalities (*e.g.*, music and speech) [2]. By taking motor noise into account for building musically-interactive robots, Oliveira *et al.* [7] utilized a template-based ego noise suppression scheme to suppress motor noise generated from periodic motions of humanoid robots while estimating the beat-times of musical pieces on-the-fly. On an interactive task requiring the simultaneous speech recognition of multiple speakers, Nakadai *et al.* [8] created a robot referee for rock-paper-scissors sound games. Their robot applied sound source localization and separation as a preprocessing to ASR, and made use of a Masking Feature Theory (MFT) algorithm for masking out unreliable features for speech recognition.

Postulating that whenever possible and needed a listening robot should also actively improve its auditory processing, Nakadai *et al.* proposed the concept of active audition, which studies strategies to increase the Signal-to-Noise Ratio (SNR) of the acoustic signals for robot audition [1]. In this pioneering work in the field, a humanoid robot was designed to actively move its head for aligning its microphones orthogonal to the processed sound source. Later, and envisioning HRI applications, Okuno *et al.* combined real-time active audition mechanisms with a visual multiple-tracking system to create a robot receptionist and a party companion robot [9]. Their active audition strategy relied on sensorimotor control for focusing attention on each individual human speaker.

By considering either the principles of active and robot audition, in this paper we propose a general framework for auditory-driven HRI applications. Based on this framework, we implemented an interactive robot dancing system capable of: *i)* dancing to the beat of the music; *ii)* understanding and responding to human speech commands; and *iii)* keeping a natural and robust interaction. The latter involved *iv)* dealing with auditory noise sources of different natures; *v)* handling continuous acoustic stimuli; and *vi)* actively ensuring a reliable auditory processing.

III. Active audition framework

The proposed active audition framework, depicted in Fig. 1, comprises a set of bottom-up layers composed of:

- **Sensing:** corresponds to the physical environment composed of interactive agents and sound sources (*e.g.*, a robot, a human interactor, the spatial environment), live acoustic signals (*e.g.*, musical stimuli, human speech, robot speech, and robot ego noise), and robot sensors (*e.g.*, microphones and motor encoders) to acquire situated information about the agents' body and the acoustic environment.
- **Filtering:** corresponds to low-level processing mechanisms (*e.g.*, SSL, SSS, and ENS) applied to the captured auditory signal for enhancing its quality and reliability as a preprocessing to the *Perception* layer.
- **Perception:** corresponds to high-level perception models applied to enhanced auditory signals towards specific perceptual tasks (*e.g.*, ABT and ASR). This layer outputs high-level features (*e.g.*, musical beats, action commands) that conduct interactive behaviors through the *Behavior* layer.
- **Binding:** this layer is responsible for transforming a set of confidences provided by perceptual models, which give information about the reliability of the acoustic signals, into cost functions; and for combining them into pseudo-continuous fitness functions that determine the policies for the behavior decision.
- **Mediation:** this layer is responsible for assuring a meaningful human-robot interaction while regulating the acoustic signals based on the assigned fitness functions. It acts as an arbiter that coordinates the robot's behaviors (*i.e.*, actions) on-the-fly necessary to maintain interaction and/or request specific actions to improve the acoustic conditions.
- **Behavior:** this layer comprises the repertoire of all the robot behaviors (*e.g.*, dancing, speaking, decreasing the music volume) used for maintaining interaction or improving the acoustic conditions.

The proposed active audition framework represents a general conception that can be extended with additional auditory perceptual modules, new behaviors, and improved mediation of the interaction. This mediation can combine additional cost functions and consider new policies with the objective of enhancing the reliability of auditory processing.

IV. Interactive robot dancing system

Based on the general active audition framework described in Fig. 1, we developed an interactive robot dancing system that integrates a set of functional modules, described below.

A. Preprocessing robot audition modules

1) Sound source localization: The SSL module is responsible for determining the location of each individual sound source based on a multi-channel microphone input. The integrated SSL implementation is based on the MUltiple Signal Classification (MUSIC) algorithm [10].

2) Sound source separation: The SSS module is responsible for splitting the mixed audio signal into the individual sound sources discriminated by the SSL. The integrated SSS implementation is based on Geometric High-order Decorrelation-based Source Separation (GHDSS) [11].

3) Ego noise suppression: Upon each separated audio signal (*i.e.*, music and speech), contaminated with ego noise from the robot's actuators, we apply ego noise suppression to separately enhance each sound source. The integrated implementation is based on the template-based ego noise suppression mechanism described in [12]. Using instantaneous joint status data, of the actuators' velocity and position, this method estimates the ego noise data from a large dataset of audio templates recorded in advance, and

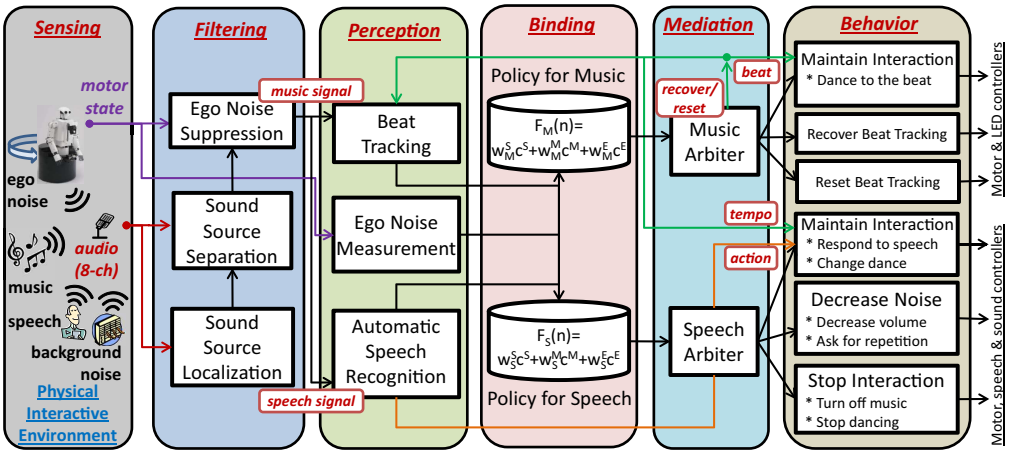


Fig. 1. Implemented interactive robot dancing system based on the proposed active audition framework.

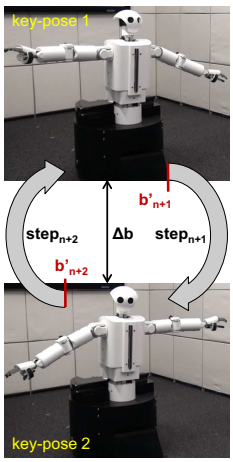


Fig. 2. Beat-synchronous robot dancing generation.

applies spectral subtraction on each separated audio spectrum to refine each independent acoustic signal. Moreover, in order to prevent dominant peaks in the audio signals caused by unpredictable bursting shudder noises generated by loose mechanical parts of the robot (see [7] and [13]), we reinforced the template-based ego noise suppression scheme with a power filter set with a high empirical threshold.

B. Auditory perception modules

1) Automatic speech recognition: The ASR module is based on Julius, an open-source large vocabulary real-time ASR engine [14]. It makes use of feature vectors with 13 static Mel-Scale Log Spectrum (MSLS) features, 13 delta MSLS features, and 1 delta power feature, all calculated on-the-fly. The ASR acoustic model is decoded through a real-time context-dependent Hidden Markov Model (HMM) algorithm. This algorithm calculates a confidence about the correctness of each recognized word given by the scoring function of the recognizer's confusion network. This confidence function, $cf_S(n)$, is used by the system's *behavior decision mechanism* (see Section IV-D) as a measure of reliability of the acoustic signal for speech processing.

2) Ego noise measurement: The level of ego-motion noise, $E(n)$, generated by the robot's actuators is proportional to the velocity of its moving joints. Thus, we compute $E(n)$ as the mean velocity of all the robot's joints, v_j, at frame n:

$$E(n) = \frac{1}{J}\sum_{j=1}^{J} v_j(n), \qquad (1)$$

where J is the total number of joints (*i.e.*, the degrees-of-freedom (DoF)) provided by the robot. This measure represents a confidence function, $cf_E(n)$, that informs the system's *behavior decision mechanism* about the reliability of the acoustic signal for any auditory perceptual task regarding the level of ego noise "contamination".

3) Beat tracking: For performing live ABT over the separated music signal we used IBT, a multi-agent-based online beat tracker first proposed in [15]. This module's architecture is composed of: an audio feature generator that parses the audio signal into a mid-level rhythmic feature; followed by an agents induction function, which (re-)generates new sets of hypotheses regarding possible beats and tempi; and followed by a multi-agent-based tracking algorithm. This algorithm assigns new hypotheses to agents, proceeds to their online ranking and killing, and outputs beats from the current best agent on-the-fly without prior knowledge (*i.e.*, without look-ahead) on the incoming signal. In addition, IBT integrates a confidence mechanism responsible for continuously monitoring the beat tracking analysis of the signal based on abrupt changes in the score evolution of the current best agent [13]. This confidence function, $cf_M(n)$, represents a measure of the reliability of the separated music signal for the task of audio beat tracking, to be used by the system's *behavior decision mechanism*. On request, IBT can either *regenerate* its pool of agents with newly induced hypotheses of beat and tempo, or *reset* itself by killing all existing agents and restarting the system with a new set of induced hypotheses.

C. Interactive robot behaviors

1) Non-verbal communication – music-driven dancing: We designed three distinct robot dance motions to be driven by the musical beat predicted by the ABT. Each is described by cyclic dance step transitions within two manually defined key-poses, which are interpolated in beat-synchrony by an online point-to-point cubic spline interpolator. All movements were designed *a priori* to be performable on "beat-time" while providing a smooth dancing performance. As depicted in Fig. 2, in order to ensure the desired beat-synchrony, the robot dancing generator triggers new step transitions at the time of each predicted beat event. In order to overcome processing and communication delays between the step transition request and the actual motion response, the predicted time of each next beat event, b'_{n+1}, is given by:

$$\begin{cases} b'_{n+1} = b_n + \Delta b - d_n \\ \Delta b = b_n - b_{n-1} \end{cases}, \qquad (2)$$

where Δb is the current IBI (Inter-Beat-Interval) estimation given by the time-difference of the last two beat events, b_n and b_{n-1}, estimated by IBT and d_n is the delay of the last

robot motion response. This delay is re-calculated at the time of every predicted beat event, b_n, as follows:

$$d_n = r_{n-1} - b'_{n-1}, \tag{3}$$

where b'_{n-1} is the timing of the previous beat event prediction, and r_{n-1} represents the timing of the motion response to the last step transition request. This response timing is given by the time-frame, n, at which the robot started moving in response to the last step transition request:

$$r_{n-1} = \arg_n[E(n)] : E(n) > s_{thres}, \tag{4}$$

where $E(n)$ represents the mean robot's joints velocity at time-frame n, given by (1), and $s_{thres} = 0.1$ is an empirical threshold value for E that marks the boundary at which the robot is considered to be stopped or moving. Based on this scheme, if the robot is moving at the time of a new step transition request, the current, unfinished, step is immediately transited to the next. If, on the other hand, the robot already finished the current step, it halts until the time of the next beat event prediction, before transiting to the next step. This strategy assures the beat-synchrony of the motion despite the motor-rate capabilities of the robot.

2) Verbal communication – speech-driven dialoguing: The robot's verbal communication is handled by a Question/Answer (Q/A) dialogue management system based on [8]. This dialogue manager receives dialog requests from the ASR module, decides on dialog states, and generates meaningful responses through action commands sent to the *behavior decision mechanism* or speech responses synthesized with VOCALOID.

D. Behavior decision mechanism

The behavior decision mechanism integrates the modules of the last three layers of the framework (*i.e.*, the *Binding*, *Mediation*, and *Behavior* layers) and is responsible for coordinating the human-robot interaction according to the reliability of the acoustic signals for auditory processing. This reliability is measured through the *confidence functions*, $cf(n)$, provided by the modules of the *Perception* layer, *i.e.*, by the ASR, the ABT, and the ENS modules (see Section IV-B). These respectively inform about the signal's conditions for speech, S, and music, M, processing, and about its level of ego noise "contamination", E. These functions are converted into discrete *costs*, C_Y, according to empirically selected thresholds, T_Y:

$$C_Y(n) = \begin{cases} 1, & \text{if } cf_Y(n) < T_Y \\ 0, & \text{if } cf_Y(n) \geq T_Y \end{cases}, \tag{5}$$

where $Y = \{M, S, E\}$ represents the acoustic modality considered by C_Y. Ultimately, these costs are weighted and combined into *fitness functions*, $F_M(n)$ and $F_S(n)$. These represent the policies which mediate different modalities of interactive behaviors, *i.e.*, respectively music-driven and speech-driven (see Fig. 3), according to the reliability of the acoustic signal for music, M, and speech, S, processing:

$$\begin{cases} F_M(n) = W_M^S C_S(n) + W_M^M C_M(n) + W_M^E C_E(n) \\ F_S(n) = W_S^S C_S(n) + W_S^M C_M(n) + W_S^E C_E(n) \end{cases}, \tag{6}$$

where W_X^Y represents the discrete weight assigned to the cost C_Y for the behaviors' modality X. These weights assume a measure of relevance about each cost for each specific modality, enabling the representation of the policies into general and extendable fitness functions. They permit the disconsideration (*e.g.*, setting the W_M^S to zero since C_S does not inform about the reliability of the signal for music processing) and/or the emphasis of specific cost functions (*e.g.*, giving higher weights for W_M^M or W_S^S since C_M and C_S are the most relevant for their own modalities). Thus, in our implementation the weights in (6) were set as: $W_M^S = 0$, $W_M^M = 2$, $W_M^E = 1$, $W_S^S = 2$, $W_S^M = 0$, and $W_S^E = 1$.

These fitness functions can assume different levels of fitnesses (*i.e.*, decisions) that would trigger a different class of behaviors for each interactive modality. According to the deliberation of the assigned behaviors these can be discriminated into a pseudo-continuous action-spectrum ranging from *active* actions – low-priority responsive actions generated to maintain the interaction; to *proactive* actions – high-priority anticipative actions generated to assure a reliable and robust interaction. These actions are depicted in Fig. 3 for each specific modality of interactive behaviors.

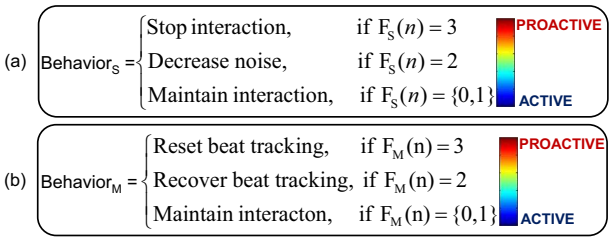


Fig. 3. Behavior decision mechanism based on different classes of actions, from active to proactive, for regulating two modalities of interactive behaviors: (a) speech-driven; (b) music-driven.

Finally, to reinforce the decision's proactivity about the speech-driven behaviors, whenever the ASR is not able to recognize a speech command it assigns $F_S(n) = 2$, regardless of the current $F_S(n)$ value.

E. Software specifications

All of the system's modules were implemented and integrated into *HARK (HRI-JP Audition for Robots with Kyoto University)*. The robot control and communication were handled by *ROS (Robot Operating System)*. The captured audio signals were processed at time increments of 10 ms, using a Complex window of 512 samples and 32% overlap for computing the audio spectrum. IBT was set with an induction window of 5 sec in length, and constrained to a tempo octave between 40 and 80 beats-per-minute (bpm). The restriction to a tempo octave was to avoid metrical-level interchanges that would compromise the stability of the system. This interval represents half[1] of the tempo range defined by the "preferred tempo-octave" (*i.e.*, 80 to 160 bpm), which fits the majority of tempi distributions [16]. The restriction to low tempi was to respect the robot's

[1]Considering the music tempo as a duple multiple of the actual perceptual tempo is acceptable for binary meters.

motor limitations (limited to ≈80 bpm) while dancing in beat-synchrony. The template database for the ego noise suppression module was built with the three dance motions described in Section IV-C.1 generated at random tempi, also in the range of 40 to 80 bpm, for a total of 5 min. The ASR engine was configured with one acoustic-model in Japanese, for the experimental testing of our system; and one other in English, for the interaction session. The Japanese model was trained with a corpus from the Japanese Newspaper Article Sentences (JNAS) containing 60 hours of speech from 306 speakers (male and female). The English model was trained with a corpus from the Wall Street Journal containing 206 hours of speech from 180 male and 181 female.

F. Hardware specifications

The implemented system was run on HRI-JP's humanoid robot HEARBO (see Fig. 2). HEARBO possesses an 8-channel omni-directional microphone array on top of its head. All audio signals were synchronously captured from the 8 channels, at a 16 kHz sampling rate. The running processes were distributed between two PCs, each possessing an Intel i7 Quad Core CPU at 2.3 GHz and 16 GB of RAM, and both connected to HEARBO via Ethernet.

V. Experimental setup

At first, we tested the implemented modules of our system by recreating the real-world acoustic challenges implicated in the proposed interactive scenario, but excluding the actual human interaction. This testing scenario consisted of the robot dancing for 10 min in beat-synchrony to a continuous musical stimulus while simultaneously recognizing a set of human utterances. The mixed auditory signal was contaminated by the robot's ego-motion noise. The musical and speech stimuli were simultaneously played from 2 loudspeakers standing 1 m away and respectively at -60° and 60° from the robot position. The music signals were recorded with a Music Signal-to-Noise Ratio (M-SNR) of -2 dB. The speech signals were recorded with a segmental Speech SNR (S-SNR) of -3 dB. For testing purposes, the dance motions were generated in beat-synchrony to the musical stimuli from the annotated beat times. Each dance motion was continuously generated for $\frac{1}{3}$ of the recording duration before changing to the next. We recorded 8 audio signals, each with 10 min, by using the speech from a different speaker per recording. All recordings were processed in a noisy room with the dimensions of 4.0 m x 7.0 m x 3.0 m and a Reverberation Time (RT20) of 0.2 sec.

A. Auditory signals

1) Musical stimuli: To reproduce the realistic scenario of continuous and dynamic musical stimuli, we used the audio data stream described in [13]. This data consisted of a single 10 min audio file built of a set of 31 musical excerpts, with 20 sec each, concatenated without any gaps This reproduces highly challenging timing and tempo transitions between the excerpts. The selected musical pieces comprised 7 different genres: *pop*, *rock*, *jazz*, *hiphop*, *dance*, *folk*, and *soul*; with tempi ranging from 80 to 140 bpm, with a mean of 109±17.6 bpm; and all with a $\frac{4}{4}$ meter.

2) Speech data: We recorded 8 audio files with the utterances of 4 male and 4 female Japanese speakers used in a typical human-robot interaction dialog, and previously used in [13]. Each audio file comprised a set of 236 different Japanese words concatenated into a continuous stream, with a silence gap of roughly 1 sec between each of them.

B. Evaluation criteria

1) Beat tracking accuracy: To quantify the standard performance of the live beat tracking on the music stream we relied on the AMLt (Allowed Metrical Levels, continuity not required), as described in [13]. Akin to [13], we considered two variants of the AMLt: $AMLt_s$, which measures the accuracy over the whole stream; and $AMLt_e$ that simulates the individual evaluation over the concatenated excerpts by measuring the accuracy of the whole stream but discarding the first 5 sec after each music transition.

2) Reaction time (r_t): To measure the reaction time, r_t, at each music transition we also followed [13] and defined $r_t = |b_r - t_t|$ as the time difference, in seconds, between the timing of the transition, t_t, and the first beat-time, b_r, of the first four continuously correct beats in the considered musical excerpt. In addition, a music transition was considered successful if the system could recover track of the beats at some point after transiting to the current musical excerpt.

3) ASR accuracy: The ASR accuracy was measured in terms of average Word Correct Rate (WCR), which is defined as the number of correctly recognized words from the test set divided by its total number of instances.

4) Dance beat-synchrony: For measuring the beat-synchrony of the generated dance, we also used the proposed variants of the AMLt score: $AMLt_s$ and $AMLt_e$. These compared the time-alignment of the annotated beat times of the musical stream (which were also used for synchronizing the robot motion in the experimental tests) with the timings of the dance step transitions. To retrieve the timings of the dance step transitions we applied a "valley-picking" algorithm on the mean velocity signal, given by (1), and retrieved the timings of the mean velocity minima.

C. Compared variants of the system

In order to assess the proposed system under the presented experimental conditions, the ABT and ASR accuracies were measured using different input signals, resultant from applying different preprocessing strategies:

- `1Ch`: audio captured from a single (frontal) microphone;
- `1Ch+ENS`: `1Ch` refined by ENS;
- `8Ch`: separated signals by applying SSL and SSS on the audio captured from an 8-channel microphone array. The separated speech and music signals are respectively sent to the ASR and beat tracking modules;
- `8Ch+ENS`: `8Ch` refined by ENS.

In addition, to observe the effect of regulating the acoustic environment for beat tracking purposes, we compared the performance of IBT over a non-regulated acoustic

signal, `IBT-default`, against the IBT performance over a regulated acoustic signal, `IBT-regulated`, through requests to *regenerate* or *reset* the beat tracker when facing unreliable acoustic conditions for music processing (see Section IV-D).

VI. Experimental results and discussion

A. Experimental results

All results represent the mean over the 8 recordings (one per speaker as described in Section V). Fig. 4 presents the mean dance beat-synchrony results, in terms of $AMLt_s$ and $AMLt_e$ scores, for the whole recording (Fig. 4a), and the distribution of the $AMLt_e$ score in function of the musical tempo in increments of 5 bpm (Fig. 4b).

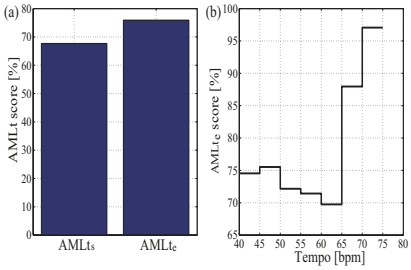


Fig. 4. Dance beat-synchrony: (a) overall results; (b) results per musical tempo.

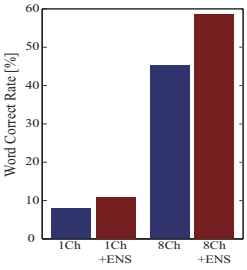


Fig. 5. ASR accuracy.

Fig. 5 presents the mean ASR results for all variants of the system. Fig. 6 presents the overall beat tracking accuracy for `IBT-default` (red) and `IBT-regulated` (blue) in terms of $AMLt_s$ (dark) and $AMLt_e$ (light) scores (Fig. 6a), and in terms of mean reaction time and number of successfully handled transitions in the tested music data stream (Fig. 6b).

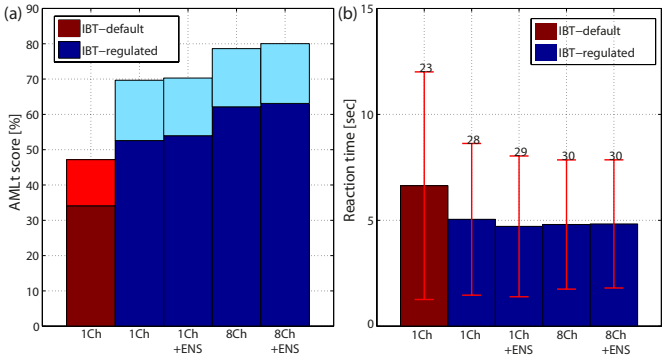


Fig. 6. Beat tracking accuracy: (a) Mean AMLt scores: $AMLt_s$ (dark) and $AMLt_e$ (light); (b) Mean reaction time (r_t) and standard deviation among the music transitions, and total number of successful transitions (on top).

B. Discussion

1) On the dance beat-synchrony results: As observed in Fig. 4a, our algorithm for generating beat-synchronous robot dancing motions was able to reproduce up to 67.7% of overall beat-synchrony, in terms of $AMLt_s$ score (*i.e.*, when continuously considering the whole 10 min music data stream), and up to 75.9% in terms of $AMLt_e$ score, when discarding the first 5 sec after each music transition. The 8.2 percentage points (pp) difference among these results is justified by the abrupt beat and tempo changes within music transitions that demand abrupt variations in the robot motor velocities at these instants. This compromise the motion's beat-synchrony until it reaches a stable state. When analyzing the dance beat-synchrony in relation to the musical tempo (Fig. 4b) we observed high discrepancies between low and high tempi, in the order of 20 pp difference in $AMLt_e$ score, with a performance threshold at around 65 bpm favoring higher tempi. The performance differences can be explained by the way we retrieve the timings of the dance step transitions, determined by the mean velocity minima. It is more accurate to detect peaky velocity transitions demanded by higher tempi (*i.e.*, by faster transitions) than flat velocity transitions demanded by lower tempi (*i.e.*, by slower transitions). Despite the results of Fig. 4b, we argue that the robot dancing motions perceptually appeared less synchronous to the beat at high tempi. This can be justified by the faster transitions, which generate quasi-continuous motions that makes it difficult to perceptually acknowledge the step transitions occurring with the beat. Additional subjective discussion on this issue is left for future work.

2) On the audio beat tracking results: From Fig. 6 we clearly observe that, in the presence of continuous and dynamic musical stimuli, robust live music processing algorithms require a running regulation of the acoustic signal conditions, *i.e.*, a running monitor that supervises the acoustic signal for disturbances while requesting the recovery of the system accordingly. This is reflected in an increase in the beat tracking accuracy of 18.5 pp in $AMLt_s$ and 22.5 pp in $AMLt_e$, when comparing the `IBT-regulated` against the `IBT-default` in the `1Ch` recorded signal. This increase in accuracy is proportionally reflected in a decrease of 1.6 sec in the reaction time at music transitions, when comparing both variants under the same `1Ch` signal condition. Moreover, `IBT-regulated` was able to recover from music transitions with a mean reaction time of 4.9 ± 2.0 sec across all signal conditions, and without statistical significances among the results (with a mean $p = 0.76 \pm 0.18$). Ultimately, we observed that `IBT-default` in `1Ch` could only manage to handle 23 out of the 30 music transitions in the data stream, whereas `IBT-regulated` could handle 28 of them in the same condition. When applying SSS after SSL (*i.e.*, on the `8Ch` signal) we improved beat tracking by 9.5 pp in $AMLt_s$ and 8.9 pp in $AMLt_e$, respectively up to 62.1% and 78.6%, and `IBT-regulated` already handled all music transitions.

Finally, by applying ego noise suppression we improved the beat tracking accuracy on `1Ch` and `8Ch` signals by on average 1.2 pp in $AMLt_s$ and 1.0 pp in $AMLt_e$. This resulted in an optimal beat tracking performance, for `8Ch+ENS`, of 63.1% in $AMLt_s$ and 80.0% in $AMLt_e$, and a mean 4.8 ± 3.0 sec in reaction time. Although the ENS was still able to slightly improve the beat tracking performance its improvement was not statistically significant either for `1Ch` or `8Ch` (respectively, $p = 0.93$ and $p = 0.79$). This is justified by the fact that the ego-motion noise was synchronized to the musical beat, which resulted in some of the beats getting suppressed along with the motion noise, hence decompensating the overall enhancement provided by ENS.

3) On the ASR results: As depicted in Fig. 5, the application of SSL and SSS as preprocessing (*i.e.*, the `8Ch`

signal) greatly improved the ASR results by on average 35.8 pp. When additionally applying ENS these results were improved by 13.3 pp, achieving an optimal WCR of 58.5%.

VII. Interactive robot dancing session

In order to test the proposed active audition framework in a human-robot interactive situation, we set up a scenario in which the humanoid robot must simultaneously dance to the beat of the music and respond to human speech commands related to the music it is listening to or to the dance it is performing. To test the full capability of the proposed framework to ensure a reliable interaction, the defined scenario included the following requirements:

- Simultaneous music and speech processing;
- Live beat/tempo tracking to continuous musical stimuli;
- Dealing with multiple noise sources of different natures;
- Autonomous beat-synchronous robot dancing;
- Q/A dialoguing;

The interaction was set up in the same room described in Section V. Akin to the experimental settings, the musical stimulus was played from a single loudspeaker standing at -60° from the robot's position, and the human speaker stood at 60° from it (see Fig. 7). Both were positioned 1 m away from the robot. The musical volume was kept the same as in the experimental tests. The human speaker also tried to reproduce the S-SNR verified in the experimental tests. The speech was performed in English. The dancing motions were interchanged on human command. IBT performed in the `IBT-regulated` mode. In order to optimize the accuracy of the ASR and ABT, based on the previous results (see Fig. 5 and Fig. 6), both tasks were processed over the `8Ch+ENS` signal conditions. In order to clearly accompany the beat tracking performance during interaction, the beats predicted by the robot generated noise clicks through one of the room's speakers (far enough to not bias the robot's beat estimations).

A. Musical stimuli

The musical stimuli used in the interaction comprised three musical pieces of Japanese Pop music from the RWC Music Database [17]. In addition, one of them was synthesized into three different mood arrangements at different tempi. The 5 selected musical pieces comprised different tempi in the range of [100-133] bpm. The music and/or mood changes were requested by the human speaker at arbitrary moments of the interaction. The names of the songs and singers were provided in advance to the dialogue management system.

B. Interaction results

A video with a full robot dancing interaction session is presented in our website[2]. Fig. 7 presents screenshots of some key-moments of the interaction. Fig. 8 depicts the main events that occurred during the presented interactive session: *(a)* the direction of the acoustic sources (*i.e.*, music at ≈-60° and speech at ≈60°) detected by the SSL; *(b)* the robot joints' angular position, in degrees; *(c)* the robot joints' angular velocity, in degrees/frame; *(d)* the Speech Decisions given by the speech-related fitness function, $F_S(n)$; *(e)* the Music Decisions given by the music-related fitness function, $F_M(n)$; and *(f)* the actual interaction events given by the *Human* speech commands (in black), the playing *Music* (in blue), and the *Robot* speech responses (in red).

[2] `http://smc.inescporto.pt/wp-content/uploads/2012/06/RoMan2012Demo.avi`

C. Discussion

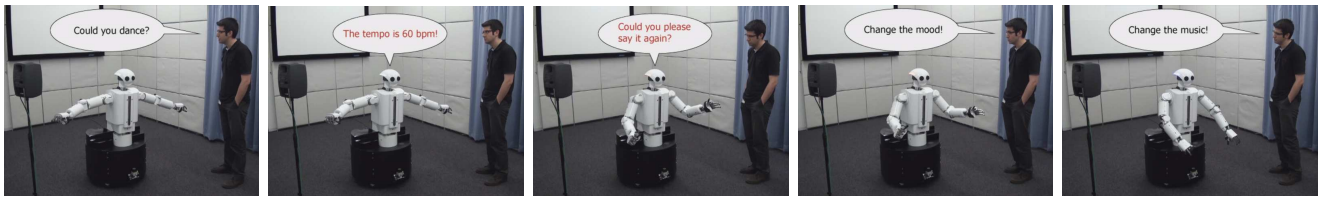
The recorded human-robot live interactive session depicted in Fig. 8 ran uninterruptedly for more than 2.5 min and was replicated a few times. This demonstrates the robustness of the implemented interactive robot dancing system on tackling the present highly challenging conditions. Moreover, as observed in the video, the human interactor was able to enjoy from a natural interaction with the robot.

Specifically, we observed the capacity of the system to handle perturbations in the acoustic signal. In terms of speech processing, we observed two critical moments where the behavior decision intervenes towards improving the noise conditions for ASR. As illustrated in Fig. 8d, these occurred at 70 sec, due to an unrecognized speech command despite the Speech Decision to "Maintain Interaction" ($F_S = 0$); and at 79 sec, due to a Speech Decision to "Decrease Noise" ($F_S = 2$) resultant from a low confidence in the ASR engine. In terms of music processing, from Fig. 8e we observed several requests to regenerate/reset the beat tracker due to the abrupt music transitions requested on human command, or due to the contaminating noise sources of different natures (*i.e.*, ego-noise and human/robot speech). The higher number of critical Music Decision events in comparison to critical Speech Decision events is justified by the differences in the resolution of both modalities (*i.e.*, continuous music stimuli against discrete speech commands).

In terms of responsiveness of the system, we observed a latency in the speech response of around 3.0 sec, which is justified by the computation of the HMM search of the ASR module. The beat tracker revealed high reaction times to music changes, up to a maximum of 11.0 sec, which were significantly higher than in the experimental results, where it revealed a maximum r_t of 7.8 sec (see Fig. 6b). These are justified by higher inconsistencies in the generated dance motions at music changes due to relying on the beats estimated by the beat tracker while it is attempting to recover. These caused more abrupt and intense ego noise variations which perturbed IBT's recovery to the music transitions.

VIII. Conclusions and future work

We proposed a general and extensible active audition framework for auditory-driven HRI. This framework simultaneously processes speech and music on-the-fly, integrates perceptual models for robot audition, supports verbal and non-verbal interactive communication, and integrates a behavior decision mechanism based on active audition for conducting a reliable human-robot interaction. This framework was applied to an interactive robot dancing system and assessed in both experimental and interactive real-world scenarios. Experimental tests on the system



(a) Human requests the robot to start dancing [t=18 s]. (b) Robot provides the musical tempo [t=29 s]. (c) Robot asks to repeat the speech command [t=79 s]. (d) Human requests the robot to change the mood [t=86 s]. (e) Human requests the robot to change the music [t=102 s].

Fig. 7. Key-moments of the recorded robot dancing interactive session.

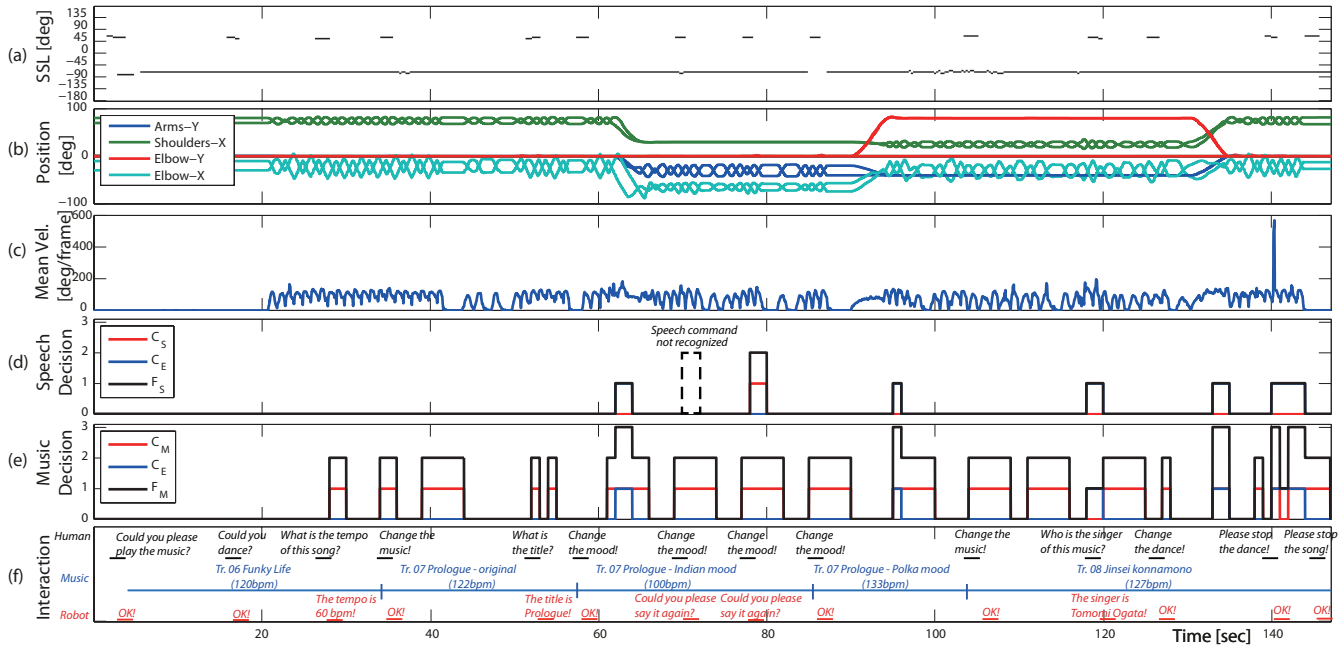


Fig. 8. Illustration of the events occurring in the recorded robot dancing interactive session.

revealed accurate beat-synchrony of the generated robot dance motions and improved beat tracking and ASR accuracies. Tests on interactive robot dancing sessions proofed the robustness of the proposed framework in the presence of highly dynamic acoustic conditions. The overall results confirmed this framework's application to general auditory-driven interactive robotic systems.

In the future we will assess the robustness and interactivity of the developed interactive robot dancing system through subjective evaluation. We should also extend the proposed framework with additional robot audition modules, robot behaviors and behavior decision policies. Finally, we will improve the regulation of the acoustic signals by integrating additional costs and more reliable confidence functions, and test the framework in other auditory-driven HRI applications.

REFERENCES

[1] K. Nakadai *et al.*, "Active audition for humanoid," in *National Conference on Artificial Intelligence*, 2000, pp. 832–839.

[2] H. G. Okuno and K. Nakadai, "Computational Auditory Scene Analysis and its Application to Robot Audition," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 124–127.

[3] G. Weinberg and S. Driscoll, "Toward Robotic Musicianship," *Computer Music Journal*, vol. 30, no. 4, pp. 28–45, Dec. 2006.

[4] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon: A Playful Robot for Research, Therapy, and Entertainment," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 3–18, Nov. 2008.

[5] M. Nakano *et al.*, "A multi-expert model for dialogue and behavior control of conversational robots and agents," *Knowl.-Based Syst.*, vol. 24, no. 2, pp. 248–256, 2011.

[6] F. Tanaka *et al.*, "Daily HRI Evaluation at a Classroom Environment – Reports from Dance Interaction Experiments," in *Conference on Human-Robot Interaction (HRI)*, 2006, pp. 3–9.

[7] J. L. Oliveira *et al.*, "Online Audio Beat Tracking for a Dancing Robot in the Presence of Ego-Motion Noise in a Real Environment," in *IEEE ICRA*, 2012, pp. 403–408.

[8] K. Nakadai *et al.*, "A robot referee for rock-paper-scissors sound games," in *IEEE ICRA*, 2008, pp. 3469–3474.

[9] H. Okuno *et al.*, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *IROS*, 2001, pp. 1402–1409.

[10] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[11] H. Nakajima *et al.*, "Blind Source Separation with Parameter-Free Adaptive Step-Size Method for Robot Audition," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 18, no. 6, pp. 1476–1484, 2010.

[12] G. Ince *et al.*, "Online Learning for Template-based Multi-Channel Ego Noise Estimation," to appear in IEEE/RSJ IROS, 2012.

[13] J. L. Oliveira *et al.*, "Live Assessment of Beat Tracking for Robot Audition," to appear in IEEE/RSJ IROS, 2012.

[14] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2009.

[15] J. L. Oliveira *et al.*, "IBT: A Real-time Tempo and Beat Tracking System," in *ISMIR*, 2010, pp. 291–296.

[16] D. Moelants, "Dance Music, Movement and Tempo Preferences," in *5th Triennial ESCOM Conference*, 2003, pp. 649–652.

[17] M. Goto *et al.*, "RWC Music Database: Popular, classical, and jazz music databases," in *ISMIR*, 2002, pp. 287–288. [Online]. Available: http://staff.aist.go.jp/m.goto/RWC-MDB/