2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
September 14-18, 2014. Chicago, IL, USA,

# Making a Robot Dance to Diverse Musical Genre in Noisy Environments

João Lobato Oliveira[1], Keisuke Nakamura[2], Thibault Langlois[3], Fabien Gouyon[4], Kazuhiro Nakadai[2], Angelica Lim[5], Luis Paulo Reis[1,6], and Hiroshi G. Okuno[5]

*Abstract*— **In this paper we address the problem of musical genre recognition for a dancing robot with embedded microphones capable of distinguishing the genre of a musical piece while moving in a real-world scenario. For this purpose, we assess and compare two state-of-the-art musical genre recognition systems, based on Support Vector Machines and Markov Models, in the context of different real-world acoustic environments. In addition, we compare different preprocessing robot audition variants (single channel and separated signal from multiple channels) and test different acoustic models, learned a priori, to tackle multiple noise conditions of increasing complexity in the presence of noises of different natures (e.g., robot motion, speech). The results with six different musical genres suggest improved results, in the order of 43.6pp for the most complex conditions, when recurring to Sound Source Separation and acoustic models trained in similar conditions to the testing scenarios. A robot dance demonstration session confirms the applicability of the proposed integration for genre-adaptive dancing robots in real-world noisy environments.**

## I. Introduction

Dance expresses the deepest parts of our being in a way no words or book could ever do. Almost every culture in the world has music and dance. Dancing, through its group synchronization and body movement, is a fun activity that powerfully bonds people together [1]. While at times shyness make it difficult to engage in collective dance, a dancing robot has the potential to entertain and unite people of various ages and backgrounds. In festivals or events, dancing robots could entrain the people around it to move their bodies, when music alone is not enough to fill an empty dance floor.

In this paper, we propose musical genre recognition for a dancing robot in noisy environments. Our long-reaching goal is a robot that can dance to live music or an arbitrary piece of music. Towards this goal, we need to make a robot recognize the genre of the music to perform appropriate movements, such as head-banging to rock music, or hip hop moves for popular music. This is difficult because, although offline genre classification for clean music signals is well-studied [2], live recognition with noisy auditory input has never been attempted. In addition, the robot's own motor noises (the so-called ego noise [3]) during dancing further degrade the music signal [4].

To this extent, we assessed and compared two state-of-the-art musical genre recognition algorithms, based on Support Vector Machines (SVM) [5] and Markov Models (MM) [6], with two different preprocessing robot audition variants (single channel and separated signal from multiple channels) and the use of acoustic models learned *a priori* under different noise conditions. We assessed all these variants in terms of genre recognition accuracy with six musical genres, using a most common dataset used in MIR (Music Information Retrieval) [7]. To verify the applicability of the proposed integration for dancing robots, we conclude this paper by introducing a demonstration session of a dancing robot that is able to quickly adapt on-the-fly to the musical genre in a real-world noisy environment.

## II. Related Research

The automatic genre recognition of musical content has been widely studied for the last two decades to give response to the increasing amount of musical data stored in music databases, catalogues, libraries, and stores, which need to be categorized [2]. More recently, and with the improvement of pattern recognition and machine learning techniques, we can already find musical genre classifiers installed in online music recommendation services [8], radio-on-demand broadcasting [9], and even in cars [10].

However, musical genre recognition has never been attempted in musical robotics under real-world scenarios, whereby experiments were yet much restrict to beat tracking [4] and mood classification [11], [12]. Moreover, when dealing with different noise conditions for robot audition, much of the work has been undertaken on environmental sound source identification [13] and speech recognition [3], [14], [15].

Considering the latter as the most investigated topic in robot audition, different strategies are used in order to enhance speech recognition under multiple noise conditions. These include the use of multiple acoustic models, trained under different noise conditions, and the use of Sound Source Separation (SSS) to recognize the speech of three different speakers [14]; the use of compensation and adaption methods to reduce the mismatch between the training and test conditions [15]; and the use of ego noise suppression strategies to tackle the unpredictable diffuse noise generated by the robot motion while performing automatic speech recognition [3] or beat tracking [4].

[1] Artificial Intelligence and Computer Science Laboratory (LIACC) – FEUP, Porto, Portugal. (joao.lobato.oliveira@fe.up.pt)
[2] Honda Research Institute Japan Co., Ltd., Saitama, Japan.
[3] Science Faculty of the University of Lisbon (FCUL), Lisbon, Portugal.
[4] Institute for Systems and Computer Engineering of Science and Technology (INESC TEC), Porto, Portugal.
[5] Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan.
[6] University of Minho, School of Engineering - DSI, Guimarães, Portugal.

Inspired by [14], we tested acoustic models learned in different noisy conditions and multi-channel SSS in order to enhance the musical genre recognition of a dancing robot while moving in the presence of background and speech noises in a real-world environment.

## III. Musical Genre Recognition

As depicted in Fig. 1, our musical genre recognition robotic system integrates two state-of-the-art genre classification algorithms, one based on Support Vector Machines and other based in Markov Models, and considers two preprocessing robot audition variants: single channel and separated signal from multiple channels through Sound Source Separation.

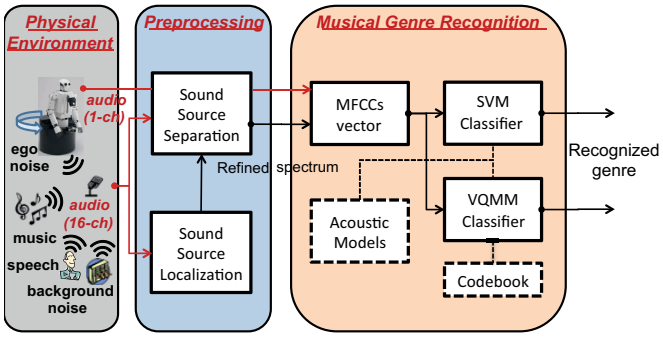



Fig. 1. System architecture.

### A. Preprocessing robot audition modules

*1) Sound source separation:* The SSS module is responsible for splitting the captured audio signal into individual sound sources discriminated by their given directions. These directions are typically measured by means of Sound Source Localization (SSL), but in order to assure a continuous signal acquisition, we directed the SSS to a particular sound source direction towards the musical source. The integrated SSS implementation applies the Geometric High-order Decorrelation-based Source Separation (GHDSS) [16].

### B. Genre classification algorithms

Genre classification typically recur to supervised machine learning algorithms that infer genre information from low-levels features extracted from the musical signal. These features may be related to different dimensions of music, including melody, harmony, rhythm, timbre, and spatial location [2].

*1) Audio Features:* In order to focus the assessment of this paper to the comparison of different genre classification algorithms and pre-processing robot audition variants, in the context of different real-world noisy conditions, we restrain our features to timbre, in the form of the most popular Mel-Frequency Cepstrum Coefficients (MFCC). These model the short-time spectral characteristics of the signal onto a psychoacoustic frequency scale. We selected the 12 first MFCCs for the feature vector which is used as input to both genre classification algorithms.

*2) SVM-based Genre Classifier:* This musical genre classification algorithm was proposed and described in [5], and implemented in MARSYAS [1]. This algorithm starts by computing a running mean, $m_{\theta(t)}$, and standard deviation, $s_{\theta(t)}$, over the past $M = 1$ sec (in frames) of the feature vector. These $m_\theta$ and $s_\theta$ are then collapsed into a single feature vector representing all extension of the considered audio clip by calculating the mean and standard deviation across the whole clip:

$$\begin{cases} m_{\theta(t)} = \text{mean}[\theta(t-M+1),..,\theta(t)] \\ s_{\theta(t)} = \text{std}[\theta(t-M+1),..,\theta(t)] \end{cases} . \quad (1)$$

This results in a 24-dimension feature vector, which is further normalized. This feature vector is used for both training and test stages of the genre classification by recurring to a multi-class Support Vector Machine. This algorithm is hereafter referred to as SVM.

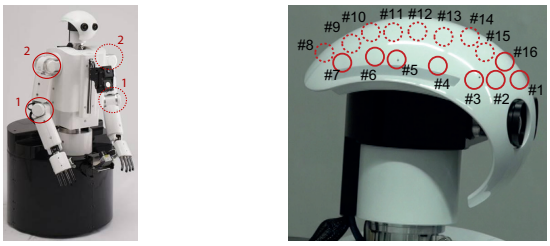
*3) Markov Model-based Genre Classifier:* This musical genre classification algorithm was proposed and described in [6]. This algorithm starts by quantizing the 12-dimension MFCC feature vector using a hierarchical clustering approach, based on Gaussian Mixture Models (GMM) and the K-means algorithm, to create clusters that can be interpreted as codewords in a dictionary.

The training data is modeled for each class as codeword transition matrices based on probability Markov Models. For classification, the incoming data is also modeled as probability transition matrixes based on Markov Models, which are then compared to each training model. The resulting classification is given by the training model that best fits the transition matrix of the incoming data. This algorithm is hereafter referred to as VQMM.

## IV. Experimental settings

### A. Hardware specifications

Our experiments were run on HEARBO, a humanoid robot from *Honda Research Institute Japan (HRI-JP)* (see Fig. 2(a)). HEARBO integrates a 16-channel omni-directional microphone array on top of its head (see Fig. 2(b)). All audio signals were synchronously captured from the 16 channels, at a 16 kHz sampling rate. All recordings and evaluation procedures were processed on an Intel Core i7 quadcore PC at 2.3 GHz, with 16 GB of RAM.




(a) Positions and number of moving joints. (b) Close-up of the head.

Fig. 2. HRI-JP humanoid robot HEARBO.

[1] See http://marsyas.info.

### B. Software specifications

All 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 whole system was processed at time increments of 10 ms, using a Complex window of 512 samples and 32% overlap (*i.e.*, hop size of 160 samples) for computing the audio spectrum.

### C. Auditory signals

*1) Musical stimuli:* In order to enable a baseline comparison with state-of-the-art genre classification algorithms, we tested oursystem with one of the most popular dataset used in the MIR community, the ISMIR2004 genre classification dataset [7]. This dataset is originally composed of 1458 full music recordings of 6 different musical genre classes annotated by experts: *classical*, *electronic*, *jazz-blues*, *metal-punk*, *rock-pop*, and *world*. We trimmed each recording to 30 sec extracted from the middle of each song. We used all 1458 clips for training acoustic models under different conditions, and selected 120 clips (20 from each genre) where both algorithms scored 100% (using the F-Score described in Section IV-E.1) under clean conditions for the recognition tests.

*2) Speech data:* The speech data consisted of 8 audio files with the utterances of 4 male and 4 female Japanese speakers used in a typical human-robot interaction dialog. Each audio file was constituted by a set of 236 different Japanese words concatenated into continuous streams, with a silence gap of $\approx$1 sec in between them. Each was individually played at a time.

### D. Periodic dance motions

We measured the effect of ego-motion noise in the musical genre recognition accuracy by using the same 3 periodic dance motions used in [4]. Each of them was composed of 2 key-poses interpolated (*i.e.*, transited) during motion generation. In order to maximize the disturbing effects of the robot's ego noise, the dance motions were designed to simultaneously move 6 joints: the shoulders *pitch* and *yaw*, and the elbows *pitch* (see Fig. 2(a)); each with a rotational variation in the range of [10-20]°, thus also maximizing the number of transitions. The dance motions were continuously generated and interleaved during recordings for generating dance sequences with a uniform number of periodic repetitions of the 3 dances. The periodic dances were generated at random tempi (*i.e.*, random velocities) in the octave of 40 to 80 bpm, which represent the maxima motor-rate frequencies achievable by our robot.

### E. Evaluation criteria

In order to assess and compare the two algorithms integrated in the system based on different preprocessing variants and different auditory conditions, we performed a 10-fold cross validation where we normalized the frequency of each genre class per fold.

*1) F-Score:* We recurred to the classic F-Score to assess the algorithms' overall classification accuracy among all genre classes, which is calculated for each fold as follows:

$$\begin{cases} F = \text{mean}[f_c, ... f_C] \\ f_c = 2 \times \frac{\text{precision}_c \times \text{recall}_c}{\text{precision}_c + \text{recall}_c} \end{cases}, \quad (2)$$

where $C$ is the number of genre classes, per-class $precision_c$ represents the fraction of music clips classified as class $c$ that were annotated as $c$, and per-class $recall_c$ represents the number of music clips annotated as class $c$ that were actually classified as $c$. The overall F-Score is calculated as the mean of the F-Scores of each fold.

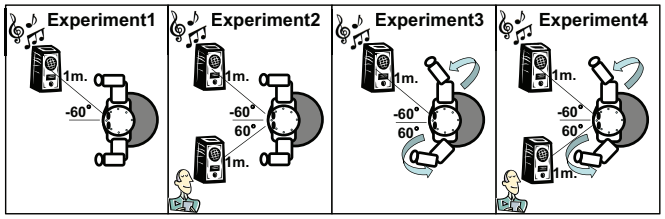



Fig. 3. Experiments for the four proposed real-world acoustic conditions with increasing levels of noise complexity.

## V. Experiments and Results

As illustrated in Fig. 3, and akin to the experiments run in [4], we created four real-world experimental conditions to assess our musical genre recognition systems in incremental levels of noise complexity:

- **Experiment1:** musical genre recognition under background noise.
- **Experiment2:** musical genre recognition under background and speech noises.
- **Experiment3:** musical genre recognition under background noise and ego noise from the robot dance motion.
- **Experiment4:** musical genre recognition under background and speech noises and ego noise from the robot dance motion.

Akin to [4], the musical stimulus was played from a single loudspeaker standing at -60° and 1 m away from the robot position in all experiments. The music signals were recorded with decreasing Music-Signal-to-Noise Ratio (M-SNR) among the four experiments, using the recording of `experiment1` as a baseline: $M\text{-}SNR = 1\,dB$ for `experiment2`, $M\text{-}SNR = 0\,dB$ for `experiment3`, and $M\text{-}SNR = -2\,dB$ for `experiment4`. For the experiments using speech stimuli (*i.e.*, `experiment2`, and `experiment4`) we played it from a second loudspeaker standing at 60° and also 1 m away from the robot.

All recordings were processed in a noisy room environment with the dimensions of 4.0 m x 7.0 m x 3.0 m and a Reverberation Time ($RT_{20}$) of 0.2 sec.

### A. Acoustic models

Besides testing different preprocessing conditions, we tested the recognition accuracy using acoustic models trained *a priori* under different noise conditions:

- `A0`: original music clips, in clean conditions.
- `A1`: music clips captured by a single (frontal #1 see Fig. 2(b)) microphone with background noise.
- `A2`: `A1` with synthesized speech noise.
- `A′2`: music clips captured by a 16-channel microphone array with background noise and synthesized speech noise, refined by SSS.
- `A3`: `A1` with synthesized ego noise.
- `A′3`: music clips captured by a 16-channel microphone array with background noise and synthesized ego noise, refined by SSS.
- `A4`: `A1` with synthesized speech and ego noises.
- `A′4`: music clips captured by a 16-channel microphone array with background noise and synthesized speech and ego noises, refined by SSS.

For synthesizing speech and ego noises in the training data, we used the data recorded under background noise as reference and merged variations of the speech and/or ego noises recorded individually without music. All models where trained with all the 1458 30 sec music clips described in Section IV-C.1.

### B. Compared variants of the system

In order to demonstrate the capability of the proposed system under the presented experimental conditions, we evaluated and compared the genre recognition accuracy of both algorithms using different input signals, resultant from different preprocessing strategies:

- `T0`: original music clips, in clean conditions.
- `Tx`: audio captured from a single (frontal #1 – see Fig. 2(b)) microphone under different noise conditions, where x represents the index of the experiment.
- `T′x`: separated audio signal, captured from a 16-channel microphone array, where x represents the index of the experiment.

All test recordings for assessing the musical genre recognition accuracy used the 120 music clips selected as described in Section IV-C.1.

### C. Results

Fig. 4 depicts the baseline genre recognition accuracy of both algorithms using a 10-fold cross validation of the whole 1458 files' dataset under the different experimental conditions. In these results we considered the same conditions of the recognition across all experiments.

Fig. 5(a) and Fig. 5(b) respectively depict the genre recognition accuracy of the SVM-based and MM-based algorithms using a 10-fold cross validation of the 120 music clips recorded under the different experimental conditions, and regarding different pre-processing auditory conditions. In these results we considered multiple acoustic models trained under different noise conditions.

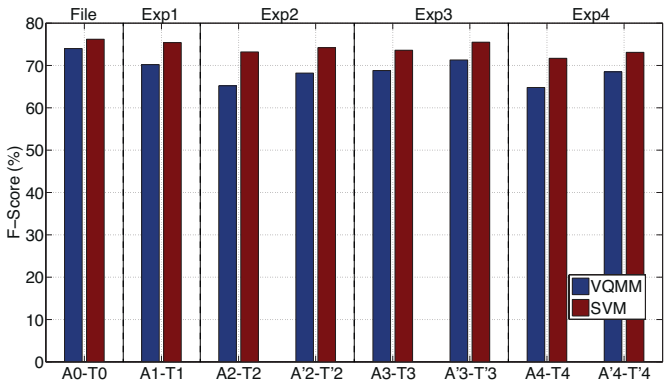



Fig. 4. Baseline musical genre recognition accuracy under different experimental conditions using acoustic models trained under the test conditions.

Fig. 6(a) and Fig. 6(b) illustrate the confusion matrix among all six genre classes when classifying respectively with SVM and VQMM under the A'4 condition.

Ultimately, Fig. 7 illustrates the results of the SVM and VQMM algorithms under the A'4 condition by using time-windows of different sizes to test the real-time accuracy of these algorithms to different amounts of information. The results were measured by splitting each 30 s instance into $30/W$ chunks, where $W$ is the size of the considered time-window (in secs), and classifying each chunk individually.

## VI. Discussion

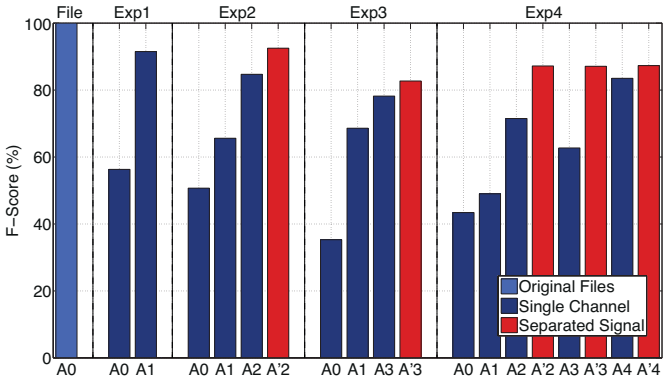
### A. On the use of noisy acoustic models

The baseline results depicted in Fig. 4 revel that, when the acoustic models exactly match the test conditions, the recognition accuracy is statistically equivalent among all experimental conditions. Despite the conditions, both algorithms slightly decreased on average solely 4.1 pp (percentage points) when compared to their 75.1% average accuracy under the clean audio files.

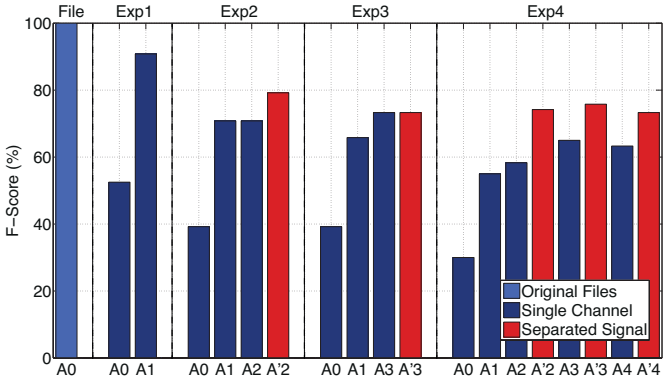
Fig. 5 depicts that the use of acoustic models based on the clean music clips results in extremely low genre recognition accuracies when tested under all noise conditions, by on average 43.3% among both algorithms. Yet, when the acoustic models are trained under synthetic noise conditions that partially simulate their real-world test equivalents, the results tend to improve in proportion to the degree of similarity of both training and test conditions. These results are maximized when the synthetic conditions used to train the acoustic models simulate the real-world test conditions, by improving the results achieved with clean acoustic models by on average 34.8 pp, despite the harshness of the noise conditions.

### B. On the noise robustness to different conditions

By looking into Fig. 4, the disturbing effect of background noise is rather small and on average in the order of 8.9 pp for both algorithms, when considering acoustic models trained under the same condition. As expected, the introduction of speech noise in the test conditions increased the disturbing

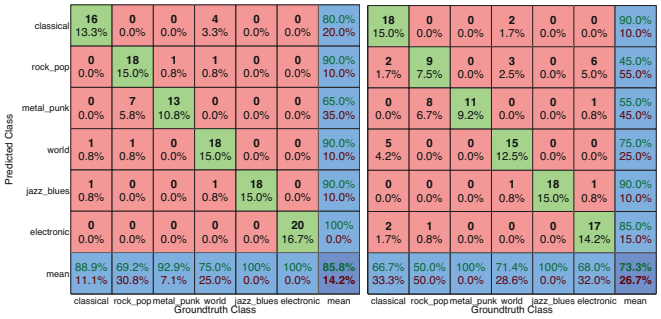



(a) SVM results.




(b) VQMM results.

Fig. 5. Musical genre recognition accuracy under different experimental conditions, considering different preprocessing variants and using acoustic models also trained under different noise conditions.




(a) SVM results. (b) VQMM results.

Fig. 6. Confusion matrix among musical genres, under A'4 conditions.

effect in the performance of both algorithms, by decreasing their musical genre recognition accuracy by an average 13.4 pp, although considering acoustic models trained with synthetic speech noise. The introduction of ego noise in the test conditions caused a slightly bigger decrease in the recognition accuracy of both algorithms, in the order of an additional 2.0 pp, when using acoustic models trained with synthetic ego noise. Ultimately, the disturbing effect when simultaneously introducing speech and ego noises in the recordings caused an average decrease of 17.8 pp below the accuracy under the `experiment1` conditions, when considering acoustic models trained with synthetic speech and ego noises.

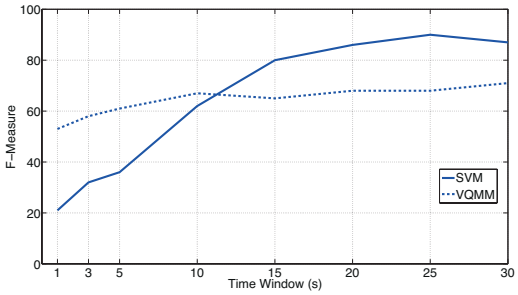



Fig. 7. Real-time accuracy to time-windows of different sizes, under A'4 conditions.

### C. On the use of Sound Source Separation

The results depicted in Fig. 4 reveal that the use of SSS is able to improve the genre recognition accuracy by on average 10.1 pp, when comparing to their single channel equivalents and when considering acoustic models trained in similar synthetic separated noise conditions. The contribution of SSS to this improvement is rather relevant when in the presence of directional noise, as is the case of speech. This is corroborated by the increasing improvement of using SSS under `experiment2` and `experiment4` when compared to `experiment3`, by on average more 9.4 pp among both algorithms. Under the most harsh conditions of `experiment4`, the recourse to SSS and an acoustic model trained in equivalent synthetic conditions enabled an average recognition improvement in the order of 6.9 pp among both algorithms in comparison to their single channel equivalents.

### D. SVM vs. MM

The baseline results depicted in Fig. 4 for both algorithms under the clean audio files suggest that both perform statistically equivalent, with a slight outperformance of the SVM by 2.2 pp. Yet, when directly comparing the baseline and experimental results of both algorithms under all real-world conditions, the SVM outperforms the VQMM by on average 6.5 pp. This suggests that although both are equivalently accurate, the SVM is more robust to noise than the VQMM, probably due to the reliance of VQMM on codebooks, which seem more prone to noise distortions.

### E. Genre Confusion and Real-Time Performance

Not unexpectedly, by looking into Fig. 6, we that verify on A'4 conditions the most typical confusions regard classical with world and jazz-blues, and metal-punk with rock-pop. These are equally observed both with SVM (Fig. 6(a)) and VQMM (Fig. 6(b)).

Regarding the real-time accuracy of both algorithms under different amounts of data, by looking into Fig. 7 we verify different behaviors between the SVM and VQMM algorithms. Although the SVM enables better offline results, when in the presence of the whole 30 sec instances, the VQMM seems more robust to small amounts of data. This might be justified by the reliance of VQMM on $12xW$ dimensional

feature matrixes while SVM relies on the same amount of data collapsed into a single 24-dimensional vector of means and standard deviations of the analyzed features.

## VII. Robot Dancing with Style Demonstration

In order to demonstrate the applicability of the proposed system in a dancing robot capable of reacting to the genre of a continuous musical stream on-the-fly with style-specific dance motions, we designed a live robot dancing scenario in the same real-world environment considered in the assessment of the system (see Section V).

This robot dancing scenario considered all the following conditions (which replicate A'4):

- Music played from a speaker standing 1 m away from the robot from its back (180°) direction.
- Audio captured from the robot single back (#8) microphone, in frontal line to the music source.
- Use of continuous music stream composed of 10 music clips, selected from the 120 files used in the system assessment, and concatenated without any gaps to reproduce unexpected changes of the musical genre.
- Real-time musical genre recognition using VQMM for fast live adaption to changes in the musical genre. To achieve such real-time processing, we followed Fig. 7 and considered audio chunks of 3 sec processed without overlap, and respond also every 3 sec with the recognized musical genre.
- Six different periodic dance motions, composed of 2 interleaving key-poses as described in Section IV-D, and each one matching their respective musical genre.
- Moving head in all dance motions to interchange the direction of the music source in relation to the back microphone.
- Periodic dance motions performed at random tempi in the interval of [40-180] bpm.

A video with excerpts of this robot dancing demonstration is sent in attachment.

## VIII. Conclusions and Future Work

In this paper, we propose the integration of genre recognition in a dancing robot with embedded microphones to enable it to recognize the genre of a musical piece while moving in a real-world noisy scenario. For this purpose, we assessed and compared two state-of-the-art musical genre recognition algorithms under different real-world noisy environments of increasing complexity. The results demonstrate that an accurate and robust musical genre recognition system demands the use of acoustic models trained in matching noise conditions. Also, the additional use of SSS as a preprocessing is able to improve the algorithms' accuracy by a total average of 43.6 pp under the most harsh conditions, when compared to tests run on single channel using acoustic models trained in clean conditions. In these conditions, when considering SSS, matching acoustic models, and an SVM genre classification algorithm, one could achieve a top mean genre recognition accuracy of 87.3%. Envisioning real-time genre recognition on the same conditions, VQMM was able to perform up to 58% accuracy by solely relying on 3-sec time-windows of music data without overlap.

Ultimately, a demonstration session confirms the applicability of the proposed integration for genre-adaptive dancing robots in real-world noisy environments.

In the future, and akin to [4], we should consider ego noise suppression as a preprocessing to genre recognition under ego noise from the robot dance motion. We should also investigate universal acoustic models for genre classification under multiple noise conditions. Ultimately, we should consider novelty detection strategies in order to enhance the reaction time to changes of the musical genre in continuous music streams, and compare the reaction time and real-time performance of the system under different conditions.

## References


[1] P. Reddish, R. Fischer, and J. Bulbulia, "Let's Dance Together: Synchrony, Shared Intentionality and Cooperation," *PLoS ONE*, vol. 8, no. 8, 2013.

[2] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 133–141, 2006.

[3] G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, and J. Imura, "Whole Body Motion Noise Cancellation of a Robot for Improved Automatic Speech Recognition," *RSJ Advanced Robotics*, vol. 25, no. 3, pp. 360–371, 2011.

[4] J. L. Oliveira, G. Ince, K. Nakamura, K. Nakadai, H. G. Okuno, L. P. Reis, and F. Gouyon, "Live Assessment of Beat Tracking for Robot Audition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, 2012, pp. 992–997.

[5] S. R. Ness, A. Theocharis, and G. Tzanetakis, "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," in *ACM International Conference on Multimedia*, 2009.

[6] T. Langlois and G. Marques, "A music classification method based on timbral features," in *International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 81–86.

[7] ISMIR2004, "Musical Genre Classification Dataset," 2009. [Online]. Available: http://ismir2004.ismir.net/genre_contest/

[8] Y. Song, S. Dixon, and M. Pearce, "A Survey of Music Recommendation Systems and Future Perspectives," in *International Symposium on Computer Music Modelling and Retrieval (CMMR)*, 2012, pp. 1–16.

[9] B. Happi Tietche, O. Romain, B. Denby, L. Benaroya, and S. Viateur, "Fpga-based radio-on-demand broadcast receiver with musical genre identification," in *IEEE International Symposium on Industrial Electronics (ISIE)*, 2012, pp. 1381–1385.

[10] H.-G. Kim and J.-M. Cho, "Car audio equalizer system using music classification and loudness compensation," in *IEEE International Conference on Convergence (ICTC)*, 2011, pp. 553–558.

[11] G. Xia, R. Dannenberg, J. Tay, and M. Veloso, "Autonomous Robot Dancing Driven by Beats and Emotions of Music," in *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Valencia, Spain, 2012, pp. 205–212.

[12] D. K. Grunberg, A. M. Batula, E. M. Schmidt, and Y. E. Kim, "Affective gesturing with music mood recognition," in *International Conference on Humanoid Robotics (Humanoids)*, Osaka, Japan, 2012, pp. 343–348.

[13] M. Janvier, X. Alameda-Pineda, L. Girin, and P. Horaud, Radu, "Sound-Event Recognition with a Companion Humanoid," in *IEEE International Conference on Humanoid Robotics (Humanoids)*, Osaka, Japan, 2012.

[14] K. Nakadai, H. G. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2003, pp. 398–405.

[15] A. D. L. Torre, J. C. Segura, C. Benitez, J. Ramirez, L. Garcia, and A. J. Rubio, *Speech Recognition Under Noise Conditions: Compensation Methods*, 2007, pp. 439–460.

[16] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind Source Separation with Parameter-Free Adaptive Step-Size Method for Robot Audition," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 6, pp. 1476–1484, 2010.