Minstrel robots: Body language expression through applause evaluation

F. Kraemer¹, I. Rodriguez², O. Parra², T. Ruiz³, E. Lazkano²

Abstract—Currently humanoid robots become technically more capable of executing complex movements, showing human-like gestures, sometimes even facial expressions, and acting in general. While this lays the basis to make robot theater enactments more and more interesting for the audience, another key-component is flexibility in the flow of an event to move on from simple pre-scripting. Here a sophisticated method is introduced relying on audio processing, clustering and machine learning techniques to evaluate audience’s applause, allowing the robot to infer self-evaluation about its actions. In a second step we use this information and a humanoid robot’s body language to alter the flow of the event and display a reaction for the audience.

I. INTRODUCTION

Basque, euskara, is the language of the inhabitants of the Basque Country. And bertsolaritzia, Basque improvised sung poetry, is one of the manifestations of traditional Basque culture that is still very much alive. Bertso-saior, Basque poetry contests in which verse-makers compete, typically feature a number of poets (bertsolari-s) that first await a set of words (i.e. rhymes), which then should be incorporated in a spontaneously made up poem. The poems are presented to the audience and may make use of one out of various melodies. Bertsolaritzia offers another sphere to develop robot body language and robot communication capabilities, and thus, to increase robot autonomy and sociability. The Bertsobot project aims to develop troubadour robots, and allows for a robot-style enactment. In its current version it features basic poetry creation and follows the different phases of an event, from finding a microphone to presenting a poem. Body language is used to make the robot’s actions more lively. A more detailed description of the state of the Bertsobot system can be found in [13].

The events usually consist of a rather formal flow of poetry recitations, reasonable appreciation from the audience, mostly by clapping and calming down to silence for the next poet’s turn. This situation has been modelled well in Bertsobot, but so far the robot had no possibility to show empathy to the audience. If the troubadour performance is to be perceived credible, lively and creative, public reaction must be perceived by the robot somehow and its behaviour must reflect the noticed sensations, either showing a proper body language or, like real troubadours do, integrating them in the next sung verse.

Despite the thrilled state of the audience, the need for concentration of human poets is very much respected by them, and thus, the crowd waits until the actor finishes to show how pleasant the verses have been, usually clapping as well as laughing when they have found it amusing.

There are several poetry disciplines around the world similar to bertsolaritzia, the Italian bards, Argentine payadors or Catalan glossators to mention some. But the closest example is the American poetry slam [14] in which poets read or recite poems and are usually judged by selected members of the audience or by a panel of judges. The winner is chosen according to the intensity and duration of the audience’s applause.

The goal of this work is to move on to a closed-loop form of the robot performance where the robot perceives audience’s feedback measuring the clapping intensity for it has sung, and then reacting through subtle gestures like it would be expected from a human poet.

II. RELATED WORK

Developing an approach to react to the audience’s feedback covers multiple fields, such as applause detection, classification and selection of the robot’s appropriate reaction in the context of the performance.

The problem of content-based audio classification and segmentation has been studied intensively outside the field of robotics and some work has specifically focused on applause. Cai et al. [5] have successfully used Mel-Frequency Cepstral Coefficients (MFCC) and a set of low-level features such as sub-band energies to find significant audience reactions including applause and laughter.

Few work has been done when it comes to observing robot induced audience expressions. Knight et al. [9] have developed a stand-up comedian robot that varies joke selection depending on pre-communicated visual feedback and noise level. Another performance robot by Katevas et al. [8] similarly features joke-telling. It incorporates visual emotion recognition and detecting the noise levels to delay the performing of the comedy script. Audience feedback is partly elicited by the robot itself leaving the spectators in a natural comedy setup without human interference.

Several authors have observed the effect of machines on humans. Nass et al. [12] found that adult humans do not credit anthropomorphic characteristics to computers, although people would still accept questions directly pointing at this and interpret machines in a humanised way. How
humanoid robots’ actions can be designed in order to produce well understandable body language and social cues has been investigated in [2], [6] and [11].

III. PROPOSED APPROACH

The presented work can generally be split up into a straight-forward workflow. In the initial step, audio processing and machine learning techniques prepare the input audio stream by first chunking it, and then classifying each chunk as being applause or not. Next, the incoming stream of classified chunks is segmented into sections of consecutive applauses, leading to a small descriptor for every evaluated applause. Based on all previous applauses of the event, the most recent one can subsequently be classified, and therefore, a corresponding robot gesture is selected. Fig. 1 summarizes those steps.

![Audio Chunk Classification Workflow](image)

Fig. 1: Approach workflow

Audio chunk classification (section IV) and applause segmentation (section V) are described and evaluated separately. Then, section VI describes how the applause detection is integrated into a live event and the robot behaviour is adjusted to react to the received applauses. It also features elaborated evaluation of the fully developed system in a real event. Finally, we conclude discussing the results and pinpointing future improvements.

IV. AUDIO CHUNK CLASSIFICATION

Applause detection can be described as a binary classification problem based on the live audio received from the audience. First, a preprocessing step chunks the audio stream into overlapping slices of about 0.1 seconds length using a Hamming-window. Next, the slice is transformed into the frequency domain using a Fast Fourier Transform (FFT) algorithm. And then, the dominant frequency band and Mel Frequency Cepstral Coefficients (MFCC) are extracted using the Essentia [4] library.

Some experiments were performed in WEKA [7] to find a suitable approach for the audio classification problem. Several supervised classification algorithms were trained with a database of 5642 audio entries (1169 labelled as applauses and 4473 as non applauses) from heterogeneous bertso-stadium events: A Support Vector Machine (polynomial kernel, epsilon=1.0E12, complexity 1), Naive Bayes, Bayesian Network (max. 3 parents), 1-Nearest Neighbour and J48 decision tree (conf. factor = 0.25, pruned). 10-fold cross-validation results can be seen in Table I.

<table>
<thead>
<tr>
<th>Performance</th>
<th>SVM</th>
<th>NB</th>
<th>BN</th>
<th>K-NN</th>
<th>J48</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC</td>
<td>0.93</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.95</td>
</tr>
</tbody>
</table>

TABLE I: Audio classification comparison

The results show no significant differences among the tested classifiers. K-NN stands out a bit but it is not very well suited for real-time problems. Alternatively, decision trees (J48) are easy to implement and computationally balanced [3]. It must be taken into account that around 20 chunks need to be classified in a second, thus, the classifier needs to give an answer in less than 50 ms. Hence, the J48 decision tree was selected as the final audio classifier. The acoustic energy calculated according to equation 1 and a binary value showing the belonging to the applause class or not for every audio chunk, will be the input for the next step.

\[ E = \sum_{0}^{T} x(t)^2 \]

\( T \) stands for the chunk duration and \( x(t) \) represents the signal value.

Fig. 2 illustrates the audio chunk classification process.

![Audio Chunk Classification Diagram](image)

Fig. 2: Audio chunk classification

A. Evaluation

The J48 based audio classifier was tested with the following three different collections of audio recordings:

- Cheering and singing from a stadium (labelled as Stadium).
- A combination with equally numbered samples from unheard human bertso-saio events, event cheers, music and whistle cheers (labelled as Pos+Neg).
- A combination of applauses from unheard human bertso-saio events superposed with stadium noise (labelled as BertsoStadium).

Table II summarizes the results obtained for these test cases.

<table>
<thead>
<tr>
<th></th>
<th>Stadium</th>
<th>Pos+Neg</th>
<th>BertsoStadium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>NO</td>
<td>NO   YES</td>
<td>NO</td>
</tr>
<tr>
<td>Precision</td>
<td>1.0</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>Recall</td>
<td>0.63</td>
<td>0.71</td>
<td>0.85</td>
</tr>
<tr>
<td>F-Meas.</td>
<td>0.77</td>
<td>0.77</td>
<td>0.80</td>
</tr>
</tbody>
</table>

TABLE II: Audio classifier results for test cases

The cheering sounds from the Stadium dataset do not contain any applause and are classified with a sufficiently high recall rate. This can be seen as a hard test case since these sounds are unlikely to be heard in a bertso-saio event. The Pos+Neg dataset shows the quality of the classifier in the event context. The good precision rate proves that it is highly applicable to the problem even for unseen data. Finally, the
superposed audio in BertsoStadium shows the limit of the classifier with poor NO-class detection for highly noisy data.

V. APPLAUSE SEGMENTATION

Once the classification of the chunks has been computed, the next step is to segment the stream of classified chunks and find the portions of applause.

Over the stream of positively or negatively classified audio chunks a sliding window is applied. If the number of positive applause classifications exceeds a certain threshold, the first positive chunk’s start time marks the start of an applause. In case the previous segment was also an applause, a continuation of the applause has been detected. The applause is identified to go on until the percentage of positive applause chunks falls below the threshold. Then, the last positive chunk of the segment marks the end time of the applause. In few occasions applauses can nearly die away and flare up again on the initiative of few individuals. In those cases segments of applauses with little temporal distance (e.g. smaller than 0.5 s) are merged into one. Finally, the energy of each segment is accumulated leading to a 2-dimensional descriptor consisting of an applause’s duration and acoustic energy.

At this point more complex descriptors would be viable, e.g. incorporating more information about applause dynamics like the time-energy relation or dominant frequency bands. Generally, the basic descriptor proved to be sufficient as we are dealing with a rather homogeneous type of applauses. It can then be used to imitate the evaluation strategy commonly used in poetry slams, allowing to judge the performance with an “applause norm”.

A. Evaluation

To judge the applause segmentation implementation, \( N = 20 \) applauses were taken from bertso-saio event videos and the detected start and end times were compared to the manually distinguished ones. That is, we considered the observed differences between the true and the program-detected times for applause starting (\( D_{S,j} \)) as well as for applause ending (\( D_{E,j} \)). For each type of differences (\( D_{j}, j \in \{S,E\} \)), the mean (\( \overline{D}_j \)), Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) statistics have been calculated as follows:

\[
\overline{D}_j = \frac{1}{N} \sum_{i=1}^{N} D_{j,i}
\]

\[
\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (D_{j,i} - \overline{D}_j)^2
\]

\[
\text{MAD} = \frac{1}{N} \sum_{i=1}^{N} |D_{j,i} - \overline{D}_j|
\]

Table III shows that both means are close to zero, showing almost no bias to the true values. The end detection errors are more spread as can be seen by its higher MSE/MAD. This is due to fading out applauses with no hard end bounds.

As a better intuitive measure also the MAD was calculated, which shows that the mean detection error can be expected to be 0.2 s to 0.3 s for start and end bounds. While these deviation errors can practically add up, they still range low enough for the purposes of this approach, when compared to applause durations usually ranging between 4 s and 10 s.

<table>
<thead>
<tr>
<th></th>
<th>Start Detection Error (s)</th>
<th>End Detection Error (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>MSE</td>
<td>0.04</td>
<td>0.22</td>
</tr>
<tr>
<td>MAD</td>
<td>0.18</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table III: Applause segmentation errors

VI. LIVE EVENT

The goal of this work is to make the robot behave accordingly to the audience’s applauses. This is a rather subjective task, first because the high variety of applauses is perceived differently even by humans. And secondly, appropriate reaction gestures must be defined, which need to be well understood by a broad audience.

Subsections VI-A, VI-B and VI-C describe our way to first introduce an objective classification method, and then, choose and execute a suited reaction. Finally, different experiments were performed in real-time in a live event to evaluate the overall robotic system and audience’s acceptance of the robotic performance. In addition, a video of some verses sung by the robot and its reaction to audience’s applauses can be seen at RSAIT’s YouTube channel1.

A. Applause Classification

The applauses are coarsely categorized as belonging to one of the following classes:

- NEGATIVE
- NEUTRAL
- POSITIVE
- VERY_POSITIVE

The NEGATIVE class is because of the social obligation of also applauding even for poor performance. These “applauses due to politeness” would actually imply a negative feedback for the robot. On the other extreme, very extensive applauses, as they occur at least at the end of an event, also call for an extra class, the VERY_POSITIVE one.

The applause segmentation step gave us a 2D descriptor containing the duration and energy of the applause. Using this descriptor, now the applause must be classified as belonging to one of the mentioned classes. In general the classification must be done in an online learning fashion. This holds to a greater extent due to the enormous variety in audience sizes, audience knowledge of the robot system, acoustic perception and emotional state of the audience, making distinct events difficult to be compared. Thus, unsupervised online learning techniques fit better to the given problem.

1RSAIT’s YouTube channel. https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ
We chose to use the k-means [10] algorithm in a non-standard way, as it is one of the simplest algorithms which uses unsupervised learning methods to solve known clustering issues. Instead of fixing the $k$ value beforehand, and training the algorithm with a database, the algorithm starts with an empty database and new data is incorporated to the training set after each applause session. Subsequently, as new applauses are being perceived, the clustering is executed again, and steadily, the number of available classes is adjusted. The first applause will always be evaluated NEUTRAL. We consider that this first applause means a welcome to the actors and that the show needs a warm up before showing emotional feedback. For the second one two classes are allowed, NEUTRAL or POSITIVE, as audience and robot are getting to know each other and this is the most uncertain learning phase. Afterwards, NEGATIVE, NEUTRAL and POSITIVE classes are offered, until after six feedback rounds sufficient knowledge has been accumulated to also make use of VERY_POSITIVE class.

As a preprocessing step to k-means both data dimensions are being normalised first, which might be handled differently depending on the event.

Fig. 3 shows an example of the classification of 41 consecutive applauses analysed from a real bertso-saio event. There is not a clear separation among classes during the initial steps of the algorithm due to the small amount of data the algorithm is fed with. This is reflected in the online classification results (3a). While the event progresses more data is available and the k-means is able to separate clusters belonging to the four classes (3b).

B. Gesture Selection

For every feedback class a set of 3 predefined gestures has been prepared, giving a total amount of 12 different gestures. Each gesture consists of several movements that must be show some fluency. After the classification of a feedback event, one gesture is randomly chosen out of the corresponding set. To avoid obvious unauthentic behaviour the last executed gesture of the selected class is excluded for the next round, so that there will never be repetitions within a short time period.

Examples are shown in Fig. 4. The first row corresponds to a negative feedback gesture in which a sad emotion can be clearly appreciated. The sequence relies on a slightly buckled bearing, a shaking of the head and the eyes fixed on the floor. The second row shows a neutral reaction; resting on the body’s left side and moving a little its right arm the robot shows indifference. The third row belongs to a positive feedback; a happy reaction can be observed in which the robot moves its hand from bottom to top as celebrating its success. The classic bow shown in the last sequence reflects one of the available very positive reactions.

All gestures usually take 3s to 5s of time and are optionally accompanied with short sounds or Basque phrases. These can range from a reserved “OK, thank you” to a cheerful “Thank you”.

Two different behaviours were implemented:

1) The **diffident behaviour** only makes use of the first three classes, leaving out the VERY_POSITIVE reaction.

2) The **more exaggerated behaviour** aims to clarify the robot’s intentions, while it might neglect the typical flow of the event. It makes use of all classes, including the most expressive gestures. Additionally, the robot puts its hand close to its ear claiming for more as soon as it perceives applause sounds.

The first one shows a more restrained or cautious character, while the second one could be categorized as haughty.

C. Evaluation

Several experiments were conducted in an event to evaluate the overall robotic system and human acceptance of the robotic stage. The audio classification and applause segmentation could be effectively proven to work well with objective offline test input, due to their subjective characteristics. However, the applause classification and gesture selection required to be evaluated online in a live event.

This event was arranged similar to a human bertso-saio, with the bertsolari robot in front of a seated audience. 17 participants (59% female) familiar with this type of events, took part in listening to the verses and reacting with applauses. After each of the robot’s counter-reactions, all participants, lecturers and researchers aged between 25-55, answered a set of questions (see Fig. 5).
The question set was designed carefully to avoid suggestive questions and to get the most honest opinions grading from 1 to 7, 7 being the most positive.

Altogether, 12 bertso-s were presented, which were selected from different sources. Four verses were automatically generated by the “automatic bertso composer” system [13], which in its current alpha state produces technically permissible but sometimes meaningless poems. Another four were created by non-professional bertsolari-s. And the rest were verses composed and sung by professional bertsolari-s on national contests. The set of verses was split up into two subsets of 6 and presented to the audience. For the first one only the more diffident behaviour was enacted, while during the second subset the more exaggerated behaviour was used.

Analysing the questionnaires we could infer a lot about the audience’s acceptance of the robotic system during the show. Fig. 6 compares the averages of the different perceptions about the event: how the individuals rated each verse, how they rated the group response and how suitable the gesture shown by the robot was (Part 1). Text labels correspond to the class of gesture made by the robot.

POSITIVE and VERY_POSITIVE reactions are accepted exceptionally well by the audience. The exception is verse number 6, which was perceived better by the audience, but only classified NEUTRAL. This is because the verse it refers to was enacted at the sixth position, when only three classes were allowed. The only verse (number 11) which was followed by a NEGATIVE reaction, but evaluated quite well by the audience can be explained by problems with the system platform, which led to bad detection of this applause in the first place.

After each subset of verses, some concluding questions about the robot’s behaviour were asked (Part 2). The comparison between the two different behaviours (Table IV) led to the following remarks: While both negative and positive
audience feedback were said to be well understood by the robot, the positive feedback was increased from 5.2 to 6.1 out of 7 possible points with the exaggerated behaviour. The second behaviour also increased the general expressiveness score from 4.8 to 5.9 and the variety improved almost 0.5 to 4.9 points.

When asked about the closeness to a real bertso-saio event, the audience rated the first set of NAO’s reactions higher (3.9 vs. 3.3), but both scores are below-average.

<table>
<thead>
<tr>
<th>Question</th>
<th>Diffident beh.</th>
<th>Exaggerated beh.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive feedback</td>
<td>5.19 ± 0.83</td>
<td>6.13 ± 0.72</td>
</tr>
<tr>
<td>Negative feedback</td>
<td>5.31 ± 1.40</td>
<td>5.14 ± 1.03</td>
</tr>
<tr>
<td>Fluency</td>
<td>4.75 ± 1.13</td>
<td>4.71 ± 0.99</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>4.75 ± 1.29</td>
<td>5.87 ± 0.92</td>
</tr>
<tr>
<td>Variety</td>
<td>4.44 ± 1.03</td>
<td>4.88 ± 1.45</td>
</tr>
<tr>
<td>Closeness</td>
<td>3.94 ± 1.12</td>
<td>3.31 ± 1.45</td>
</tr>
</tbody>
</table>

TABLE IV: Public response average values for the Part2 of the questionnaire

In order to get more insight about the quality of the gesture set, the audience was confronted to each gesture in a random sequence while asked for a gesture tag. Fig. 7 shows the measured classification rates. Only case 5 was wrongly classified as being NEUTRAL for the big majority of the audience, while the gesture was POSITIVE. Cases 2, 3, 6, 8 and 12 do not show a clear definition, as the results show a tie between the right tag and a neighbour class.

![Gesture Classification Rates](image)

Fig. 7: Audience gesture classification rates for the 12 gestures available. The execution order was randomly selected.

VII. CONCLUSION AND OUTLOOK

Without questioning for additional information or unusual behaviour of the audience, this work already allows robots to react and alter their behaviour during an event according to a specific audience’s natural feedback. We have found that the audience felt better understood when the robot exaggerated its behaviour. It is inconceivable for a bertsolari to show any kind of arrogance in such a traditional cultural event that requires extreme concentration. As a consequence, this result makes us think that we should detach the robot event from its human counterpart. Moreover, it appears that instead of imitating the original event in detail, a new identity should be established for the robot.

The classification of the applauses could also be achieved by a Gaussian Mixture Model allowing for a more sophisticated classification using priors. A comparison with the already implemented system should be carried out. Another extension may be to deduce the number of needed gesture classes from an error measure over the current classification. Future work will also include investigation about audience-sensitive planning and integration of further, but more subtle human feedback, like emotions in general through facial expressions or reservation through delayed reactions. Also we will research more in the field of gesture evaluation of movement styles combined with probabilistic methods. The robot’s empathetic level would be improved by applying sentiment analysis to the verses sung by the opponents and combining the results with the proposed gesture selection mechanism.

REFERENCES