Audience mood estimation for the Waseda Anthropomorphic Saxophonist 5 (WAS-5) using cloud cognitive services

Estelle Randria, Sarah Cosentino, Jia-Yeu Lin, Thomas Pellegrini, Salvatore Sessa, Atsuo Takanishi

Audience mood estimation for the Waseda Anthropomorphic Saxophonist 5 (WAS-5) using cloud cognitive services

Estelle I. S. Randria (Université Paul Sabatier), Sarah Cosentino (Waseda University), Jia-Yeu Lin (Waseda University), Thomas Pellegrini (Université de Toulouse), *Salvatore Sessa (Waseda University), Atsuo Takanishi (Waseda University)

1. Introduction

Human social communication is complex, and comprehends two communication channels: rational and emotional. The first one is conscious and result of cognitive, rational processes. The second one is often subconscious and the result of emotional instinctive processes.

Social communication is also adaptive, its evolution depending on both the rational and the emotional responses of the counterpart [1]. In particular, rational communication uses some specific type of language, verbal or non-verbal, while emotional communication largely relies on subconscious body language cues and facial expressions [1]–[3]. Music is a form of social communication [4]: a musical performance evolves in what we can call the "Music social space", the environment in which the musician performs. In the music social space, the musician must communicate with, and adapt to the signals coming from, the others around him. Musical performances are complicated communication paradigm: as performers usually cannot speak during musical performances, they must be able to understand gestural cues from other performers for synchronization. Moreover, they should be able to recognize and adapt to feedback from the audience. In fact, musicians use gaze, specific gestures, and facial expressions to communicate information about music arrangement and timing to other partner performers. In addition, they often accompany instrument playing with empathetic body language and facial expressions to convey and enforce emotions evoked by music to the audience [5].

Musical robots have been developed with the objective of overcoming human physical limitations in instrument playing, so to achieve excellent skills in the specific played instrument [6]. Nowadays, there are also several examples of interactive musical robots [7]–[9]. However, a musician robot completely integrated in the music social space must be able not only to send and recognize synchronization gestures from partner performers, but also to acknowledge emotional signals from the audience, and to change its performance in an adaptive way.

The Waseda Anthropomorphic Saxophonist robot Ver.5 (WAS-5) shown in Fig. 1 has been designed to mimic the functions of the organs involved in the instrument playing (oral cavity, lips, and fingers). Its dimensions are 30% larger than an average human [10]. WAS-5 can play the saxophone at the same level of an intermediate human player. However, if we want WAS-5 to perform at the same level as a human performer, it has to be perfectly integrated in the music social space. This means that the robot must be able to execute and interpret gaze, body movements, and facial expressions to synchronize with partner performers and communicate emotionally with the audience. The current version of the robot can communicate with basic body language and facial expressions (i.e. bending, frowning, and winking) to send synchronization signals to partner musicians and accompany music with emotional cues [11]. Moreover, it can analyze other performers' movement data from cameras and wearable inertial measurement units, and adapt its performance accordingly [12], [13].

The objective of this research was to verify the feasibility to recognize audience emotions in order to generate emotional musical behaviors for the saxophonist robot.

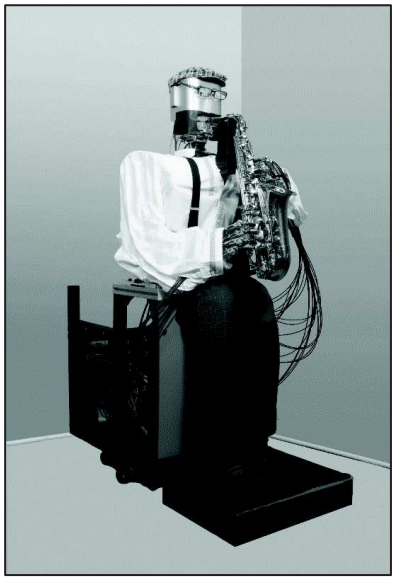

Fig. 1 Waseda Anthropomorphic Saxophonist 5 (WAS-5)

2. Face Expression and emotion detection

The objective is to create the system which is represented in Fig. 2, WAS-5 acquires images of audience from cameras and the cloud services are used to identify faces and ambience. Depending on the ambience a proper music (MIDI file) is selected and played by the robot.

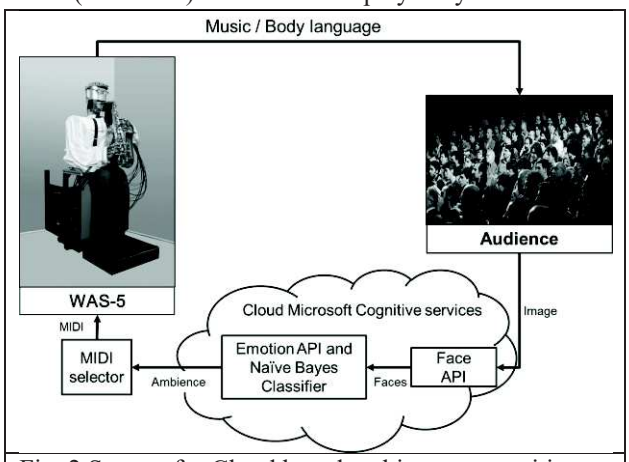

Fig. 2 System for Cloud based ambience recognition

In this section, the definition of emotions on which our research is based, as well as the chosen emotion recognition method are presented, explained, and validated.

2.1 Emotion definition

There is still no generally accepted definition of emotion [14]. However, it is accepted by now that emotions are states of interrelated, synchronized physiologic and behavioral changes in response to an external or internal stimulus [15]–[17]. The external manifestation of emotion is called affect; a pervasive and sustained emotional state, mood [17]. Psychologists have extensively studied emotions, and several different models have been devised. In particular, these models try to define which emotions can be considered basic, as unique, distinct psychophysiological states, and which ones can instead be considered a mix of two or more basic emotions. Determining basic emotions is a complex task, as deciding the criteria for defining an emotion as basic are difficult to settle [18], [19].

Plutchik considered eight basic emotions: acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise. He elaborated the well-known wheel of emotions [20]. Ekman first considered six basic emotions, the "Big Six": happiness, sadness, fear, surprise, anger, and disgust. He then expanded his model to 15 by adding contempt (often considered as the seventh basic emotion), amusement, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame [21]. Weiner and Graham considered as basic emotions only sadness and happiness [22]. James considered as basic emotions fear, grief, love and rage [23].

2.2 Emotions and facial expressions

As emotions are linked to specific psychophysiological states, there are different methods to monitor and measure physiological parameters to estimate the linked emotion. However, the most ecological methods –thus most attractive– rely on audio and visual information. In particular, facial expressions are the most direct and clear display of emotions [28]. Following this approach, Ekman and Friesen developed the Facial Action Coding System (FACS) [16], a system to taxonomize human facial movements, and linked each emotion in Ekman's model to a specific set of facial movements. From then on, most of the successful works on automatic emotion recognition uses facial expressions and are based on Ekman's basic or extended emotions models.

2.3 Face expression recognition system

As WAS-5 is equipped with two front cameras, we also decided to analyze facial expressions emotion recognition from facial expressions, based on Ekman's basic emotions model. Moreover, as we plan to use emotional recognition as a feedback from the robot audience, we expect to have to analyze simultaneously facial features of several people, and this would involve high computational load. For this reason, we resolved to use cloud computing to be able to scale easily computational power and expressions database. We chose Microsoft Azure as this cloud computing service already contains two APIs for face detection and emotion detection: Microsoft Cognitive Services - *Face API* and *Emotion API* [29]. Face API analyzes an image and returns the number and the position of the faces in the image. Emotion API analyzes a face image and returns coefficients in the range [0-1] for all the seven basic emotions and neutral of the Ekman's model. The more a coefficient is near to one, the stronger is the corresponding emotional component in the facial expression. The displayed emotion can be estimated using these coefficients, and in the most simple of the cases, the estimated emotion equals the emotion with the highest coefficient.

3. Experiment and Results

3.1 Face API test

The first step to examine the feasibility of our system is to test the performances of the system in tracking and analyzing simultaneously the emotions of many people. We tested separately the performances of the Face and Emotion APIs. In particular, the performances in face detection of Face API are strongly influenced by the

orientation of the face compared to camera. Orientation of an object in front of the camera is expressed by tilt (angle around the horizontal axis) roll (angle around the frontal axis) and pan (angle around the vertical axis).

As the audience might move during the robot performance, pan might vary accordingly, and being the cameras fixed on the robot, tilt will result when the height of camera and the person face are mismatched. To test performances depending on tilt and pan variations we used an existing facial recognition database, the Head Pose Image Database [35]. This database contains 2790 neutral expression face images with a variation of tilt and pan angles between -90° and +90°. In total, the database has 15 folders, each folder corresponding to one subject. Each subject took two series of 93 pictures with different head poses. We created a dictionary of tilt and pan combinations associated to the occurrence of face detection for each specific combination. Results showed that face detection is over 90% within the range [-30°, +30°] of both tilt and pan (Fig. 3). From the results, we can also observe that tilt is more critical for recognition than pan, so the camera should have adjustable height for a more reliable recognition. We also found that accessories that cover part of the face like hat, sunglasses, or scarfs, strongly influence face detection. In particular, accessories that cover the lower part of the face are critical and do not allow the face detection with this API.

3.2 Emotion API test

We tested the performances of the Emotion API. We combined the HeadPoseImage database with another two existing databases for specific emotional face expression recognition: the Cohn-Kanade database [36], and the Cohn-Kanade Extended database [37].

We tested the system with a subset of these combined databases, containing seven folders, each one corresponding to a basic emotion in Ekman's model: happiness, sadness, fear, surprise, anger, disgust, and contempt. The subset contained in total 210 mixed, black and white, and colored pictures, with no tilt and pan. In these conditions, success rate of emotion recognition is 68%.

To improve the emotion recognition, we introduced two correction factors.

1) Naïve Bayes Classifier

We divided the database subset in training and test subsets and we tested the results of the classifier. The classifier uses as input data the scores provided by the Emotion API.

To see if the use of a Naïve Bayes classifier helps to obtain better recognition results we divided the same subset of 210 images from the databases in training data and validation data, and we observed the evolution of the success rate depending on the size of the training data.

The success rate does not increase linearly with the size of the training set. That may reflects the presence of a bias in the training or validation data set, due to differences in image format from the two databases. However, in real-time images are taken in a variety of situations, so variety in the training data is desirable. The best result is obtained when the training set is the biggest (90% of 210 pictures are used), and the success rate is higher than without the classifier: 76.2% against 68% without the classifier. The use of the Naïve Bayes Classifier seems to bring an improvement on the emotion recognition process in the case where the training set is big. While testing the Emotion API, we noticed that Neutral, Happiness and Surprise had a recognition rate of 100%. The API returned coefficients of above 0.9 for Neutral and Surprise and above 0.95 for Happiness. We decided then to apply thresholding before the Naïve Bayes Classifier, to discriminate directly these cases and determining directly these three emotions bypassing the classifier. We observed that using a threshold brings a success rate above 98%.

4. Discussion, conclusion and future works

During improvisation, a musician can adapt his performance depending on the feedback from the audience, according to established criteria. The artist can decide to play to enhance the emotion perceived, by giving the same emotional connotation to the played melody, or try to influence the perceived emotion by giving to the melody the desired emotional connotation.

There are two main aspects to implement emotional interactive improvisation in our robot:

- General audience mood: its evaluation is complex when different emotions are detected on many people

PAN / TILT	-90	-75	-60	-45	-30	-15	0	15	30	45	60	75	90
90	x	x	x	x	x	x	3	x	x	x	x	x	x
60	3	3	13	53	53	80	83	73	67	33	10	3	0
30	10	43	77	93	100	100	100	100	93	83	70	37	17
15	23	53	90	100	100	100	100	100	100	100	77	53	30
0	33	60	93	100	100	100	100	100	100	100	70	50	17
-15	17	40	87	100	97	100	100	100	100	90	60	23	7
-30	7	20	83	93	97	93	100	97	93	80	40	17	3
-60	0	7	27	37	47	63	73	77	53	47	17	0	3
-90	x	x	x	x	x	x	20	x	x	x	x	x	x

Fig. 3 Recognition rate for tilt and pan combinations

- Music feedback: decide the emotional feedback criteria, enhancing or influencing the audience mood

In this paper we mainly focus on the identification of the general audience mood. In particular, we evaluated of a cloud based system for the face and emotion recognition. For real-time multiple faces detections the main limitation is to track faces with tilt and pan among a certain range. This condition can be fulfilled under specific conditions: people in the audience should be facing the robot, and the camera should be placed in a position allowing filming the most people possible regardless of their height.

The combination of Emotion API, thresholding and Naïve Bayes classifier in an algorithm to evaluate the general mood of the audience provided satisfactory results, yet a much bigger and diverse training set should be used for a better estimation of the emotions with negative valence, in different environmental conditions. However, in the case of many different displayed emotions, with a large audience, a more conservative approach is to rely only on valence recognition.

Acknowledges

Research supported by ST Microelectronics K.K. and Microsoft. The authors would like to express their thanks to the Italian Ministry of Foreign Affairs, General Directorate for Cultural Promotion and Cooperation for its support to RoboCasa; Tokyo Womens Medical University/Waseda University Joint Institution for Advanced Biomedical Sciences (TWIns) and International Center for Science and Engineering Programs (ICSEP) of Waseda University.

REFERENCES

[1] J. M. Darley and R. H. Fazio, “Expectancy confirmation processes arising in the social interaction sequence,” *Am. Psychol.*, vol. 35, no. 10, pp. 867–881, 1980.

[2] M. J. Snyder, “Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in...,” *Annu. Rev. Psychol.*, vol. 50, no. 1, p. 273, Feb. 1999.

[3] R. R. Provine, *Curious Behavior: Yawning, Laughing, Hiccupping, and Beyond*. Harvard University Press, 2012.

[4] M. Chanan, *Musica Practica: The Social Practice of Western Music from Gregorian Chant to Postmodernism*. Verso, 1994.

[5] S. Kawase, T. Nakamura, M. R. Draguna, K. Katahira, S. Yasuda, and H. Shoda, “Communication channels performers and listeners use: a survey study,” in *Proceedings of ICoMCS December*, 2007, p. 76.

[6] G. Weinberg and S. Driscoll, “Toward Robotic Musicianship,” *Comput. Music J.*, vol. 30, no. 4, pp. 28–45, Dec. 2006.

[7] G. Weinberg, S. Driscoll, and M. Parry, “Musical interactions with a perceptual robotic percussionist,” in *IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005*, 2005, pp. 456–461.

[8] G. Weinberg, B. Blosser, T. Mallikarjuna, and A. Raman, “The creation of a multi-human, multi-robot interactive jam session,” in *Proceedings of the Ninth International Conference on New Interfaces for Musical Expression*, 2009, pp. 70–73.

[9] M. Cicconet, M. Bretan, and G. Weinberg, “Visual cues-based anticipation for percussionist-robot interaction,” in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 117–118.

[10] G. Xia *et al.*, “Expressive Humanoid Robot For Automatic Accompaniment,” in *13th Sound and Music Computing Conference and Summer School (SMC 2016)*, 2016, pp. 506–511.

[11] K. Matsuki, K. Yoshida, S. Sessa, S. Cosentino, K. Kamiyama, and A. Takanishi, “Facial Expression Design for the Saxophone Player Robot WAS-4,” in *ROMANSY 21-Robot Design, Dynamics and Control*, Springer, 2016, pp. 259–266.

[12] Sarah Cosentino, “Non-verbal interaction and amusement feedback capabilities for entertainment robots,” Waseda University, Tokyo, Japan, 2015.

[13] C.-H. Hjortsjö, *Man’s face and mimic language*. Studen litteratur, 1969.

[14] P. Ekman and W. Friesen, *Facial Action Coding System*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[15] “Microsoft Cognitive Services.” [Online]. Available: https://www.microsoft.com/cognitive-services/en-us/. [Accessed: 22-Feb-2017].

[16] N. Gourier, D. Hall, and J. L. Crowley, “Estimating face orientation from robust detection of salient facial structures,” in *FG Net Workshop on Visual Observation of Deictic Gestures*, 2004, vol. 6.

[17] T. Kanade, J. F. Cohn, and T. Yingli, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000, pp. 46–53.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94–101.