Dancing the Night Away — Controlling a Virtual Karaoke Dancer by Multimodal Expressive Cues

(Short Paper)

Matthias Rehm, Thurid Vogt, Michael Wissner, Nikolaus Bee
Multimedia Concepts and Applications
Faculty of Applied Informatics
University of Augsburg
Eichleitnerstr. 30
86159 Augsburg, Germany
NAME@informatik.uni-augsburg.de

ABSTRACT

In this article, we propose an approach of nonverbal interaction with virtual agents to control agents' behavioral expressivity by extracting and combining acoustic and gestural features. The goal for this approach is twofold, (i) expressing individual features like situated arousal and personal style and (ii) transmitting this information in an immersive 3D environment by suitable means.

Categories and Subject Descriptors

I.5.5 [**Pattern Recognition**]: Implementation—*interactive systems*

General Terms

Human factors

Keywords

multimodal interaction, expressivity recognition

1. INTRODUCTION

Immersive 3D environments like Second Life have become important for the social life of many people. Although it is possible to express certain features of personal style by individual designs of one's avatar, these environments still lack ways of expressing individual or situated behavior. In this article, we propose an approach to control agents' behavioral expressivity by extracting and combining acoustic and gestural features. The goal for this approach is twofold: (i) expressing individual features like situated arousal and personal style, and (ii) transmitting this information in an immersive 3D environment by suitable means. The first goal is achieved by recognizing individual interaction parameters from speech and gestural data. The second goal is achieved by controlling the dance movements of a virtual character and thus making the information available to onlookers. The chosen scenario is that of a karaoke show. Expressive acoustic and gestural cues naturally occur in such a karaoke performance. Thus, this scenario allows for a natural setting to test our approach. Moreover, movements in general and dance movements in particular are ideal candidates to express individual style and/or emotional content (e.g. [3], [4]). Recognizing expressive information from speech and gestural data lifts the burden from the user to consciously select the necessary information e.g. from a drop-down menu of available animations. Instead, his natural behavior is utilized to control the movements of the virtual agent by selecting appropriate animations and modifying these animations according to the recognized expressivity.

Cite as: Dancing the Night Away: Controlling a Virtual Karaoke Dancer by Multimodal Expressive Cues (Short Paper), M. Rehm, T. Vogt, M. Wissner, N. Bee, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons(eds.),May,12-16.,2008,Estoril,Portugal,pp.1249-1252.

2. RELATED WORK

Movement patterns are often seen as a source of information about the emotional state of the user (e.g. [3], [4]). Gallaher focuses instead on expressing personality traits aka personal style by movement features like speed or spatial extent [5]. Pelachaud has shown that users are able to detect these cues in virtual agents [9] and Bevacqua et al. [1] analyse user's expressive movements from video streams to mirror these movements with a virtual character. In previous work [11], we used the Wiimote controller to extract information about the user's cultural background based on his expressive gestural behavior. Price, Douther, and Jack [10] argue that current immersive 3D environments have some significant limitations due to a lack of atmosphere and feeling and propose choreographic techniques as adequate means to overcome these limitations because they claim that there exist accepted movement patterns which can be utilized for this end. A similar argument can be found from James et al. [6]. They propose a method to analyse certain movement features concerned with spatial and temporal relationships between different individuals to control an interactive performance. Reidsma et al. [12] describe a system that extracts beat information from diverse sources like video or audio to modify stored movements of an agent. Although they target their work to entertainment only, they present a first system that combines information from different sources to influence the movement patterns of a virtual agent. Whereas movement patterns are used to extract information about personal style, analysis of expressive features of the speech signal most of the time deals with the recognition of user's emotional or at least arousal state (e.g. [8]). Emotions are often triggered by outward stimuli and thus can be seen as a

Figure 1: Processing steps for expressivity recognition

source of information about the situational context the user is involved in. Until now, only few approaches have dealt with online recognition of emotional states (e.g. [7]). In the next section, we show how this information is extracted from the speech signal.

3. RECOGNITION OF ACOUSTIC EXPRESSIVITY

For acoustic expressivity detection, we follow a strategy that we recently used for emotion recognition from speech [14]. It involves the three steps of signal segmentation, acoustic feature extraction, and classification (see Figure 1). A special demand here is that expressivity recognition has to be done in real-time, which has not been done often in previous work. First, the continuously incoming acoustic signal from the microphone needs to be broken down into meaningful units of analysis. Here, we estimate pauses in the signal by a voice activity detection algorithm and thus analyse only the voiced segments. Acoustic expressivity is detected from a variety of acoustic features. First, pitch, energy, MFCCs, the frequency spectrum, duration, pauses and voice quality are computed. From these basic time series, further series are derived of e.g. all values, only local maxima, only local minima etc. Lastly, statistical functions such as mean, max, minimum, standard deviation are applied to the whole series. The resulting 1316 values form the feature vector. Obviously, this vector contains a multitude of features, among them possibly redundant or irrelevant ones for the task. However, as the feature extraction needs to be done in real-time and thus fully automatically, we assume the big number of features to be responsible for the high degree of automation not being disadvantageous compared to offline classification of affective states, where often also partly manually extracted features are used and where there are no time constraints. For classification, a Naïve Bayes classifier was used. Although this is a very simple classifier, it has the advantage of being very fast without performing much worse than more sophisticated classifiers such as support vector machines. As Naïve Bayes is a statistical classifier, it needs labeled acoustic training data to be build. The classifier was trained on the Berlin corpus by Burkhardt et al. [2] and has excellent recognition rates (see Table 1 left column) for the two-class problem of distinguishing between low and high activation.

4. RECOGNITION OF GESTURAL EXPRESSIVITY

Bevacqua et al. [1] define gestural expressivity following Gallaher's suggestions [5] by six parameters: (i) *Overall Activation* is the number of gestures in a specific time, (ii) *Spatial Extent* describes how much space a gesture needs, (iii) temporal extent is the *Speed* of movements, (iv) *Fluidity* is the smoothness of movements, (v) *Power* is the strength of a gesture, and (vi) *Repetition* is the amount of repeated parts of a gesture. Thus, expressivity does not define what kind of gesture is done but how a gesture is done. We adopt this approach and analyze the user's gestures by supplying him with the Wiimote controller (see Figure 2 left) which uses accelerometers to sense its movements in 3D space along 3 axis. Figure 2 (middle) gives an impression of a typical signal for the three accelerometers. To analyse this signal, we created WiiGLE (Wii-based Gesture Learning Environment), a classification toolbox for acceleration-based gesture recognition. So far, it features three classification methods – Dynamic Time Warping, Naïve Bayes, and Multilayer Perceptron – as well as two different feature sets for shape- and motion-based features, and allows for recording arbitrary sets of gestures from different users, as well as training and using the different classifiers in combination with the two feature sets.

Three of the above mentioned six parameters for gestural expressivity have already been integrated in the proposed system: power, speed, and spatial extent. To analyse fluidity and repetition of a gesture, it becomes necessary to opt for a more dynamic classification method like HMMs to capture the time dynamics of the gestures. Because expressivity is concerned with how a gesture is realized we did not only treat expressivity recognition as a classification problem but also tried an approach that treats expressivity as a problem of feature calculation (see [11]). The results are given in Table 1 (Calculated) and show that this approach works well for the power parameter, which is more or less identical to the amplitude of the signal. For the other two parameters this approach is not satisfying. The machine learning approach with the gesture classfication toolbox leads to better results (Table 1 motion-, shape-based). From the available classificaton algorithms (Dynamic Time Warping (DTW), Naïve Bayes, and Multilayer Peceptron (MLP)) we chose the Naïve Bayes classifier. DTW basically compares the input signal with stored test patterns and decides for the pattern with the closest distance to the signal. This works well for a small set of gestures but poses problems if not the gesture itself is the target but the way the gesture is realized. The Naïve Bayes approach has the advantage over the MLP of being considerably faster which is crucial for the realtime application.

For the given two-class problem for each expressivity parameter, two feature sets for motion- and shape-based classification are employed. Motion features are directly calculated on the acceleration signal of each channel. Features include the absolute minimum and maximum, the mean and median, amplitude and other basic features. 16 motion features are taken into account, no feature reduction has yet been done. To derive shape features, the acceleration information on the three axis is used to estimate the rough shape of the gesture on a two dimensional plane. Although this information is very inaccurate due to some obvious problems, the results for this feature set are very promising (see Table 1) at least for the expressivity recognition task. Feature calculation on the gesture shape relies on the features described by Rubine [13] and includes features like the length of the gesture or the size of the bounding box. Three classifiers are needed for the task of recognizing gestural expressivity, one for each parameter. These classifiers are trained on the two-class problem of distinguishing between low and high values for the expressivity parameters. The training

	Acoustic Expressivity	*Gestural Expressivity*								
		Calculated			Motion-based			Shape-based		
	Activation	Power	Speed	Sp. Ext.	Power	Speed	Sp. Ext.	Power	Speed	Sp. Ext.
low	95.4%	100%	94.8%	62.4%	99.7%	93%	96.3%	99.7%	94%	96.7%
high	94.0%	100%	67.1%	59.5%	99.7%	93%	96.4%	99.7%	94%	96.7%
all	94.5%	100%	81.0%	61.0%	99.7%	93%	96.3%	99.7%	94%	96.7%

Table 1: Recognition results for acoustic and gestural expressivity.

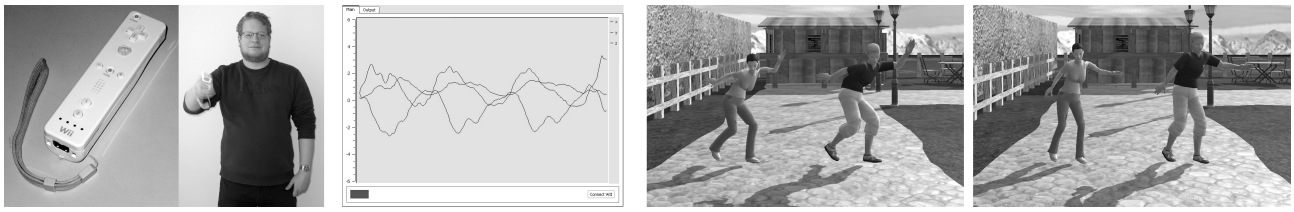

Figure 2: Modifying agents' dancing movements by high and low acoustic and gestural expressivity.

set consisted of 1260 gestures from 7 users, i.e. 420 examples per class. A ten-fold cross-validation of Naïve-Bayes classifiers for motion and shape features was done.[1] The results of this approach are given in Table 1. For the two-class problem, the Naïve-Bayes classifier gives acceptable results regardless if motion-based or shape-based features were employed. For the speed parameter the shape-based classifier is slightly better but the difference is not significant.

5. CONTROLLING A CHARACTER BY MULTIMODAL EXPRESSIVITY

Our prototype realizes a karaoke system with virtual characters that dance during the karaoke performance. Their movements are controlled by the information about situational context and personal expressivity which is derived from the user's singing and his gestural movements. The animations of the agent are taken from a library of 30 different dance movements that are modified to reflect the current context as well as individual style preferences. Animations can be played faster or slower thus realizing the control over the speed parameter of gestural expressivity. To modify the spatial extent of a movement, a technique called motion blending is used. The original animation is modified by combining it with a second one. For instance, an animation with high spatial extent can be blended with a second animation with no movement. The result is an animation identical to the first one but with lower spatial extent. Whereas gestural expressivity can be interpreted as a trait of personal style, acoustic expressivity is treated as situation- and context-dependent. In the given scenario, acoustic expressivity of the user is dependent on the chosen karaoke song which is either a high arousal song like "Livin' La Vida Loca" or a low arousal song like "Lemon Tree". Of course the user can always choose to sing a melancholic song with high arousal. Thus, two different types of information have to be combined to create an individual and context-specific influence on the motion behavior of the virtual agent. Gestural expressivity is used to reflect the personal expressive style of the user in agent's movements. The speed of his movements influences the speed of the dance movements of the agent, the spatial extent of his movements also influences the spatial extent of the agent's movements (see Figure 2 (right) for an example). The additional context information gained from the acoustic signal is utilized to further modify the control parameters of the agent's movements. This modification can be expressed for each parameter g_i (speed, power, spatial extent) by the following formula which determines the control parameters ($\vec{r}$) for a given combination of acoustic and gestural features:

$$\vec{r} = (r(a, g_{Speed}), r(a, g_{Power}), r(a, g_{SpExt}))^T \quad (1)$$

$$r(a, g_i) = \begin{cases} f(g_i) & : \text{ if } a = g_i \\ m(a) * f(g_i) & : \text{ otherwise} \end{cases} \quad (2)$$
$$\text{with} \quad 0 < f(g_i) < 2 \quad \text{and} \quad 0 < m(a) < 2$$

a and g_i are the result of the acoustic and expressivity recognition. Because the animations in the library have different baselines in regard to the expressivity parameters, it is necessary to specify the two function $m(a)$ and $f(g_i)$ for each animation to map the recogntion results to control values for the graphics engine. Consider an animation that is neutral in regard to speed, power and spatial extent. Then, $m(a)$ and $f(g_i)$ are defined as follows:

$$m(a) = \begin{cases} 0.5 & : \text{ if } a \in \text{Class}_{low} \\ 2 & : \text{ if } a \in \text{Class}_{high} \end{cases} \quad (3)$$

$$f(g_i) = \begin{cases} 0.5 & : \text{ if } g_i \in \text{Class}_{low} \\ 2 & : \text{ if } g_i \in \text{Class}_{high} \end{cases} \quad (4)$$

If the animation has for instance a higher baseline for speed, then $f(g_{Speed})$ would result in lower values. This is an adhoc solution tailored to rapdily test the recognition techniques. The next step must be to automatically access the expressive qualities of the animations to apply a model-based approach for the expressive modifications.

Let's assume the user is performing a sad song, i.e., the acoustic expressivity should show low arousal for this song. Nevertheless, the user is generally quite expressive in his gestures and shows high spatial extent. In this case the control parameter for the agent's movements is set to a lower value to take the context information into account. The control values for high and low gestural activity depend on the animation in the library. Very expressive dance movements which already have a high spatial extent as a baseline

[1] using the WEKA analysis tool: http://www.cs.waikato.ac.nz/ml/weka/

have a low control parameter for high gestural expressivity but a high parameter for low gestural expressivity whereas the opposite is true for less expressive original movements. Because the library of dance movements was not created by us but is a commercial product based on motion capture data, the animations do not have a common baseline in terms of expressivity making this item specific modification necessary.

6. EVALUATION

Evaluation of the system is pending at the moment. The evaluation design is centered around two hypotheses: (H1) users experience control over the agent's expressive features, and (H2) the audience interprets the agent's expressive behavior as intended by the user. The two hypotheses correspond to the two goals given in the introduction. If the user shall be able to express situational and individual information through the movement behavior of an agent, the user must also experience control over what the agent is doing which is stated by the first hypothesis. On the other hand, if the user expresses contextual and personal information by the agent's movements, the audience should be able to interpret these movements in the intended way which is stated by the second hypothesis.

To examine hypothesis one, a questionnaire is designed to assess the user's subjective feeling of control over the agent's behavior. For investigating hypothesis two, a web-based experiment will be realized which displays the user's input (acoustic as well as gestural) and two videos exhibiting a dancing agent. One video is the one corresponding to the user's input, the other video is randomly taken from a given corpus. The participant has to decide which video best represents the user's input.

7. CONCLUSION

We presented a first prototype of a system that integrates information from multiple input channels to analyze the user's personal and situational expressive behavior and to utilize this information to control the movements of a virtual character. By this approach, the user is able to let an interlocutor know about his personal style and his current situational context by very natural means, i.e. without consciously thinking about the information. To allow for testing the claims we make for this system, we realized a virtual karaoke dancer to display the user's expressive behavior. This scenario has two advantages: (i) the situational context can be controlled by us (the song specifies the level of arousal that is conveyed by the acoustic input), and (ii) it is a natural situation for expressive movement behavior, thus does not create an awkward situation for the user. The Wiimote is not the ideal input device because the user has to handle it manually all of the time. But it allows for rapidly testing the use of acceleration based recognition. Having shown this, the next step has to be to integrate more unobstrusive sensors.

8. ACKNOWLEDGMENTS

The work work described in this paper is partly funded by the German Research Foundation (DFG) under research grant RE 2619/2-1 (CUBE-G) and by the EU under research grant IST-34800 (CALLAS). WiiGLE is available upon request from the authors.

9. REFERENCES

[1] E. Bevacqua, A. Raouzaiou, C. Peters, G. Caridakis, K. Karpouzis, C. Pelachaud, and M. Mancini. Multimodal sensing, interpretation and copying of movements by a virtual agent. In *PIT*, pages 164–174, 2006.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Proceedings of Interspeech 2005*, 2005.

[3] G. Castellano, R. Bresin, A. Camurri, and G. Volpe. User-Centered Control of Audio and Visual Expressive Feedback by Full-Body Movements. In *Affective Computing and Intelligent Interaction*, pages 501–510. Springer, Berlin, Heidelberg, 2007.

[4] E. Crane and M. Gross. Motion Capture and Emotion: Affect Detection in Whole Body Movement. In *Affective Computing and Intelligent Interaction*, pages 95–101. Springer, Berlin, Heidelberg, 2007.

[5] P. E. Gallaher. Individual Differences in Nonverbal Behavior: Dimensions of Style. *Journal of Personality and Social Psychology*, 63(1):133–145, 1992.

[6] J. James, T. Ingalls, G. Qian, L. Olsen, D. Whiteley, S. Wong, and T. Rikakis. Movement-based interactive dance performance. In *Proceedings of ACM Multimedia*, pages 470–480, New York, NY, USA, 2006. ACM Press.

[7] C. Jones and J. Sutherland. Acoustic emotion recognition for affective computer gaming. In C. Peter and R. Beale, editors, *Affect and emotion in human-computer interaction*, Heidelberg, Berlin, 2007. Springer.

[8] P. Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183, 2003.

[9] C. Pelachaud. Multimodal expressive embodied conversational agents. In *Proceedings of ACM Multimedia*, pages 683–689, 2005.

[10] R. Price, C. Douther, and M. A. Jack. An investigation of the effectiveness of choreography for the portrayal of mood in virtual environments. In *Proceedings of AGENTS '00*, pages 54–55, New York, NY, USA, 2000. ACM Press.

[11] M. Rehm, N. Bee, B. Endrass, M. Wissner, and E. André. Too close for comfort? Adapting to the user's cultural background. In *Workshop on Human-Centered Multimedia, ACM Multimedia*, 2007.

[12] D. Reidsma, A. Nijholt, R. Poppe, R. Rienks, and H. Hondorp. Virtual rap dancer: invitation to dance. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, pages 263–266, New York, NY, USA, 2006. ACM Press.

[13] D. Rubine. Specifying gestures by example. In *SIGGRAPH '91: Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, pages 329–337, New York, NY, USA, 1991. ACM Press.

[14] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia & Expo (ICME)*, Amsterdam, The Netherlands, 2005.