VoicePlay – An Affective Sports Game Operated by Speech Emotion Recognition based on the Component Process Model

Gerhard Hagerer, Florian Eyben, Dagmar Schuller, Klaus R. Scherer, Björn Schuller
audEERING GmbH, Germany
Contact: gh@audeering.com

***Abstract*—In the present work we outline the first computer game which is operated by real-time emotion recognition from speech. Urgency is detected from the voice of a player and conveyed to control parameters like speed during a race or accuracy while shooting on a target for a biathlon sports simulation. Moreover, the game showcases the world's first automatic speech emotion recognition which is based on the Component Process Model, a fundamental theory of cognitive psychology.**

1. Introduction

Computational intelligence has reached quasi-human qualities when estimating emotional characteristics such as arousal from speech [1], [2]. With many real-life systems being capable of deployment outside of laboratory settings [3], emotion recognition tasks are ready to expand into consumer products. In that regard, the entertainment and sports sectors, which play a major role for fun and well-being, can benefit from affective computing.

Therefore, we present a first of its kind show-case which demonstrates how automatic vocal affect recognition can be used as natural and fun input modality for virtual sports gaming. We show a real-time software prototype of such a game running on PC and smartphone platforms. Further, it is the first real-time embedded show-case application of our novel acoustic emotion recognition technology based on a solid psychological model instead of big-data and machine learning. The demonstrator is the first implementation in that regard, and as such shows a fully working example of applied psychological research.

The remainder of this paper is structured as follows: Section 2 briefly describes the underlying VocEmoApI emotion recognition technology, Section 3 describes the gaming prototype, and Section 4 gives evaluation results of a preliminary user study before we conclude our presentation in Section 5.

2. Speech Emotion Recognition

In order to provide a control variable to the game in real-time, voice segments are analysed by our VocEmoApI emotion recognition software, resulting in continuous estimates of the player's affective state. VocEmoApI detects vocal markers which are caused by changes in physiological processes due to appraisal checks in the cognitive affective process [5], [6] – i. e., the cognitive process which happens when we process emotion eliciting events in our mind.

The acoustic voice analysis is based on an extension of the Geneva Minimalistic Acoustic Parameter set [7] implemented in audEERING's openSMILE toolkit [8]. The inference of scores for appraisal criteria is based on empirical correlations between vocal markers and appraisal process criteria. Scherer's component process model describes four major appraisal dimensions [5], [9], [10], [11], [12]: the *Novelty*, the intrinsic *Pleasantness* or goal conduciveness, the ability of the person to *Control* the event, and the resulting *Urgency* for action and behavioural excitation. For our sports gaming context, urgency is the most meaningful appraisal criterion and thus was chosen as only variable to control the game (Section 3). On the acoustic side, urgency, which leads to increased vocal activation, causes – among several other parameters – e.g., the speaking speed, volume, and pitch to change.

3. The Game

We present a computer game in which the operation mode is based on emotional urgency conveyed by the player's voice. The player is engaged in a biathlon sports competition which naturally consists of two parts: skiing and rifle shooting. In the first part (Figure 1, left), a skier avatar needs to be 'cheered' to the target line as fast as possible. In the second part (Figure 1, right), the player has to aim a shot at the centre of a target disc. Affect-wise, this setting is highly interesting, as it requires the player to change between high and low vocal urgency. Likewise, it offers an opportunity to train one's control of vocal emotional display by switching between these different ends of expressed urgency.

Technically, voice activity is detected in a first step by audEERING's noise-robust voice activity detection [13] based on recurrent neural networks with Long Short-Term Memory. Secondly, voice segments are analysed by our VocEmoApI emotion recognition software in real-time (see Section 2), resulting in continuous estimates of the player's emotional state.

For the *ski racing* (Figure 1, left), high urgency in the voice makes the skier moving faster, which is necessary to

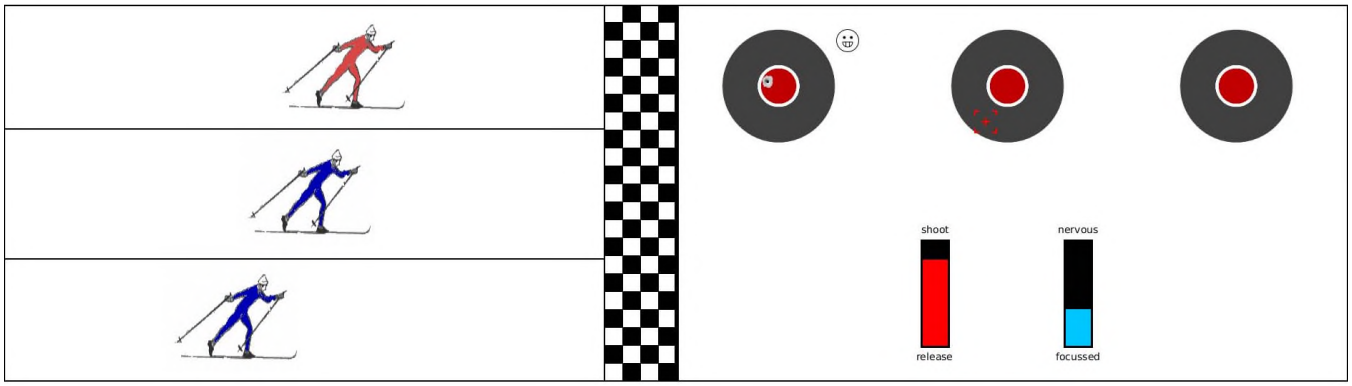

Figure 1. Screenshots of both parts of the vocal affect controlled biathlon game. Left: Ski racing of the player (red) vs two computer-controlled avatars (blue). Right: Target shooting. Urgency is detected from the voice in real-time leading to faster (high urgency) or slower (low urgency) skiing, and to more precise (low urgency) or more off-target (high urgency) shots.

Figure 2. Players playing the affect controlled biathlon game at a public event [4].

to win this part of the game. Maximum points are rewarded for being fastest of the three skiers, and no points for being the last. This part of the game induces strong vocal emotions by creating an urgency for action in the player if the skier moves too slow.

In contrast, the *shooting part* of the game demands the opposite behaviour of the player (Figure 1, right). Calmness and focus is of essence for aiming precisely at a target. Thus, speaking calm causes the cross-hairs to move slower and closer to the centre of the bull's eye. The shot is fired after a fixed duration of speech activity, removing any urgency from the player. The score depends on the distance of the shot from the centre of the target.

The game additionally provides a leaderboard listing the player names sorted by their achieved points. This leads to more competition and thus more engagement in social settings, which in turn encourages players to show even stronger emotions when competing against each others.

4. Evaluation

The game was first showcased at a public demonstration event of audEERING's VocEmoApI technology [4] – see the some of the players in action on the photographs in Figure 2. More than 50 participants participated at the event. Most of them played the game and were fascinated by the way they could influence the game solely by the tone of their voice. 21 players filled in questionnaires with feedback of the game. They were from various nationalities (including English, Italian, German, and Chinese) and aged from approx. 20 to 50 years with a mean around 35 years; approx. 30% were female players, and 70% male. They rated on a scale of 1 (worst) to 5 (best) how well they thought A) the system did pick up their emotional tone, and B) how much fun the game was.

The mean rating for question A (emotion recognition) is 4.05 with a standard deviation of 1.09, which indicates excellent emotion recognition performance as perceived by the players of the game. The mean rating for question B (fun) is 4.57 with a standard deviation of 0.73, which proves that the game was well received due to its novel and innovative input mechanism, which has not been available for gaming so far.

5. Conclusion

This paper introduced the world's first computer game fully operated by emotions expressed by voice. It showcases the real-world applicability of ground-breaking, fundamental psychological research and theories through the VocEmoApI technology. A brief description of the VocEmoApI technology and the relation to Scherer's Component Process Model was given. Preliminary player feedback indicates a 4.05 out of 5 (best) score for the perceived accuracy of the emotion recognition and a 4.57 out of 5 (best) score for the fun-factor of the game.

We aim to make the game available as an Android App and add more levels in the future with further affect related game elements.

6. Acknowledgements

The project leading to this results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 680883).

References

[1] F. Eyben, M. Unfried, G. Hagerer, and B. Schuller, "Automatic Multilingual Arousal Detection from Voice Applied to Real Product Testing Applications," in *Proceedings 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2017.

[2] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[3] F. Eyben, B. Huber, E. Marchi, D. Schuller, and B. Schuller, "Real-time robust recognition of speakers' emotions and characteristics on mobile platforms," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 778–780.

[4] audEERING GmbH, "VocEmoApI presented to affective computing industry leaders," http://bit.ly/vocemoapi, April 2017, [Online; accessed 06-July-2017].

[5] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal processes in emotion: Theory, methods, research*, vol. 92, no. 120, p. 57, 2001.

[6] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological psychology*, vol. 87, no. 1, pp. 93–98, 2011.

[7] F. E. et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Jun. 2016.

[8] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM'MM 2013*. Barcelona, Spain: ACM, Oct. 2013, pp. 835–838.

[9] K. R. Scherer, "Introduction: Cognitive components of emotion." in *Handbook of the Affective Sciences*. Oxford University Press, 2003, p. 563.

[10] ——, "The dynamic architecture of emotion: Evidence for the component process model," *Cognition and emotion*, vol. 23, no. 7, pp. 1307–1351, 2009.

[11] K. Gentsch, D. Grandjean, and K. R. Scherer, "Cumulative sequential appraisals generate specific configurations of facial muscle movements: Evidence for the component process model of emotion." *PlosOne*, vol. 10, no. 8, p. e0135837, 2015.

[12] T. Bänziger, G. Hosoya, and K. R. Scherer, "Path models of vocal emotion communication," *PlosOne*, vol. 10, no. 9, p. e0136675, 2015.

[13] G. Hagerer, V. Pandit, F. Eyben, and B. Schuller, "Enhancing lstm rnn-based speech overlap detection by artificially mixed data," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.