Design and Evaluation of a Gesture Controlled Singing Voice Installation

Cornelius Poepel
Hochschule Ansbach
cornelius.poepel@hs-ansbach.de

Jochen Feitsch, Marco Strobel,
Christian Geiger
Fachhochschule Düsseldorf
{jochen.feitsch,marco.strobel,geiger}
@fh-duesseldorf.de

ABSTRACT

We present a system that allows users to experience singing without singing using gesture-based interaction techniques. We designed a set of body-related interaction and multimodal feedback techniques and developed a singing voice synthesizer system that is controlled by the user's mouth shapes and arm gestures. Based on the adaption of a number of digital media-related techniques such as face and body tracking, 3D rendering, singing voice synthesis and physical computing, we developed a media installation that allows users to perform an aria without real singing and provide the look and feel from a 20th century performance of an opera singer. We evaluated this system preliminarily with users.

Keywords

Gesture based musical interfaces, 3D character performance, singing voice synthesis, interactive media installation.

1. INTRODUCTION

Singing is a foundational and very old form of human expression. However, the authors see todays singing culture as more restricted than it was in former times. It is assumed that a demand exists to get closer to this central human ability and to deal with it in an experimental and playful way.

Our development does not focus basically on a simulation of the acoustic process of singing. We present a complete audiovisual setup including a singing interface to control a synthetic singing voice. Our goal is it to allow the performer to experience a credible feeling of singing with a voice different from the user herself. Further more, by making use of the singing system we seek to offer an entertaining experience for both the performers and listeners.

By installing an interface to the synthesis system we separate the voice from the body of the performer. In doing so, the singing voice becomes a more instrumental character.

Research outcome in music psychology shows that the visual appearance in performances play a role (in some cases a major role) for the perceived musical expression [1]. In an exaggerated way one might say that listening to musical expression is a process that is - in parts - done with the eyes. Since our system focuses the experience of singing on both performers and listeners our setup includes visual objects and a visualization of a singing avatar. Additionally, haptical feedback of singing vibrations is incorporated in order to enhance the bodily experience of singing.

2. RELATED WORK

Sundberg describes the necessary foundations in the research area of singing voice synthesis [13]. VocaWatcher generates realistic facial expressions of a human singer and controls a humanoid robot's face [9]. This is one of the first systems we identified that focuses on imitating a real singer based on acoustic and visual expressions. HandySinger provides hand puppets as an expressive and easy-to-use interface to synthesize different expressional strengths of a singing voice [14]. Experiments confirmed that it is very easy to gesture with a hand-puppet interface. Cano et al. described a system that allows to morph between the user's singing voice and the voice of a professional singer [2]. The system is used within a Karaoke performance setting.

D' Allessandro et al. [4] described an interesting way to interact with a Vowel-Consonant-Vowel singing voice synthesizer providing a fine grained control over voice parameters like pitch, vowels and strength. By creating Digiartic, a bi-manual device featuring two pen tablets, they provide a handbased voice synthesis. The non-dominant hand controls manner and place of articulation and consonant voicing while the dominant hand creates intonation and voice strength.

The non trivial mapping of gestures to sounds provides a real challenge if the dimensionality of the sensor-captured data is high. Fasciani et al. [6] applies an artificial neural network approach to simplyfy the mapping and increase the usability of their voice-controlled interfaces

Several projects studied the synthesis of sound with mouth shapes or with gestures. At NIME 2003 Lyons et al. presented a vision-based mouth interface that used facial action to control musical sound [11]. A headworn camera tracks mouth height, width and aspect ratio. These parameters are processed and used to control guitar effects, a keyboard and looped sequences. Closely related to our approach is the ”Artificial Singing” project, a device that controls a singing synthesizer with mouth movements that are tracked by a web camera [10]. The user's mouth is tracked and mouth parameters like width, height, opening, rounding, etc. are mapped to a synthesizer's parameters e.g. pitch, loudness, vibrato, etc. De Silva et al presented a facetracking mouth controller [5]. The image-based recognition tracks the user's nostrils and mouth shape and maps this to a syrinx's model to generate sounds. Recently, Cheng and Huang published an advanced mouth tracking approach that combines realtime mouth tracking and 3D reconstruction of the mouth movement in real-time [3].

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

3. SYSTEM OVERVIEW

The installation consists of a 3x3 video wall with 46" monitors, the motion tracking system Noraxon MyoMotion (www.noraxon.com) for full body skeleton tracking, and a PrimeSense Carmine for facial tracking. To enhance the user experience of being an opera singer from the 1920, the user wears a tailcoat and a top hat including bone conduction headphones to place the generated tenor's voice "inside" the user's head without blocking surroundings sounds. This is important in order to hear the background music of the sung track. Additionally, a bib is equipped with several exciters and vibration modules to induce the vibration of the singing voice into the user's thorax. Using bone conduction and vibrational feedback, the illusion of singing is supported.

A gramophone is connected to the computer as a resonating body for audio playback. It is also used as an input device for choosing between several songs by placing a shellac disc (gramophone record) on the gramophone and as a switch to trigger the performance to start. A glove is equipped with switches that can be activated by closing thumb and index finger. Closing the switch is used to alternate between two sets of vowels that could not be detected by mouth shape alone (see subsection "Tangible Illustration").

The processing is done on two network-connected computers. One operates the facial tracking hardware and software and the synthesizer modules. The other runs the main application with a skeleton tracking system and rendering. The computer running the main application processes all tracking data, uses it to fully animate a virtual opera singer in 3D, controlled by the user, and sends relevant data to the synthesizer module.

Initially, the user steps in front of the main computer which renders the content on the video wall. She can now take a picture of herself to make the avatar's head look like her, or customize it further. After that she starts the performance mode by putting the gramophone's pickup arm on top of the selected record and steps onto the stage. By moving her arms and shaping vowels with her mouth, she can not only control the movements of the virtual tenor, but also make him sing. The user's goal is to reproduce the original singing of the selected track as accurately as possible. With an increased positive rating the 3D avatar's face morphs from the original user face to the face of a famous opera singer. For an additional increase of a 1920s performance experience we applied a shader that emulates the visual presentation of an old silent movie film. In addition a simple visualization is shown to tell the user whether he is singing correctly or not and how to adjust his performance while singing an aria. This only concerns the pitch; volume and vowels are up to the user's desire and experimentation.

4. USER TRACKING

After the avatar generation step the user can initiate the performance mode in which face and body of the user are tracked. MyoMotion is a portable, wireless and expandable motion tracking system that can provide three-dimensional orientation information of two to 36 IMU (Inertial Measurement Unit) sensors. Our setup utilizes five sensors to track the upper body: one for each upper arm, one for each forearm and one at the back to root the coordinate system to the user.

4.1 Facial tracking and calibration

The facial motion tracking system used is an adaption of faceshift (www.faceshift.com). This face tracking approach provides sufficiently accurate data based on a user profile that has to be created for each user in advance. Furthermore we introduced an additional calibration step to increase the recognition accuracy. A neural network is used to map the tracking data provided by faceshift to corresponding vowels. To do this we feed the neural network with data taken while the user forms each desired mouth shape for several seconds. There are four mouth shapes that have to be trained to the neural network: "A", "E", "O" and closed mouth. The second vowel set is triggered by hand gesture using the same mouth shapes.

Tracking data includes head pose information, arm gestures, body position, facial blend shapes (also called coefficients), eye gaze and additionally specified virtual markers. The coefficients are fed to the trained neural network and the probability values of each vowel are sent directly to the audio synthesizer, using the OSC protocol (Open Sound Control), where they are processed for sound generation (see section six "mapping and sound synthesis"). In addition to the audio processing the head pose, blend shapes and eye gaze data are used to animate the avatar's facial expression in real time. The head pose is used to rotate the neck bone and the eye gaze to rotate the specific eye while the blend shapes are used to change the look of the avatar's face using the morph capabilities of the FaceGen SDK (www.facegen.com).

4.2 Full body tracking

The full body tracking uses the provided joint data from the body tracker and maps it onto the avatar. In addition, tracking data from hand, elbow and shoulder joints are sent to the singing voice synthesizer to control the volume and pitch parameters. The arm stretch is calculated by adding shoulder-to-elbow and elbow-to-hand distance, devided by the distance of shoulder-to-hand. This gives a normalized value. The hands' position height (y-axis) is subtracted from the shoulders' positions height and is divided by the maximum arm stretch to get a normalized range of approximately -1 to 1. This value is used to determine the pitch value. These calculations are done separately for each arm, choosing the larger of both values.

5. TANGIBLE ILLUSTRATION

The newly designed interaction techniques include the application of a gramophone, suitable clothes and appropriate acoustic, visual and haptic feedback for the performing user: We realized the selection of a dedicated piece of music using an old gramophone with shellac discs. Currently we provide two different suitable arias for selection (Ave Maria – Schubert and Nessun Dorma – Puccini). The user can select his favorite background music by placing the relevant disc on the gramophone.

Real singing results in hearing one's own voice inside head and body and in vibrational feedback caused by lung pressure and movement of the syrinx's vocal folds. To provide a similar user experience we developed a prototype vibrational vest by equipping a bib with vibrational actors, exciters and a sound device that triggers the actors based on sound input.

Two vowel sets were defined that can be switched by the user by performing an additional interaction technique. This was necessary as it is difficult to detect all vowels based on facial depth tracking only. There are two alternative techniques, eye brow raising and using a simple glove switch; both gestures are not realistic behavior but work well after a short training period. Using the additional interaction technique, the vowels "Ä" [æ:], "I" [i:] and "U" [u:] are synthesized instead of ”A” [a:],”E” [e:] and ”O” [o:]. More details can be found in [8].

6. MAPPING AND SOUND SYNTHESIS

By combining mouth shapes and arm gestures, the user alters several parameters of the singing voice. The singing voice is fully synthesized, providing a considerable amount of flexibility.

6.1 Arm gestures and "harmonizer"

The volume is controlled directly by the user’s arm stretch: outstretching the arms increases, bending them reduces the volume. The sung pitch is determined by the arm’s height.

The obvious method of mapping the arm height directly to the pitch results in a hard to play, Theremin-like behavior. As an enjoyable user-experience is a main concern of this project, this method would have been counterproductive. Two alternative methods were designed to solve this problem:

Method one quantizes the arm-height into 25 fixed steps plus filtering out small arm movements to prevent changing rapidly back and forth if the arm position is borderline between two steps. Every step is mapped to a specific MIDI-pitch, ranging from 41 (ca 87 Hz) to 65 (ca 349 Hz). This way, playing is a little easier, but still so hard that the user can become frustrated.

As the installation is destined especially for musically inexperienced users, the second method of pitch selection only pays attention to the direction of the next desired note: moving the arms up triggers the next higher note, moving the arms down triggers the next lower note. For this the system consults the song’s melody (provided as MIDI-Track) and chooses the next higher respectively lower note on the basis of the song’s correct melody or current key. We designed the system to choose notes that always sound more or less suitable to the backing music of the aria. For a more detailed overview of the harmonizer module refer to [7].

6.2 Mouth Shapes and Synthesis

The user can control the sung vowel by forming it with his mouth. The trained neural network determines the likeliest vowel and sends it to the sound synthesis. To prevent rapid changes between vowels due to noise and inaccuracy of the vowel tracking, five consecutive detections of the same vowel are necessary to trigger a vowel change. This produces a short but acceptable delay. To change the vowel, the sound synthesis modifies the filters used for the formant synthesis. A short interpolation between the current vowel and the target vowel prevents sound artifacts. This interpolation must not be too long, as an audible vowel-slide sounds odd.

The basis of the singing synthesis is a vowel synthesis via formants. The vowel’s formant frequencies used in this work are averaged values obtained from [12]. Max/MSP in combination with Java Script is used for the audio processing. The synthesis engine using additive synthesis is controlled via MIDI (provided by the user’s arm height) to determine the base frequency of the voice.

Band pass filters produce the desired formants necessary for the target vowel. Two formants are sufficient for the signal to be recognizable as a human’s vowel, but three formants were identified to be useful to achieve a more human sound. A fourth formant at 2,7 kHz is used to produce a tone color similar to an opera singer. A formant around 3 kHz in the human voice is called ”singers’ formant” and can only be found in the frequency spectrum of highly trained singers.

7. EVALUATION

To evaluate the system six subjects tested the entire procedure, including face calibration, training of the neural network and singing first freely and then performing both Ave Maria and Nessun Dorma. Afterwards they were interviewed and filled in the AttrakDiff^{TM} (attrakdiff.de) questionnaire. Three of the subjects had a considerable amount of singing experience (singing in a choir or band), the others had none.

AttrakDiff^{TM} is used to measure the attractiveness of interactive products using a dedicated questionnaire with opposite attributes for each question. Attributes refer to different evaluation categories. The pragmatic quality (PQ) describes the usability, and how effective the user could achieve her goals, the hedonic quality – stimulation (HQ–S): shows how the product supports the user in her desire to make progress by providing novel, interesting and inspiring functionalities, contents and ways of interaction and presentation. The hedonic quality – identity (HQ–I) gives the amount of the user’s identification with the product and the attractiveness (ATT) shows the global rating of the product.

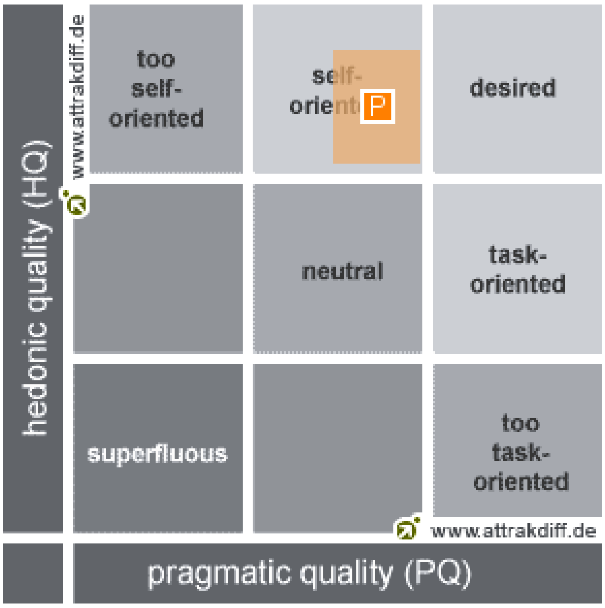

Figure 1: AttrakDiff^{TM} (attrakdiff.de) test results

PQ and HQ are independent factors and add approximately even to the ATT level. The installation was classified as ”self-oriented” (see figure 1). This means for the HQ that the user has a good identification with the product and is stimulated and motivated.

The average values of the evaluation categories are plotted on Figure 2. All four categories are above average. Especially the HQ–S achieved a very high scoring and the overall attractiveness is strong as well, showing that the goal of creating an enjoyable user experience could be achieved. The users also had a quite strong identification with the installation, what is to be seen as a common effect when having an enjoyable experience. Compared to the other categories, the PQ is quite low. Improving the usability would therefore raise the overall attractiveness furthermore.

To identify the specific needs in order to improve the usability of the installation an analysis of the interviews was important. The following main statements could be extracted from the interviews:

- All six test users found using the installation enjoyable

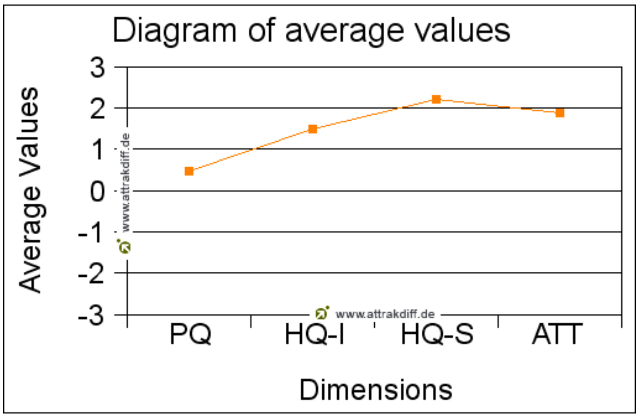

Figure 2: Mean values of the four categories

and entertaining. Three of the users reported their technical fascination as an argument for a positive user experience, the others simply enjoyed the experience itself.

- The sound synthesis was uniformly referred to as at least believable. Only one subject mentioned that she rarely identified herself with the artificial character of the singing voice.
- The three test persons without singing experience could imagine, that the installation conveyed a sense of physical feeling as if actually singing. The three users with singing experience didn't find the physical feeling realistic.
- All but one tester found using the installation very challenging. However, all but one of the subjects were positive about being able to learn using it properly after a couple of attempts. One subject found it easy to use from the very beginning.
- The consensus was, that a better visualization would greatly enhance the usability of the installation.
- No user found the calibration and training length disturbingly too long.

Summarizing, this preliminary test reveals a positive tendency but due to the small number of subjects the test was not designed as a statistically reliable evaluation.

8. CONCLUSION

The described prototype works as expected with all components described in this paper but current limitations concerning the set-up / calibration time for a user and fault tolerance of some system parts need to be resolved before we conduct larger user tests. Moreover, some items such as the glove and vest have to be improved to provide a robust demonstration for a larger audience. However, we plan to experiment with expanded possibilities and additional functions. Since the present singing synthesizer only works with vowels, it would be very interesting to see how the system works when consonants are added. This correlates with the wish for increased possibilities for articulation.

On the list for future developments are different and shapeable consonants, additional arias of course and the addition of different voices, e.g. a soprano voice as well as voices with different personal characters i.e. more extroverted or introverted. In far away scenario it might be possible to record the voice of a singer, extract meaningful characteristics defining the identity and personality of the voice and applying these characteristics to the singing synthesizer.

9. REFERENCES

[1] K.-E. Behne and C. Wöllner. Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication. *Musicae Scientiae*, 15(3):324–342, 2011.

[2] P. Cano, A. Loscos, J. Bonada, M. D. Boer, and X. Serra. Voice morphing system for impersonating in karaoke applications. In *Proc. of the 2000 Int. Computer Music Conference*, pages 109–112, Berlin, Germany, 2000.

[3] J. Cheng and P. Huang. Real-time mouth tracking and 3d reconstruction. In *3rd Int. Congress on Image and Signal Processing (CISP)*, volume 4, pages 1524–1528, 2010.

[4] N. D'Alessandro, C. d'Alessandro, S. L. Beux, and B. Doval. Real-time calm synthesizer new approaches in hands-controlled voice synthesis. In *Proc. of the 2006 Conference on New Interfaces for Musical Expression*, pages 266–271, Paris, France, 2006.

[5] G. C. de Silva, T. Smyth, and M. J. Lyons. A novel face-tracking mouth controller and its application to interacting with bioacoustic models. In *Proc. of the 2004 Conference on New Interfaces for Musical Expression*, pages 169–172, Hamamatsu, Japan, 2004.

[6] S. Fasciani and S. Wyse. A self-organizing gesture map for a voice-controlled instrument interface. In *Proc. of the 2013 Conf. on New Interfaces for Musical Expression*, pages 507–511, Daejeon, Korea, 2013.

[7] J. Feitsch, M. Strobel, and C. Geiger. Singing like a tenor without a real voice. In *Advances in Computer Entertainment*, pages 258–269, Twente, Netherlands, 2013.

[8] J. Feitsch, M. Strobel, S. Meyer, and C. Geiger. Tangible and body-related interaction techniques for a singing voice synthesis installation. In *Proc. of the 2014 Conference on Tangible and Embedded Interaction*, pages 157–164, Munich, Germany, 2014.

[9] M. Goto, T. Nakano, S. Kajita, Y. Matsusaka, S. Nakaoka, and K. Yokoi. Voca-listener and voca-watcher: Imitating a human singer by using signal processing. In *Proc. of the 2012 IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 5393–5396, 2012.

[10] A. Hapipis and E. R. Miranda. Artificial singing with a webcam mouth-controller. In *Proc. of the 2005 Int. Conference on Sound and Music Computing*, 2005.

[11] M. J. Lyons, M. Haehnel, and N. Tetsutani. Designing, playing, and performing with a vision-based mouth interface. In *Proc. of the 2003 Conference on New Interfaces for Musical Expression*, pages 116–121, Montreal, Canada, 2003.

[12] G. E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

[13] J. Sundberg. The KTH synthesis of singing. *Advances in Cognitive Psychology*, 2(2-3):131–143, 2006.

[14] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure. Handysinger: Expressive singing voice morphing using personified hand-puppet interface. In *Proc. of the 2005 Conference on New Interfaces for Musical Expression*, pages 121–126, Vancouver, Canada, 2005.