The voice painter

Amalia de Götzen   Riccardo Marogna   Federico Avanzini
Sound and Music Computing group
Dep. of Information Engineering, University of Padova, Italy
E-mail: {degotzen, marognar, avanzini}@dei.unipd.it

Abstract

Very often when looking at a painting or touching a sculpture it is possible to experience the meaning of being a perceiver as enactor of perceptual content. The piece of art that we try to explore forces us to move around the object in order to discover new meanings and sensations. We need to interact with the artistic object in order to completely understand it. Is it possible to think about sound and music in this way? Is the auditory and musical experience prone to such investigation? This paper tries to describe an enactive system that, by merging the concepts of autographic and allographic arts, transforms the spectator of a multimodal performance into the performer–perceiver–enactor. The voice painter, an instrument to paint with our voice movement in a closed loop interaction, offers a new artistic metaphor as well as a potentially useful tool for speech therapy programs.

1. Introduction

The topic of this work is inherently multidisciplinary: in order to discover and emphasize the enactive approach in artistic productions (dance, music, painting, sculpting etc.) by means of new technologies it is necessary to bridge the gap between technology and art, taking into consideration suggestions and needs from artists and constraints and possibilities from technicians. Arts and enaction can be considered strictly related even in classical works, where the technology is simply a brush and some colors, or the material of a sculpture and the tools used by the artist to build his piece of work. Two well known examples are the works of Close in painting [8] and Serra in sculpting [18].

However, new technologies for artistic expression that contemplate multimodal interaction give to the artist new tools and new ways to think about their work, involving the final users in an enactive interaction while experiencing a specific work of art. The idea of the *voice painter* device presented in this paper is to have simple – non-intrusive – technology that stimulates the user-perceiver-enactor to use her/his body and in particular his voice in order to create a painting, and discover the relation between her/his actions and the signs that she/he can produce. We think that this kind of approach is genuinely enactive and that it suggests a “third path” between allographic and autographic arts.

The paper is organized into three main sections. Section 2 summarizes the state of the art and main views about how enactive experiences inform artistic representations. In Sec. 3 the voice painter system is presented and described in its technical details while Sec. 4 is devoted to discussing applications of the system.

2. Autographic and allographic arts

One of the main categorizations between different forms of arts is the one introduced by Goodman [11] which defines ‘autographic’ and ‘allographic’ arts:

> the former cannot be noted and do not contemplate performance, while the latter can be translated into conventional notation, and the resulting ‘score’ can be performed with a certain freedom of variation.

Painting and music are the two artistic expressions that are generally used to exemplify this distinction. It is difficult to determine the rules that generated a given painting, there is no notation that can help someone else to produce an exact replica of an original piece of art: it is even possible to define every copy a ‘forgery’. In music the point of view is totally different: every copy/performance of a piece is a possible interpretation. Notation allows many different musicians to play a given piece of music: the ‘discrete’ musical signals are first notated by the composer and then interpreted by the musicians. One can say that while autographic arts are one-stage arts, allographic arts are two-stage arts. The distance between these two forms of art can be dramatically reduced in modern performances, where, for instance, a painting can be seen as the result of a live performance: a dancer that paints with her/his body or a musician that controls some multimodal device that produces a video output while playing.

2.1 Enaction in Arts

One of the main achievement of the ENACTIVE project has been a deep and fruitful reflection on the role of enaction in the artistic creation process [3]. The topic is particularly hard to address since it links together abstract concepts which are difficult to define (enaction, creation): the Enactive/07 Conference in Grenoble has collected several contributions that can be analyzed in order draw a sort of "red line" that goes across different artistic expression with the common intention of exploring the enactive creative process.

The enactive theory of perception states that it is not possible to disassociate perception and action schematically, since every kind of perception is intrinsically active and thoughtful. In this view, experience is something that an animal *enacts* as it explores its environment [24, 19]. In this view, the subject of mental states is the *embodied* animal, situated in the environment.

Enactive knowledge can be acquired also when discovering a painting or a sculpture if the perceiver is immersed in this action-perception loop. The typical example of an enactive art is music: a violin player needs to feel and to hear the sound in order to adjust the performance. Following this perspective many enactive artistic applications created with the support of technology explore virtual instruments through different kinds of gestures or postures. At the same time these applications have to consider the specific feedback received e.g. when exploring a surface or when using a bow on a string [5]: we perceive through our hands and fingers a specific haptic sensation that stimulates the user/player to react in order to understand. Virtual musical instruments are then augmented with haptic devices that can render the surfaces and the forces involved while playing a real instrument.

2.2 Painting with voice

The use of the voice as enactive instrument is rather unexplored, particularly so in the context of artistic applications. Voice is a universal human-to-human communication means and is used to convey also non-verbal, paralinguistic elements including emotion, prosody, stress. Moreover voice and speech are always accompained by other non-verbal communication channels, such as facial expression, gesture, and body language, forming a single system of communication [17].

These observations provide the motivation for the development of an interface that uses vocal expression as a tool for creating visual signs. The central idea is to exploit the most relevant features of vocal expression and map them into graphic features, thus creating a simple and versatile instrument that can be used by an experienced performer as well as by a naif user. Similar ideas have been recently explored in [13] and [16]. In [13] the aim is the development of a drawing system for users with motor impairments, therefore voice is used as a controller (a "vocal joystick") which assigns different non-verbal sounds to different directions or actions. Due to the specific application field, the user voice is the only allowed input. Since the present work is focused on a different application field, it does not have this constraint and full body movement is used as a second input. Moreover, the mapping presented in [13] does not follow a *phonesthetic* [15] approach similar to the one presented here (except for loudness mapping).

The work presented in [16] shares more similarities with our main concept. In the concert performance *"Messa di Voce"*, the sounds produced by two vocalists are augmented in real-time by interactive visualization software. The mapping approach is similar to the one presented here, although in some scenarios more *iconic* mappings are used instead. The main difference lies in the localization/tracking system: while in [16] this is based on cameras and computer vision techniques, we introduce a microphone array-based localization system: in this way tracking is based exclusively on the incoming voice direction, and a single audio input system provides all the information for the mapping procedure.

In our scenario, the player will be able to paint on a black screen using her/his voice. The mouth can be considered the brush: in order to draw on the entire screen surface the user will be forced to move, therefore involving the whole body rather than just the voice as an input instrument. The action-perception closed loop is then recreated with the help of a system that will be presented in the next section.

3. System description

The system integrates an 8-microphone array and TDOA (time delay of arrival) estimation technique for localization and tracking of the user position (see Fig. 1). The user is supposed to move in the active area at a given distance from a screen. The graphic rendering is projected onto the screen in order to satisfy a full correspondence between voice source position and rendering position. In this way, the user is supposed to have

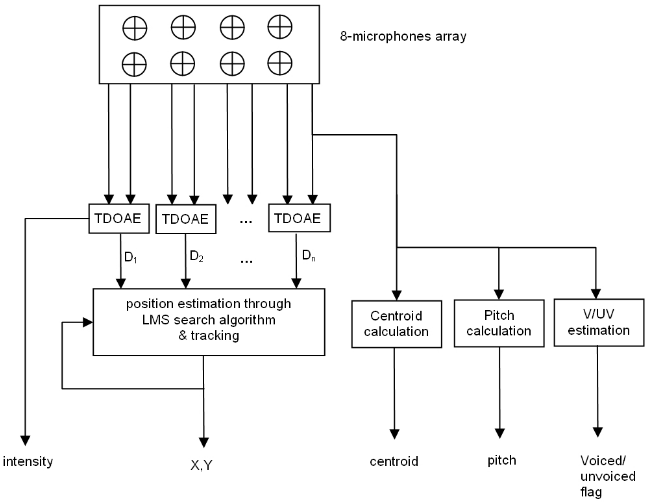

Figure 1: Block scheme of the system.

a more natural interaction with the virtual canvas, since the visual feedback is temporally and spatially correlated with the voice event.

3.1 Voice tracking system

The audio system performs real-time user localization and tracking through a 2-stage algorithm. First, TDOA estimation is performed for each pair of microphones. Then, the obtained vector of delays is processed in a LMS (*Least Mean Square*) algorithm in order to extract the estimated position.

The TDOA estimation stage implements an ATF (Acoustical Transfer Function) ratio estimation algorithm [10]. In a previous work [6] this method was recognize to achieve the best performances compared to other algorithms. The main idea behind the algorithm is to exploit some peculiar voice features in order to extract the signal of interest from the background noise and track it in a robust way. We made the assumption that background noise can be impulsive (e.g., a closing door) or stationary (fans, air conditioning systems,...). Noise events are therefore stationary or very short in time, while human voice is quasi-stationary (it can be considered stationary in a short-time frame of $20 - 30$ ms, while it changes its statistics from frame to frame). The algorithm exploits this assumption in order to estimate the TDOA for a given pair of received signals. Therefore, a vector of delays (VOD) is obtained.

In the second stage, the estimated VOD is compared to the elements of a pre–computed VOD matrix. This matrix is obtained through a discretization of the vertical plane of interest in the active area, and the subsequent calculation of the VOD for each position. Searching for the best fitting VOD in the search matrix with LMS criterion results in an estimated (x, y) position in the vertical plane of interest.

Although this stage achieves a robust position estimation (since it tends to ignore erratic delays and searches for a coherent figure), its main drawback is the poor performance due to the time-consuming search through the entire VOD matrix. This limitation can be improved, however, using a tracking technique. Since the voice source is supposed to move slowly in the active area with respect to the localization algorithm, position estimates that are close in time are supposed to be close in space too. Therefore, the second stage of the algorithm is improved by restricting the search to neighboring positions in an area surrounding the previous estimation. The exhaustive matrix search is then performed only at voice attacks. The described localization technique was implemented as an external module written in C for the Pure Data platform [20]. In this way, the algorithms have been highly optimized for real-time processing.

3.2 Feature extraction

Human voice can be characterized by several features, and human beings are able to control vocal emission in order to modify most of these features. We identified some prominent features which were relatively straightforward to be extracted with real-time algorithms. They were:

- intensity, computed as the RMS value of the squared voice pressure signal;
- spectral centroid, i.e. the center of gravity of the spectral magnitude computed over an audio frame;
- a voiced/unvoiced flag, depending on whether the utterance is associated to pseudo-periodic vocal fold vibrations or not;
- pitch, i.e. the subjective attribute of sound height.

Although the first two features are defined as usual and easy to extract, the other two are not so trivially estimated and therefore they are briefly described here. The voicing flag indicates the presence of a voiced signal, i.e. a signal containing periodicities due to vocal fold vibrations. This kind of signals can be detected with various approaches. We implemented a technique which combines zero-crossing detection and cepstrum extraction [1]. Pitch estimation is facilitated in this case by the relatively simple harmonic structure of voiced utterances, so that the problem reduces to estimation of the

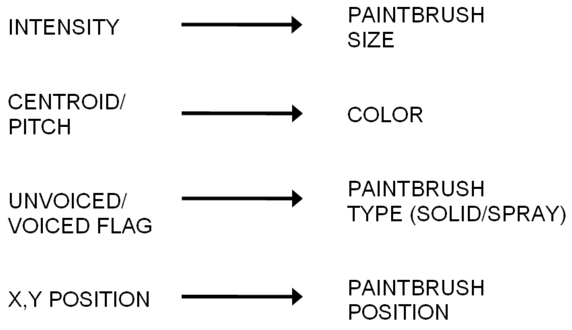

Figure 2: Mapping of voice features into graphic features.

fundamental frequency. This is estimated using an algorithm that extracts and matches harmonic spectral components on successive frames of the vocal signal [21].

3.3 Graphic rendering and feature mapping

Real time graphic rendering was performed using GEM [26], an OpenGL library designed to be integrated on Pure Data. The main idea was to construct a mapping of vocal features into well-recognizable graphic features (size, color, geometry). The choice of the mapping is crucial in designing the interface and dramatically influences the ability to control the digital instrument. Mapping strategies are the subject of several works in HCI literature and particularly for the design of digital musical instruments [14, 25].

As a starting point, the initial mapping was organized as in Fig. 2. It can be noticed that, although some features have a somehow immediate and intuitive mapping, for some other features such as pitch or voicing the corresponding graphic effect is quite arbitrary. For example, establishing a correspondence between sound frequency content and light frequency may not necessarily seem natural to the user, even though it may seem logical on numerical grounds.

The resulting visual effect is a sort of *abstract sketch* which can contain well defined geometric elements produced by short voiced segments, and/or particle-like signs due to unvoiced segments. The intensity and pitch/centroid information results in different color gradients and sizes. A snapshot of the resulting graphic rendering is shown in Fig.3.

Different and possibly more intuitive color mappings could be obtained e.g. in a RGB color field, by associating lower pitches to “hot” colors and higher pitches to “cold” ones. Similarly, HSV or HSL scales could be used to fit the pitch class scale. There is hardly any literature about applications that use this particular mapping and compare different possible strategies, therefore a simple one-to-one mapping was used in the first implementation.

An evaluation of this mapping can be based only on several subjective tests which are planned in the near future. Users will be asked to paint with their voice for 10 minutes to experience the interface. At the end of this preliminary training session they will be asked to perform several tasks using different mapping strategies. All tasks will concern the reaching of specific colors using the voice painter. Finally, the users will be asked to assess the quality of the mapping strategy at hand. The data will be then analyzed: timings, precision and paths will be collected and correlated to provide a quantitative assessment of each mapping strategy.

4. Discussion

A first informal test of the system was conducted with several users during various demo sessions (including the Enactive/07 Conference [3]). Users were allowed to interact with the virtual canvas without any hint about the graphic mapping and without any specific task (the only suggestion was: “use your voice to paint”). The goal of the test was to evaluate:

1. the way the user approaches the canvas and explores its voice features;
2. how many vocal features the user is able to identify in the mapping;
3. whether she/he is able to control an identified feature in order to obtain a desired graphic effect.

At the end of the test the user was asked to judge if the experience was natural for her/him, and in which cases.

This preliminary test showed that most of the features were identified by the users (except for the mapping pitch/centroid→color, which was not clearly recognized in several cases). It also confirmed that some mapping assumptions were natural for most of the users (e.g. intensity→size, mouth position→paintbrush position, voicing type→paintbrush type).

4.1 Robust speaker localization/tracking

Current developments are targeted at improving the robustness of the localization and tracking subsystem, especially in noisy, real-life scenarios.

The TDOA estimation and LMS algorithms described in Sec. 3.1 only consider impulsive or stationary noise background, and the performance is degraded

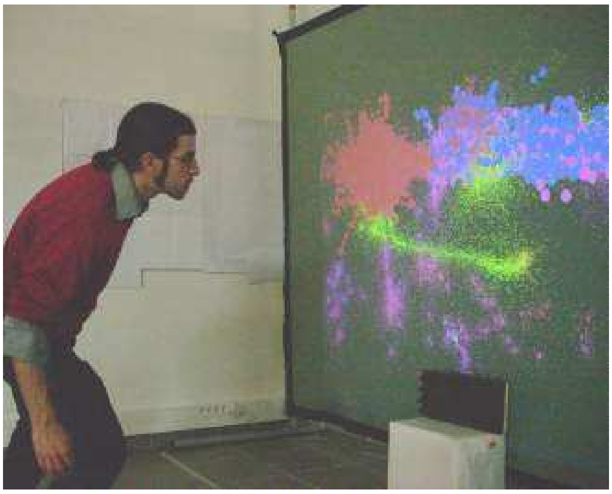

Figure 3: Performing with the voice painter.

in the presence of low SNR values especially when the noise is non-stationary. In order to improve robustness the first improvement in the tracking procedure is the inclusion of a more general voice activity detection (VAD) algorithm, that employs higher order statistics in order to allow detection of presence/absence of human speech in any region of audio [23].

A bimodal system is also being developed, in which localization and tracking is performed not only from the voice signal, but also using video. A preliminary implementation has been realized using the EyesWeb software platform for gesture analysis [7].

One more challenging issue concerns the problem of simultaneous localization of multiple acoustic sources, which would open up new possibilities for applications in cooperative scenarios. This will require the use of advanced statistical methods such as particle filtering [2].

4.2 Applications in speech therapy

Besides applications in performing arts and entertainment, the proposed system has the potential to be employed as an assistive technology in the field of speech therapy. HCI techniques are increasingly used as a method to teach and reinforce vocalization and speech skills in various contexts. It is generally acknowledged that interfaces with visual and multimodal biofeedback can influence the communication of individuals and can help facilitate in particular the speech and vocalization education process for children with communication skill deficits, both by motivating and rewarding vocalization and by providing information about the acoustic properties of vocalizations [4, 9, 12, 22].

Areas of application include *speech disorders* due to physical impairments or problems in motor planning and coordination (dysarthria, dyspraxia). Previous works have shown that computer-based speech training systems can help to improve articulation [4], and to attain correct production of specific sounds [22]. A second potential area of application concerns *speech delay* problems (i.e. situations in which speech development follows the usual patterns but at a slower rate than normal). In this case previous works have demonstrated the utility of systems that, through real-time analysis of vocalizations and appropriate biofeedback, reinforce of the production of syllabic utterances associated with later language and cognitive development [9]. A third potential area of application concerns problems associated to communication, social functioning, and expression, such as autistic spectrum disorder (ASD). Applied behavior analysis techniques are typically used, in which target behavior is rewarded e.g. with toys and the rewards are then gradually removed over time. This approach has been recently adopted in the design of computer-assisted systems that encourage playful behavior via technology, with the additional advantage that technology and computers reduce the apprehension caused by human-to-human interaction [12].

5. Conclusions

This paper presents an interface that allows explorations in different directions. The voice painter was born with artistic applications in mind: to create an instrument that could mediate between allographic and autographic arts, allowing sophisticated performances based on a musical notation (used then to paint) as well as improvisations or simple entertainment. The instrument has been first demonstrated at the Enactive/07 Conference and has spurred an active interest from the public and in particular among visual artists. Several different applicative scenarios have provided ideas for further development of the instrument: besides artistic applications, those related to speech theraphy and communication disorders are the most promising ones.

6. Acknowledgments

We would like to thank Barbara Mazzarino and Giovanna Varni for their useful suggestions and their help in developing the first implementation of the system, presented at the Enactive/07 Conference in Grenoble.

References

[1] S. Ahmadi and A. S. Spanias. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. *IEEE Trans. on Speech and Audio Processing, vol. 7, no. 3*, May 1999.

[2] F. Antonacci, D. Riva, M. Tagliasacchi, and A. Sarti. Efficient localization and tracking of two acoustic sources using particle filters with swarm intelligence. In *Proc. EURASIP European Sig. Process. Conf.*, Poznan, 2007.

[3] Association ACROE, editor. *Proc. 4th Int. Conf. on Enactive Interfaces (ENACTIVE/07...Enaction_in_Arts)*, Grenoble, 2007.

[4] O. Bälter, O. Engwall, A.-M. Öster, and H. Kjellström. Wizard-of-oz test of artur: a computer-based speech training system with articulation correction. In *Proc. 7th Int. ACM SIGACCESS Conf. on Computers and accessibility (ASSETS'05)*, pages 36–43, Baltimore, 2005.

[5] C. Cadoz. Musical creation process and digital technology. the supra–instrumental gesture. In *4th International Conference on Enactive Interfaces (Enactive'07)*, Grenoble, France, November 2007.

[6] G. Calvagno, C. Trestino, M. Romanin, and R. Marogna. A dsp implementation of a real-time speaker localization system using atf-ratio time delay estimation. *EDERS*, pages 1–10, Sept. 2006.

[7] A. Camurri, B. Mazzarino, and G. Volpe. Analysis of expressive gesture: The EyesWeb expressive gesture processing library. In A. Camurri and G. Volpe, editors, *Gesture-based Communication in Human-Computer Interaction*. LNAI 2915, Springer Verlag, 2004.

[8] C. Close, R. Storr, K. Varnedoe, D. Wye, and G. D. Lowry. *Chuck Close*. The Museum of Modern Art, New York, 2002.

[9] H. Fell, C. Cress, J. MacAuslan, and L. Ferrier. visiBabble for reinforcement of early vocalization. In *Proc. 6th Int. ACM SIGACCESS Conf. on Computers and accessibility (ASSETS'04)*, pages 161–168, Atlanta, 2004.

[10] S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with application to speech. *IEEE Trans. Signal Processing, 49 (8)*, 2001.

[11] N. Goodman. *Languages of Art*. Hackett, 1985.

[12] J. Hailpern, K. Karahalios, J. Halle, L. DeThorne, and M.-K. Coletto. Visualizations: speech, language & autistic spectrum disorder. In *Proc. ACM Computer-Human Interaction Conf. (CHI'08)*, pages 3591–3596, Firenze, 2008.

[13] S. Harada, J. O. Wobbrock, and J. A. Landay. Voicedraw: a hands-free voice-driven drawing application for people with motor impairments. In *Assets '07: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 27–34, New York, NY, USA, 2007. ACM.

[14] A. Hunt, M. M. Wanderley, and M. Paradis. The importance of parameter mapping in electronic instrument design. In *NIME–02: Proceedings of the 2002 Conference on New Instruments for Musical Expression*, pages 149–154, Dublin, Ireland, 2002. Media Lab Europe.

[15] W. Khöler. *Gestalt Psychology*. Liveright Publishing Corporation, New York, 1927.

[16] G. Levin and Z. Lieberman. In-situ speech visualization in real-time interactive installation and performance. In *Proceedings of The 3rd International Symposium on Non-Photorealistic Animation and Rendering*, Annecy, France, June 2004.

[17] D. McNeill. *Gesture and Thought*. University of Chicago Press, Chicago, 2005.

[18] K. McShine, L. Cooke, J. Rajchman, B. Buchloh, and R. Serra. *Richard Serra Sculpture: Forty Years*. The Museum of Modern Art, New York, 2007.

[19] A. Noe. *Action in perception*. MIT press, Cambridge, Mass., 2005.

[20] M. Puckette. Max at seventeen. *Computer Music J.*, 26(4):31–43, 2002.

[21] M. Puckette and T. Apel. Real-time audio analysis tools for pd and msp. In *Proc Int. Computer Music Conf.*, pages 109–112, San Francisco, 1998.

[22] L. I. Shuster, D. M. Ruscello, and A. R. Toth. The use of visual feedback to elicit correct /r/. *Am. J. Speech-Language Pathology*, 4:37–44, 1995.

[23] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Trans. Speech Audio Process.*, 8(4):478–482, 2000.

[24] F. Varela, E. Thompson, and E. Rosch. *The Embodied Mind*. MIT Press, Cambridge, MA, 1991.

[25] M. M. Wanderley and N. Orio. Evaluation of input devices for musical expression: Borrowing tools from HCI. *Computer Music J.*, 26(3):62–76, 2002.

[26] J. M. Zmölnig. Gem, http://gem.iem.at.