PERFORMANCE ANALYSIS AND SCORING OF THE SINGING VOICE

OSCAR MAYOR[1], JORDI BONADA[1], AND ALEX LOSCOS[2]

[1] *Music Technology Group (MTG), University Pompeu Fabra, Barcelona, SPAIN*
oscar.mayor@upf.edu
jordi.bonada@upf.edu
[3] *Barcelona Music and Audio Technologies (BMAT), Barcelona, SPAIN*
alex@bmat.com

In this article we describe the approximation we follow to analyze the performance of a singer when singing a reference song. The idea is to rate the performance of a singer in the same way that a music tutor would do it, not only giving a score but also giving feedback about how the user has performed regarding expression, tuning and tempo/timing characteristics. Also a discussion on what visual feedback should be relevant for the user is discussed. Segmentation at an intra-note level is done using an algorithm based on untrained HMMs with probabilistic models built out of a set of heuristic rules that determine regions and their probability of being expressive features. A real-time karaoke-like system is presented where a user can sing and visualize simultaneously feedback and results of the performance. The technology can be applied to a wide set of applications that range from pure entertainment to more serious education oriented.

INTRODUCTION

Singing voice is considered to be the most expressive musical instrument. Singing and expressing emotions are strongly coupled, making clearly distinguishable when a singer performs sad, happy, tender, or aggressive.

Many people have been working in the scientific field of performance analysis of the singing voice including solo or polyphonic singing voice transcription [1,2], score alignment [3,4] and expressivity [5] but there is a lack in references about automatic expressive detection or transcription and expression categorization of singing performances as it is focused in this article.

We present here a tool for the automatic evaluation of a singing voice performance with precise note segmentation and expression detection. The application includes a friendly graphical interface for visualization of the analyzed song descriptors including pitch, vibratos, portamentos, scoops, attacks, sustains and releases.

We also compare our system with existing systems for evaluation of singing music including musical teaching applications and musical karaoke-like games.

1 SYSTEM OVERVIEW

In our system, the analysis of the singing voice includes first a note segmentation which consists on aligning the singing performance to a reference midi and then an expression segmentation which is basically an expression transcription of the performance, segmenting each note in sub-regions (attack, release, sustain, vibrato or transition) and assigning an expressive label to each region. All these processes are based on a set of descriptors and features extracted from the input audio.

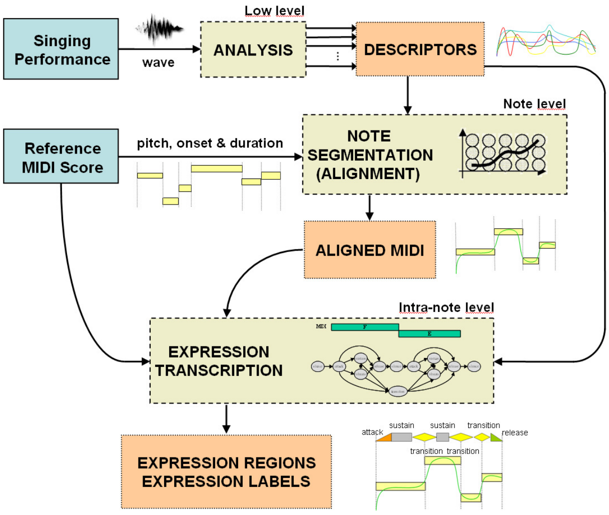


Figure 1: Overview of the performance analysis system

In figure 1 we can see an overview of the performance analysis of the singing voice that is being performed by our system. Firstly we decide which features are more relevant in the singing voice and then we try to derive a set of heuristic rules, based on the analysis descriptors (pitch, energy, spectral coefficients, mel cepstrum coefficients and its derivatives) that can uniquely identify each expression. This set of heuristic rules constitutes the base of a hypothetic probabilistic model based on Hidden Markov Models (HMM). Later on, the performance is automatically segmented into notes and expression regions based on this hypothetic model with no training process [7] involved at all.

2 ANALYSIS DESCRIPTORS

First of all the singing voice is analyzed (see figure 2) and some descriptors in time domain are extracted including zero crossing, amplitude and energy and its derivative. Then a frequency domain analysis is performed and some spectral descriptors are computed like LF energy, HF energy, filter bank (40 coefficients), Mel Cepstrum (24 coefficients), spectral flatness, and delta timbre calculated from the average of the derivatives of the Mel Cepstrum coefficients. Mel cepstrum derivative is very relevant to detect timbre changes in the singing performance since in singing voice, note onsets commonly match changes in phonetics. After the frequency domain analysis some spectral algorithms are applied to extract high level descriptors from the performance including a spectral peak and pitch detection algorithm, vibrato detection based on the pitch analysis and some harmonic descriptors are computed like stability and sinusoidality. These descriptors will be the base to establish a criterion to segment into notes and detect expression from the performance.

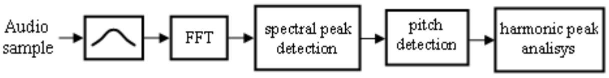


Figure 2: Analysis process

One of the first steps we need to do before the automatic analysis of the singing performance is to manually analyze and categorize different expressive aspects or expressive executions in singing performances. In most cases the expression can be categorized using only pitch and amplitude evolution along the time, but if we want to achieve better results to improve current existing systems we need to take into account more analysis descriptors. We would need to label different expressive resources and extract the descriptors that better describe the performance and distinguish the most, one performance from another. This set of descriptors will be used in the real-time performance analysis of the singing voice.

3 NOTE ALIGNMENT

We are performing a note segmentation with prior knowledge of the reference midi melody the user is supposed to be singing, so we are aligning the midi notes to the notes in the singing performance. As a result of the segmentation we will have the same notes of the midi reference but the onset and duration will be adjusted or aligned to the performance of the user and silences between notes will kept. Thus, if the performance adds or drops notes that are not present in the midi reference, the alignment will try to be as many similar as it can be to the reference but without adding or dropping notes. In figure 3 we can see the results of a note alignment where three notes in the original score (MIDI notes) are aligned to the user pitch (User notes).

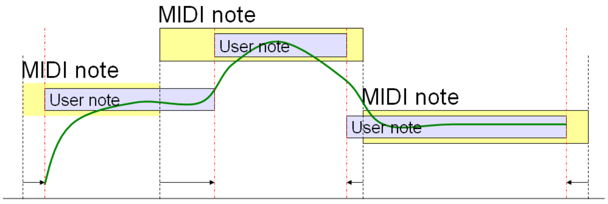


Figure 3: Note alignment results

Based on the analysis data observation, we decide which features are more relevant in the singing voice and then we derive the set of heuristic rules that best identify uniquely each expression. This set of rules is used as a probabilistic model in the note segmentation and expression transcription algorithms to automatically segment the performance into notes and expression regions.

Note alignment is performed using segmental HMMs based on hypothetic probabilistic models. A sequence of note and silence states given by a MIDI score represents the melody of the song (see figure 4) and heuristic rules determine the most probable path from all possible paths in the Viterbi matrix. The resulting score is the same as the reference MIDI but with the notes shortened or lengthened and the onsets and pitch shifted to better fulfill the rules applied in the segmentation/alignment algorithm. You can see this in detail in [6].

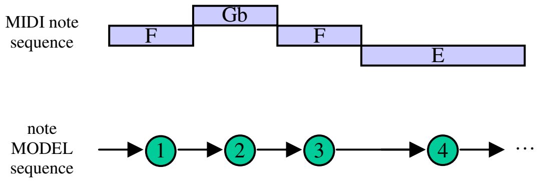


Figure 4: MIDI note and note model sequence.

The probability associated to each node computes the best path to reach that node, starting from the beginning of the song. We distinguish two types of probability:

transition probability and cost probability. The transition probability in our case is always 1, therefore the path probability only depends on the cost probability. The cost probability is computed using heuristic rules which observe the voice descriptors. Given an observation window (from start to end frame) and a given note model, a set of rules are applied to the voice descriptors and each rule computes a rule probability. The cost probability is then the multiplication of all these rule probabilities.

4 EXPRESSION CATEGORIZATION

Expression categorization and transcription of the performance is carried out as well using segmental HMMs based on hypothetic probabilistic models. Expression paths are modeled as sequences of attack, sustain, vibrato, release, transition states and their possible connections. Besides, different labels can be assigned to each state to distinguish between different ways of performing. For instance, in case of a transition, some possible labels include scoop-up, portamento or normal. These paths are considered by the expression recognition module and the path with highest probability among all is the one chosen. The probabilities are based in heuristic rules based on the analysis descriptors. You can see this in detail in [6].

In figure 5 all the possible expression paths for a two notes segment are shown and the best expression path is highlighted. These paths are evaluated by the expression recognition module to pick among them the highest probability one. Cost probabilities and transition probabilities are calculated in the same way that for the note alignment process.

In order to build the expression model sequence while performing in real-time, first we need to complete the alignment of the last performed note. Thus, in real time context, the performance analysis has a one note latency before it can show any feedback to the user.

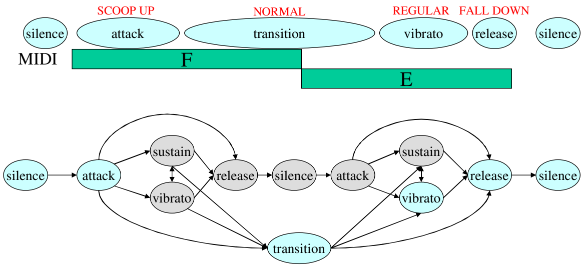


Figure 5: Expression path with expression labels

Once the best expression type path has been chosen, the most probable label for each expression type has then to be estimated.

5 PERFORMANCE RATING/SCORING

After doing the analysis of the user performance, we compute:

- *Elemental ratings* based in high level descriptors like pitch and volume.
- *Timing ratings* based on note segmentation and alignment.
- *Expression rating* based on the expressive transcription and categorization explained in previous sections.

Results of the performance analysis are given as a score from 0 to 100 including fundamental performance ratings and the expression performance rating (see figure 6).

The fundamental performance ratings are calculated comparing the pitch, volume and timing between the user and the reference singer and also between the user and the midi notes. This is done in order to give two separate ratings, one for mimicry and other for comparison with a standard execution.

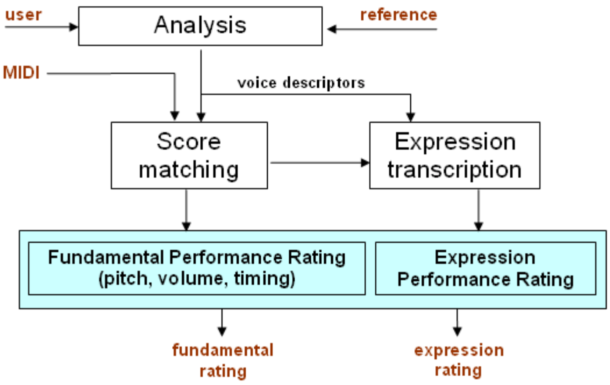


Figure 6: Performance rating overview

We are not giving much relevance to this topic as our application is more focused in giving constant feedback to the user about the performance rather than giving a global accumulated score, which does not help the singer to improve the performance. We have chosen instead to give more relevance to the display of the segments where the singer performs worse; giving the chance to repeat those parts and improve singing skills.

6 APPLICATIONS

The technology that we have developed can be applied in many fields from entertainment and games to more education focused applications.

6.1 Singing Education

In the classical singing education, the master-apprentice model is used where teacher gives instructions and feedback on the performance to the student about:

- Acoustic quality
- Physiological aspect of the performance (posture of the vocal apparatus)
- Tunning
- Timing

The way that the teacher gives feedback to the student uses imagery (ex: “sing as if through the top of your head”) and this yields into a problem of ambiguous interpretation and a big time lag between student’s performance and teacher’s feedback.

There are some existing systems for computer singing education that offer real-time visual feedback trying to solve the problem of classical singing education stated above. These systems include pitch trace, spectrogram, larynx parameters, etc. Some examples of these systems are: SINGAD (SINGing Assessment and Development) [8], WinSINGAD [9], ALBERT (Acoustic and Laryngeal Biofeedback Enhancement in Real Time) [10] [11] and SING & SEE [12].

Our system can also be used to solve the lack of visual feedback problem offering pitch, visual note segmentation and expression detection in real-time. Moreover our system walks a step further allowing to compare the user performance with a performance of a professional, for instance the performance of the teacher. These features convert it into a powerful tool for singing education. The visual feedback that our system gives to the user is explained with details in section 7.1.

6.2 Entertainment

Singing voice automatic scoring has become quite popular in the past few years in games like *Singstar* [13], *Ultrastar* [14], *Karaoke Revolution* [15], Lips [16] and Rock Band [17]. However, the algorithms applied in these videogame applications are rude and far too distant from current voice analysis research in the scientific community. With our system we offer a more complex analysis of the singing voice, not only focused in pitch and timing characteristics but also detecting expressivity in the performance, which can be perfectly applied to video games. In section 7.1 we enumerate some musical video games that incorporate karaoke style gaming and we compare them with our system.

7 SOFTWARE IMPLEMENTATION

The software that has been developed includes an offline tool for research development and a real-time scoring application for user singing evaluation.

The off-line tool can be used for manual segmentation and expressive label edition of notes and intranote regions. This tool gives feedback by showing the probability of each heuristic rule in the manual segmentation given by the user. With the tool, the user can change the segmentation and see whether the global probability improves or not. This tool is also used to display the descriptors calculated in the analysis process so the user can view the values of these descriptors at any time point and change the heuristic rules to and improve the automatic note segmentation and expression transcription results. In the bottom window in figure 7 we can see the display of values of some analysis descriptors along time, each descriptor with a different color. Above this window we can see the pitch curve of the performance, the results of the note-segmentation and expression transcription as well as the MIDI score.

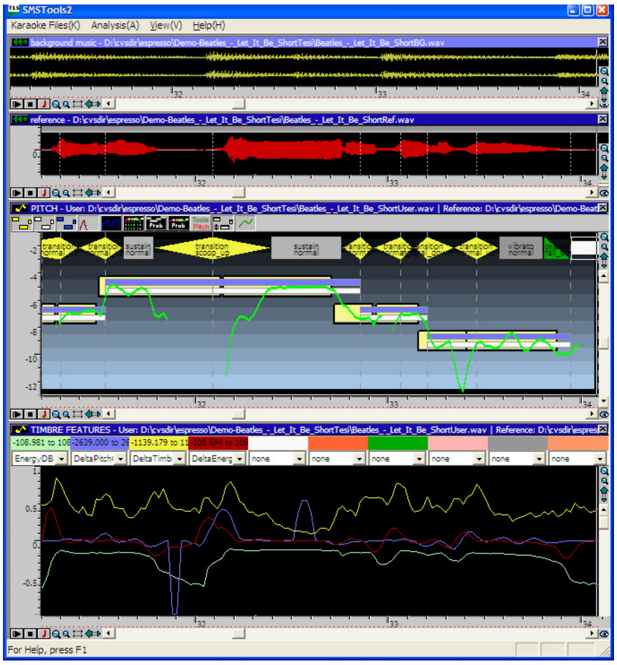

SMSTools2
Karaoke Files(K) Analysis(A) View(V) Help(H)
background music - D:\cvsdir\espresso\Demo-Beatles_-_Let_It_Be_ShortTesi\Beatles_-_Let_It_Be_ShortBG.wav
32
33
34
reference - D:\cvsdir\espresso\Demo-Beatles_-_Let_It_Be_ShortTesi\Beatles_-_Let_It_Be_ShortRef.wav
0
PITCH - User: D:\cvsdir\espresso\Demo-Beatles_-_Let_It_Be_ShortTesi\Beatles_-_Let_It_Be_ShortUser.wav | Reference: D:\cvsdir\espresso\Demo-Beat
Prob
Prob
Tools
Pitch
transition normal
transition normal
sustain normal
transition scoop_up
sustain normal
transition normal
transition normal
transition
transition normal
vibrato normal
-2
-4
-6
-8
-10
-12
TIMBRE FEATURES - User: D:\cvsdir\espresso\Demo-Beatles_-_Let_It_Be_ShortTesi\Beatles_-_Let_It_Be_ShortUser.wav | Reference: D:\cvsdir\espres
-108.981 to 108
-2639.000 to 26
-1139.179 to 11
-108.694 to 108
EnergyDB
DeltaPitch
DeltaTimb
DeltaEnerg
none
none
none
none
none
none
1
0.5
0
-0.5
For Help, press F1

Figure 7: Offline GUI tool

The real-time tool allows the display of some analysis descriptors like pitch and note and expression transcription as well as the reference song’s notes and lyrics to guide the performance while the user sings. The display proposal is based on karaoke-style-games with significant additional information about the performance is shown to the user including some partial and global scoring. The real-time scoring tool is explained with more detail in section 7.1 together with a comparison with other real-time scoring games.

7.1 VISUAL PERFORMANCE FEEDBACK

One of the big drawbacks in karaoke games and commercial karaoke video systems is the lack of visual feedback that users get about the performance. Many systems just show the lyrics and only in some cases the user gets vague information about the note pitch and duration of the notes that has to perform. While performing the song, these notes get highlighted when you sing them in tune (see figure 8).

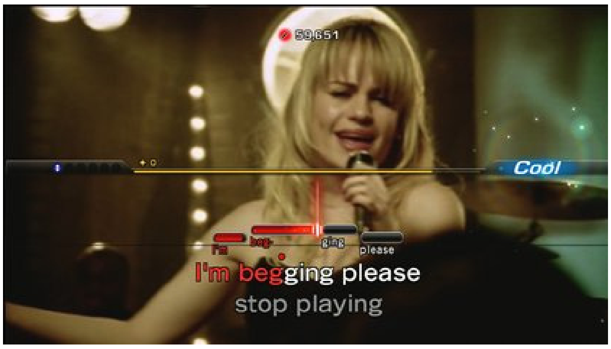


Figure 8: Lips game

This information is not enough if you want to use the karaoke as a virtual singing teacher, and also, when we sing out of tune we don't have enough feedback to determine how far we are from the correct pitch or tempo of each of the notes. Some systems also give information, when the user sings out of tune, about the note performed, so the user is able to know if is singing below or above the target desired pitch (see figure 9).

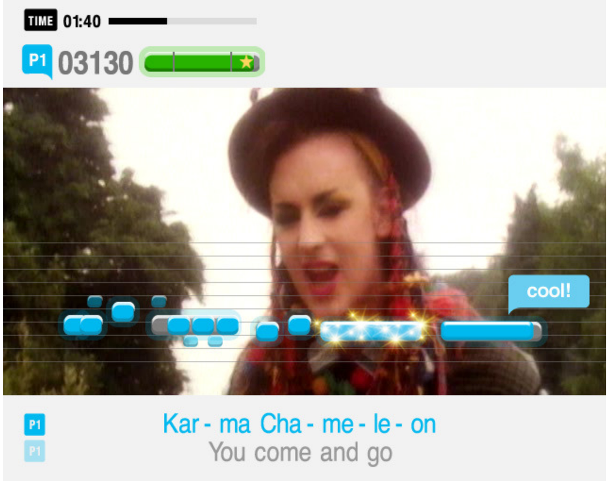


Figure 9: Singstar game in solo mode

In such kind of systems, only information about the song phrase that the user is singing and all the reference notes of the current phrase are shown in the screen and are replaced by new notes when the new phrase comes in. Other systems adopt a scrolling representation of melody, where notes flow from right to left of the screen as in a platform game, while the user is singing a song. In figure 10 you can see an example of scrolling karaoke game. In this example the user only gets feedback about the pitch of the performance at the current time and past information is lost as the notes are sung so the user gets limited feedback about out of tune errors.

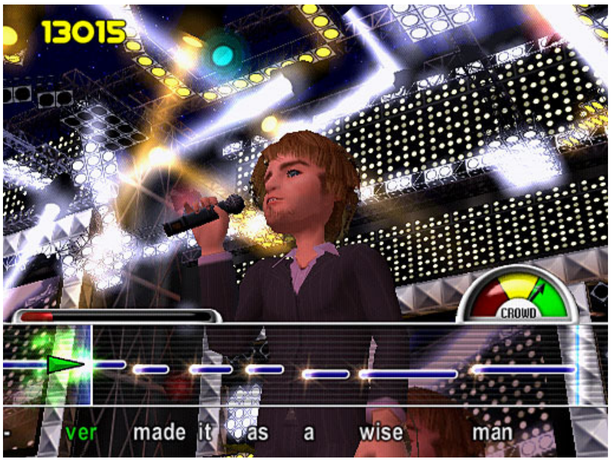


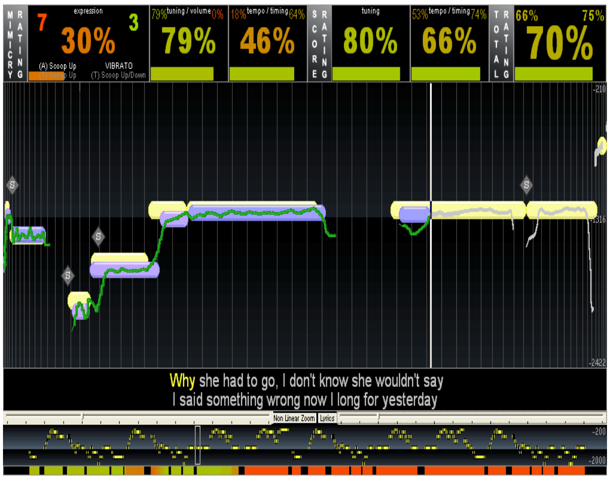
Figure 10: Karaoke Revolution game

Our singing scoring tool adopts a hybrid method by adding more visual feedback for the user; at the same time it gives the possibility to replay the performance to view the parts of the song where the user has mistaken the most. In figure 11 we can see the visual feedback that offers our system, in the lower part a global view of the song is given which will allow the user to review the performance after singing and see the errors committed in any part of the song. These parts will be marked with red color so the user can visualize the conflicting parts of the song and quickly go to them. We also allow repeating certain parts of the song to improve results.

In the middle part of the screen the lyrics are shown and, above them, the midi notes of the score and the transcribed notes from the performance of the user are represented in a scroll from right to left while performing the song. Also the fundamental pitch of a reference singer and the user are shown in real-time, so while performing you can see how close you are to the reference, not only at a note level but also with more detail at a frame level, comparing your pitch with the reference one. Visualizing the pitch also allows you to improve vibrato executions and other expressive aspects where the pitch shape evolution is fundamental like scoop-ups attacks, fall-down releases and different kind of transitions. This also helps the user to be in tune as the desired scenario for pitch is to be drawn over the midi target notes, when this happens the user is in tune and the more distant the pitch is from the midi notes, the more out of tune the user is. This visual feedback allows the user to rapidly correct the tuning while singing.

While the user sings, some detected expressive resources performed by the reference singer are marked on the screen with different signs. The user has to imitate them to score high in the expressive rating. If the user performs well regarding expression, these marks get highlighted. If the user performs new expressive resources not performed by the reference singer, user marks appear in the screen.

In the above part of the screen, the scoring of the song is shown divided in expressive rating, mimicry rating, score rating and total rating, which is an average of the previous.



Why she had to go, I don't know she wouldn't say
I said something wrong now I long for yesterday

Figure 11: Singing tutor application

8 EVALUATION

Five commercial pop songs have been used to evaluate the system and some amateur singers have been asked to sing the songs. The recorded performances have been analyzed and note segmentation and expression transcription have been performed. From these analysis results, more than 1500 notes have been evaluated achieving more than 95% accuracy in the note segmentation, using as reference manual segmentation by a musician and allowing a tolerance of 30 milliseconds, so boundaries automatically segmented within this margin are considered as correct. For the expression transcription evaluation, there is no a simple way to evaluate the results, as sometimes there are many ways to correctly transcript the same performance, for instance putting a very-short sustain or not when the stable part of a note is about 50ms, or putting a long sustain fall-down and a short normal release instead of a short sustain and longer release fall-down. For this reason, the expression transcription evaluation has been done manually by an expert. More than 2500 expression sub-regions have been manually checked and more than 95% of the correctly note-segmented sub-regions were considered to be correctly transcribed by the expert.

9 CONCLUSIONS

This research can either be applied to education, to entertainment or to the so called edutainment (something in between the previous two). In the education context, the impact will be significant as this research can be applied to music schools in order to give an evaluation tool to analyze the performance of the user, not only giving information about tuning and tempo but also about expression and compare it to reference performances. In the entertainment arena, this research can be applied to gaming, the most obvious application is karaoke but there are many other musical applications.

10 ACKNOWLEDGEMENTS

The work described in this paper has been supported and funded by the Music Technology Group in the Pompeu Fabra University and by Yamaha Corp.

REFERENCES

[1] Viitaniemi, T., Klapuri, A. & Eronen, A. “A probabilistic model for the transcription of single-voice melodies”. Proceedings of the 2003 Finnish Signal Processing Symposium, FINSIG'03, Tampere, Finland, 2003.

[2] Ryynänen, M. P. & Klapuri, A. P. “Modelling of note events for singing transcription. In Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing”, Jeju, Korea, 2004.

[3] Cano, P., Loscos, A. & Bonada, J. ”Score-Performance matching using HMMs” Proceedings of the International Computer Music Conference, Beijing, China, 1999.

[4] Dannenberg, Roger B. & Ning Hu. ”Polyphonic Audio Matching for Score Following and Intelligent Audio Editors”. Proceedings of the 2003 International Computer Music Conference San Francisco, USA, 2003.

[5] Sundberg, J. “Expression in music: Comparing vocal and instrumental performance”. Karevold, H Jörgensen, IM Hanken, E Nesheim, eds, Flerstemmige innspill 2000, NMH-publikasjoner 2000:1, Norges Musikkhögskole, Oslo, 2001.

[6] Mayor, O., Bonada, J., Loscos, A. "The singing tutor: expression categorization and segmentation of the singing voice”. 121st AES Convention, San Francisco, USA, 2006.

[7] Cano, P., Loscos, A. & Bonada, J. “Score-Performance matching using HMMs” Proceedings of the International Computer Music Conference, Beijing, China, 1999.

[8] Welch, G. F. “The assessment of singing” Psychology of Music, 22, 3-19, 1994.

[9] Howard, D.M., Williams, J, Brereton, J., Welch, G.F., Himonides, E., Howard, A., and DeCosta, M, “A mirror for sound: Introduction and practical exploration of the WinSingad software for teaching singing”, 3rd International Conference on the Physiology and Acoustics of Singing, PAS3, 33, York, 1-905351-04-6, May, 2006.

[10] Howard, D. M. “Human Hearing Modelling Real-Time Spectrography for Visual Feedback in Singing Training”, Folia Phoniatrica et Logopaedica, 57(5/6):328-341, 1021-7762, September, 2005.

[11] Rossiter, D. & Howard, D.M. “ALBERT: A real-time visual feedback computer tool for professional vocal development”, Journal of Voice, 10(4):321-336, 1996.

[12] Pat H. Wilson, Kerrie Lee, Jean Callaghan, C. William Thorpe. “Learning to sing in tune: Does real-time visual feedback help?”, Journal of interdisciplinary music studies, volume 2, issue 1&2, art. #0821210, pp. 157-172, 2008.

[13] SingStar: karaoke game for Sony’s Playstation 2 & 3. It allows you to evaluate how good you are when you sing by analyzing your voice pitch. http://www.singstargame.com/

[14] Ultrastar: open source PC conversion of famous karaoke game Singstar. Written in Delphi/Kylix. http://sourceforge.net/projects/ultrastar/

[15] Karaoke Revolution: Karaoke game developed by Harmonix Music Systems and Blitz Games for many consoles. The scoring system is based on pitch and rhythm and the game includes a crowd meter to rate the performance.

[16] Lips: karaoke game for the Xbox 360 developed by iNiS and published by Microsoft Game Studios. The game will feature the use of a motion sensitive microphone, and supports the use of songs already owned by the user.

[17] Rock Band: a series of music video games developed by Harmonix Music Systems and MTV Games that allows for up to four players to virtually perform rock music songs on lead guitar, bass guitar, drums, and vocals. http://www.rockband.com