INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Evaluation of singing synthesis: methodology and case study with concatenative and performative systems

Lionel Feugère[1], *Christophe d'Alessandro*[1], *Samuel Delalez*[1], *Luc Ardaillon* [2], *Axel Roebel*[2]

[1]LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
[2]IRCAM, CNRS, Sorbonne Universités UPMC, 75004 Paris, France

lionel.feugere,cda,samuel.delalez@limsi.fr, luc.ardaillon,axel.roebel@ircam.fr

Abstract

The special session Singing Synthesis Challenge: Fill-In the Gap aims at comparative evaluation of singing synthesis systems. The task is to synthesize a new couplet for two popular songs. This paper address the methodology needed for quality assessment of singing synthesis systems and reports on a case study using 2 systems with a total of 6 different configurations. The two synthesis systems are: a concatenative Text-to-Chant (TTC) system, including a parametric representation of the melodic curve; a Singing Instrument (SI), allowing for real-time interpretation of utterances made of flat-pitch natural voice or diphone concatenated voice. Absolute Category Rating (ACR) and Paired Comparison (PC) tests are used. Natural and natural-degraded reference conditions are used for calibration of the ACR test. The MOS obtained using ACR shows that the TTC (resp. the SI) ranks below natural voice but above (resp. in between) degraded conditions. Then singing synthesis quality is judged better than auto-tuned or distorted natural voice in some cases. PC results show that: 1/ signal processing is an important quality issue, making the difference between systems; 2/ diphone concatenation degrades the quality compared to flat-pitch natural voice; 3/ Automatic melodic modelling is preferred to gestural control for off-line synthesis.

Index Terms: singing synthesis, singing quality assessment, computer music

1. Introduction

The special session Singing Synthesis Challenge: Fill-In the Gap is following previous singing synthesis challenges held in 1993 [1] and 2007 [2]. The aim is to gather different research teams working on singing synthesis, using common material for comparing approaches, methods and results. This year, the proposed challenge is to fill-in the gap in well-known songs, i.e., to synthesize a new, especially written couplet including new lyrics, to be inserted in the song. It is anticipated that both Text-to-Chant (TTC) systems and Singing Instruments (SI) will take part to the challenge.

In TTC, the singing voice signal is computed from a symbolic description of the song: a text for lyrics and a musical score [3]. TTC appeared first in experimental studio works, thanks to the "Chant" program [4]. "Chant" is based on a formant synthesizer and synthesis by rules. The following generation of voice synthesis systems was based on recording, concatenation and modification of real speech samples. A remarkably successful TTC system is Yamaha's Vocaloid [5].

Singing instruments, or performative singing synthesis systems allow for real-time, possibly on stage, synthetic singing production. The performer interprets the musical score, playing with some sort of prepared singing material. Following the development of new interfaces for human-computer interaction, SI have recently been issued by different research groups, including parametric, concatenative and statistical synthesis methodologies [6, 7, 8, 9, 10, 11, 3].

The preceding singing synthesis challenges have been rather informal as far as evaluation is concerned: a post-session participant voting procedure was used rather than controlled listening tests. It seems important to propose more formal methods for assessing the quality obtained with the current systems and for establishing the baseline quality for future systems. In the present paper, the question of formal singing synthesis assessment methodology is addressed along with a case study using two systems and a total of 6 system versions. The paper is organized as follows. In the next section, the singing assessment methodology is proposed. In section 3, the different TTC and IS systems tested are described. Section 4 presents perception tests and the results obtained. Section 5 concludes.

2. Singing synthesis assessment methodology

Subjective testing is the most appropriate methodology for assessment of singing synthesis quality. Quality evaluation is a multidimensional task, encompassing sound quality (signal concatenation, signal modelling), and expressivity (interpretation rules, voice quality, performative control). Both global and analytic evaluation methodologies are needed.

2.1. Absolute Category Rating

Absolute Category Rating (ACR) is the most obvious method for subjective quality assessment of synthetic singing. It is designed for evaluation and comparison of the quality of systems by listening to the systems output separately. The comparison between systems is therefore indirect. This gives a global evaluation of the output, without taking in consideration the system's internal functioning and without trying to understand the source of its defects. Subjects listen once to each stimuli and are asked to report a Mean Opinion Score (MOS) on a 5-points scale.

2.2. ACR test calibration: reference conditions

The ACR test is calibrated by using common references. This allows for comparison of the different systems on a common basis, and the repeatability of the test in the future, for measuring the progress. References are made of natural speech, either in clean form ("top condition") or in intentionally degraded form. Three degraded natural speech conditions (DC) are obtained from natural speech. They can be downloaded from the

http://dx.doi.org/10.21437/Interspeech.2016-1248

URL given in the last section.

DC1 Pitch degradation was done with the Antares Autotune-Evo vst pluggin, providing unnaturally hard-tuned stimuli. The parameters *return speed*, *humanize* and *natural vibrato* were all set to 0. DC1 is a middle quality condition.

DC2 Ableton Live's Overdrive effect was used to degrade the voice spectrum. *Filter freq* was set to 1kHz, *Filter width* to 9, *Drive* to 60 %. Other parameters were left to built-in preset values. DC2 is a middle quality condition.

DC3 Temporally degraded stimuli were made with Ableton Live's time stretching tools. Natural voices were warped with option *Beats*. Original signals were stretched to obtain twice longer modified signals. These signals have been consolidated (an Ableton Live's option that saves a signal as it is after modification), and their durations have been divided by 2. The degraded stimuli have the same duration as natural ones, but with a degraded phoneme quality. DC3 is bottom quality condition.

2.3. Paired Comparisons

Paired Comparisons (PC) involve a simple choice: two stimuli A and B are presented, and the subjects must express their preference for stimuli A or B. The attention of subjects is directed to specific features, both by explicit instructions and by presentation of selected short utterances focusing on these features. The features studied here are the quality of articulation (consonantal transitions) and the quality of melodic ornamentation (pitch vibrato and pitch transition between notes).

2.4. Singing material

The "fill-in the gap" task consists of the singing voice synthesis for a selected karaoke version of the two famous XXth century songs: *Summertime* music by George Gershwin (1934), *Autumn Leaves* music by Joseph Kosma (originally *les feuilles mortes* (1946)). Original lyrics (in English and French) were written for the singing synthesis challenge (the French lyrics are used herein). These data are publicly available [12]. Two singers (a female soprano and a male tenor) recorded the two songs *InterspeechTime* (117 beats per second, swing) and *Interspeech-Leaves* (142 bps, swing). They also recorded the lyrics on a same note (flat pitch) and with regularly-timed syllables (regular rhythm). This is useful for testing concatenation quality.

2.5. Dimensions tested

Several features of the systems are evaluated, with the help of ACR and PC.

Concatenation The segmental basis of the signal is built by diphones concatenation (Con-), or is the natural signal recorded with flat pitch and regular rhythm (monocord-isochron: Mi-)

Melodic modeling : offline automatic parametric modeling of pitch and durations is applied to Con- and Mi- signals.

Gestural control Gestural control of melody and rhythm is applied to Con- and Mi-.

Time and frequency scaling algorithms Three time and frequency scaling algorithms are tested: PAN, SVP for the automatic TTC system and RT-PSOLA for the Caliphony system Cal. Note that PAN was used to create the monocord-isochron file needed to perform Con-cal.

This results in 6 systems (Mi-PAN, Mi-SVP, Con-PAN, Con-SVP, Mi-Cal, Con-Cal) and 4 control conditions (Nat, DC1, DC2, DC3), i.e. 10 conditions for each feature tested. Note that the gesture-controlled synthesis systems (Mi-cal, Con-cal), as well as the natural voices were singing from the score, while the TTC system computed the signal from a score file corresponding to the notes and the lyrics.

3. Singing synthesis systems

3.1. Concatenative synthesis system

The synthesis system used in this work is an extension of the one presented in [13]. It is based on diphone concatenation, and is composed of: a control module, in charge of generating the control parameters from the input text and MIDI score; a unit selection module, which selects the units to be concatenated from a database; and a synthesis engine, in charge of the concatenation and transformations processes, based on the selected units and the generated control parameters. Those modules are organized in a modular way, so that it is possible to integrate different methods for each module. In this work, 2 different synthesis engines, SVP and PAN have been assessed.

3.1.1. Databases

In order to synthesize any possible lyrics, the minimum requirement for our systems database is to cover all the diphones (about 1200 for French). A set of 900 words has been chosen for ensuring this coverage. Those words are sung on a single pitch with constant intensity. The database is segmented in both phonemes and diphones, where the diphones boundaries lie in the stable part of each phonemes. Those segmentations are used during the synthesis to select from the database the units to be concatenated and compute the required time stretching factors. Two databases have been used in the presented work. The 1st one is a tenor male singer, and the 2nd one is a female soprano. Both databases have been recorded with a pop-like voice timbre, with few vibrato.

3.1.2. SVP

The SVP synthesis engine is based on superVP [14, 15], an advanced phase vocoder, using shape-invariant processing [16]. This engine processes the units in the time-frequency domain for transposition and time-stretching, and some phase and envelope interpolation is done at the junctions between the selected units in order to avoid discontinuities, as explained in [13].

3.1.3. PAN

The PAN synthesis engine is based on an enhanced version of the SVLN analysis/synthesis method [17]. Improvements are on one hand the refined and extended glottal pulse estimation method described in [18] and on the other hand a new approach to extract and synthesize the unvoiced signal component [19].

3.1.4. Control module

The control module generates the target pitch (F_0) curve and phonemes durations from the input text and score. Other parameters, such as intensity, have not been modeled in this work. The F_0 curve generation is based on the approach presented in [13], where the expressive fluctuations of the F_0 (such as vibrato, overshoot, preparations, ...) are modeled with B-splines using an intuitive parametrisation. The curve is temporally segmented in basic units (attack, sustain, transition, and release), each having its own set of parameters. Those parameters are extracted from recordings of real singers, along with the contexts associated with the score of the recording, to form a database of parametric templates. At synthesis stage, parametric templates are selected in this database, for each F_0 segment, using decision trees, according to the target contexts of the score to be synthesized [20]. A similar procedure is used to choose the phonemes durations.

3.2. Singing instrument: The Calliphony system

The *Calliphony* system allows performative time and pitch scale modifications of pre-recorded voice. Pitch is controlled manually with a stylus on a Wacom graphic tablet and rhythm is controlled with an expression foot pedal. It has been programmed in the Max environment [21]. A real time version of the TD-PSOLA algorithm [22] (RT-PSOLA [23]) has been implemented in Java and integrated into Max/MSP. Period markers obtained with Praat were used.

3.2.1. Pitch control

Pitch of a pre-recorded voice signal is modified with the position of the stylus in the *x axis* of the tablet. The user can visually target notes on the tablet thanks to a so called *tablet mask* installed on the tablet. The same pitch control strategy is used in the Cantor Digitalis [24].

3.2.2. Rhythm Control

Rhythm of the original signal is modified with an Eowave usb expression pedal. The pedal has two extreme positions: upper and lower positions. The user points a syllable vocalic part by placing the pedal in any extreme position. Vowel-Consonant-Vowel transitions are performed by moving the pedal from one extreme position to another. Thus, consonants are pointed around the central position of the pedal, in order to allow fast rhythm control and to prevent foot movements with too large amplitude.

4. Evaluation tests

25 subjects were hired to participate to a listening test in an isolated room. All of them are either musician or have an activity related to sound listening (a mean current practice of 6 hours a week). None of them reported any hearing issue and they were not working on the current project. They were paid for the experiments.

A computer interface was especially designed for this study. Subjects were asked to listen a short excerpt (or a pair of short excerpts) of singing synthesis and to score (or give a preference) for each of the excerpt (pair of excerpts). Listening can be repeated with a play button. A button allowed to validate the choice and to go to the next stimuli. A training session, featuring examples of all the conditions for both singers, was offered prior to recording the results.

4.1. Experiment 1: ACR

4.1.1. Protocol

For the first experiment, InterspechTime is split in 4 excerpts of 4-bars and InterspeechLeaves is split in 8 excerpts of 4-bars but only the 4 first excerpts are used. The first experiment is an ACR with the following question: "Globally, how did you appreciate the quality of what you have just heard?" (in french in the experiment: "*Globalement, comment appréciez-vous la qualité de ce que vous venez d'entendre ?*"). The possible score ranges are: bad (1), poor (2), fair (3), good (4), excellent (5). The original terms used in the experiment are: *médiocre* (1), *faible* (2), *moyenne* (3), *bonne* (4), *excellente* (5).

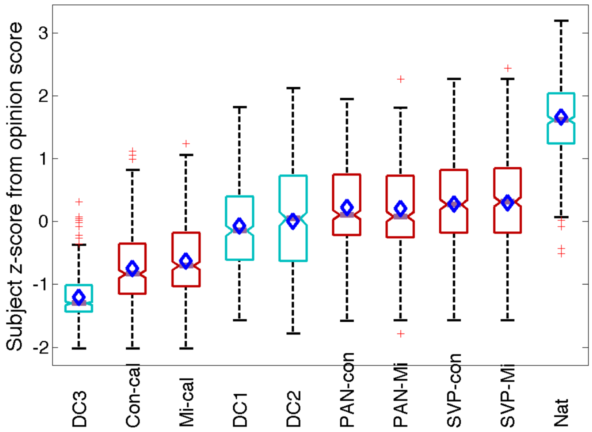

Figure 1: *Z-scores computed from subject's opinion scores. Diamond represents the z-score mean.*

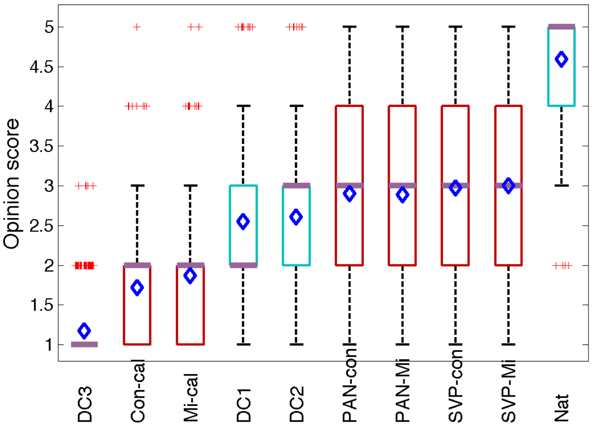

Figure 2: *Opinion score distributions. Diamonds are the MOS.*

4.1.2. Results

MOS and associated standard deviation are given in table 1 for each system. A z-score computation on each subject was done in order to normalize the mean and dispersion of the results.

Dispersion of the opinion score in term of z-score is displayed in Figure 1 for each system. Statistical significance is studied using a Tukey's honestly significant difference criterion from the Matlab *multcompare* function. As expected, the two extreme conditions DC3 (MOS=1.2) and natural speech (MOS=4.6) are significantly different from the other conditions ($p < 10^{-6}$). The 8 other conditions are distributed in four groups. The first group is made of the TTC systems, with a MOS between 2.9 and 3.0. This groups is homogeneous, with no significant differences between conditions. The second group is made of the control conditions DC1 and DC2, with a MOS between 2.5 and 2.6, without significant differences be-

Table 1: *Experiment 1. MOS (on a 1-5 scale) and standard deviation for each system.*

	DC3	Con-cal	Mi-cal	DC1	DC2	PAN-con	PAN-Mi	SVP-con	SVP-Mi	Nat
MOS	1.2	1.7	1.9	2.5	2.6	2.9	2.9	3.0	3.0	4.6
std	1.1	0.8	0.9	1.0	1.1	1.1	1.1	1.1	1.1	0.7

tween the two. The third and fourth groups are the Calliphony systems, with a MOS of 1.7 for the one using concatenation and 1.9 for the one played from speech transformation, and with a small significant difference ($p = 0.04$).

In addition, z-score for all groups are significantly different from z-score of other groups ($p < 0.05$). The ACR test leads to the following conclusions:

Concatenation Surprisingly, there is no difference in MOS between concatenation and flat-pitch regular rhythm recorded speech. This demonstrates the high quality of the concatenation system.

Melodic modeling is also very well scored.

Gestural control of melody and rhythm scored above DC3, but below all other conditions.

Time and frequency scaling algorithms No significant difference is found between PAN and SVP. RT-PSOLA is scoring above DC3, but below all other conditions.

This first test gives a clear picture of the perceived quality for the different systems, but it is difficult to figure out which part of the appreciation concerns the signal quality or the melodic rules quality.

4.2. Experiment 2: PC

4.2.1. Protocol

The second experiment is a PC, split in two parts. The first part deals with quality of lyrics articulation while the second deals with quality of melodic ornamentation (vibrato and portamento). Three short excerpts (a few seconds) were chosen for each dimension. The participant was asked to choose his preferred item in the pair by the following question: "Choose the item for which you appreciate the quality of lyrics articulation the most" (articulation dimension) or "Choose the item for which you appreciate the quality of ornamentation (vibrato, portamento) the most" (in french in the experiment: "*Choisissez l'extrait dont vous avez le plus apprécié la qualité d'articulation des paroles*" or "*Choisissez l'extrait dont vous avez le plus apprécié la qualité d'ornementation (vibrato, portamento)*"). All the terms "articulation", "vibrato", "portamento" were explained before. No training session was needed as all the subjects were already familiar with the voices, owing the first experiment. No control conditions were used for this experiment. Only selected pairs of systems were tested (see Table 2).

4.2.2. Results

Result of the PC test are reported in Table 2. Significances are analyzed using a chi-square test. The results show a good agreement with the ACR test, but it refines the analysis.

Concatenation Transformed natural voice (Mi-) is always preferred to transformed concatenated voice (Con-) for articulation, except if Mi- is associated with Calliphony (-cal).

Melodic modeling is equivalent for the different TTC versions (not depending on signal processing or concatenation).

Table 2: *Experiment 2: Percentage of preference of the column system over the line system, for each pair. A star means that the proportion is significant compared to a 50% proportion in the same conditions (i.e. there is no preference). First line: articulation; second line: melodic ornamentation.*

	SVP-Mi	PAN-Con	PAN-Mi	Con-cal	Mi-cal
SVP-con	68%* 58%*	56% 57%		15%* 29%*	40%* 34%*
SVP-Mi					20%* 28%*
PAN-con			71%* 48%	13%* 31%*	35%* 33%*
PAN-Mi					17%* 37%*
Con-cal					71%* 55%

Gestural control is always outperformed by melodic modeling. However, gestural control of transformed natural voice is close (but significantly different) to TTC concatenation.

Time and frequency scaling algorithms Again, no significant difference is found between PAN and SVP. RT-PSOLA is never preferred.

5. Conclusion

The proposed methodology includes both global and analytic evaluation methods. Degraded conditions are useful for comparing systems, because they introduce anchor points in the ACR procedure. Three types of degradation that are likely to occur in singing synthesis systems have been chosen: pitch degradation, spectral degradation and phoneme degradation. These anchor points give a scale for system evaluation and will be useful for measuring the progress of singing synthesis systems. The PC test is useful for unveiling details otherwise masked in the ACR test.

Application of this methodology to two systems gave a clear picture of their perceptual merits. The TTC system sounded better than all the degraded conditions, although it was clearly different from natural singing. The Si is at this point in time of lesser quality than TTC, probably because of signal processing quality problems. Sound examples corresponding to this paper can be downloaded at `http://groupeaa.limsi.fr/staticaa/chanter/IS16/FeugereDDAR16_sounds.zip` or can be played online at `http://chanter.limsi.fr/doku.php?id=evaluations:start`. Quality assessment must be considered as an important issue in singing synthesis research, and this work is a first step in this direction.

Acknowledgements This research is in the framework of the ANR (Agence Nationale de la Recherche) ChaNTeR project (ANR-13-CORD-011).

6. References

[1] "Session synthesis of singing," in *proceedings of the Stockholm Music Acoustics Conference (SMAC 1993)*, 1993, pp. 279–294.

[2] "Synthesis of singing challenge, special session at interspeech 2007,," in *8th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, 2007.

[3] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Processing Magazine*, vol. 32, no. 55-73, 2015.

[4] X. Rodet, Y. Potard, and J.-B. Barrière, "The CHANT project: From the synthesis of the singing voice to synthesis in general," *Computer Music Journal*, vol. 8, no. 3, pp. 15–31, Autumn 1984.

[5] H. Kenmochi and H. Oshita, "Vocaloid – commercial singing synthesizer based on sample concatenation," in *Interspeech*, 2007.

[6] M. M. Wanderley, J.-P. Viollet, F. Isart, and X. Rodet, "On the choice of transducer technologies for specific musical functions," in *Proc. of the 2000 International Computer Music Conference (ICMC2000)*, 2000, pp. 244–247.

[7] L. Kessous, "Contrôles gestuels bi-manuels de processus sonores," Ph.D. dissertation, Université de Paris VIII, 9 novembre 2004.

[8] M. Zbyszynski, M. Wright, A. Momeni, and D. Cullen, "Ten years of tablet musical interfaces at cnmat," in *Proceedings of the 7th Conference on New Interfaces for Musical Expression (NIME'07)*, New York, USA, 2007, pp. 100–105.

[9] N. D'Alessandro, P. Woodruff, Y. Fabre, T. Dutoit, S. Le Beux, B. Doval, and C. d'Alessandro, "Real time and accurate musical control of expression in singing synthesis," *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 31–39, March 2007.

[10] S. Le Beux, L. Feugère, and C. d'Alessandro, "Chorus digitalis : experiment in chironomic choir singing," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, P. of the conference ISSN: 1990-9772, Ed., Firenze, Italy, 27/08 au 31/08 2011, pp. 2005–2008.

[11] M. Astrinaki, N. D'Alessandro, B. Picart, T. Drugman, and T. Dutoit, "Reactive and continuous control of HMM-based speech synthesis," in *IEEE Workshop on Spoken Language Technology (SLT 2012)*, Miami, Florida, USA, December, 2-5 2012.

[12] "ChaNTeR project," https://chanter.limsi.fr/, 2014.

[13] L. Ardaillon, G. Degottex, and A. Roebel, "A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls," in *INTERSPEECH 2015*, Germany, 2015.

[14] M. Liuni and A. Roebel, "Phase vocoder and beyond," *Musica/Tecnologia*, vol. 7, no. 73-89, 2013, http://www.fupress.net/index.php/mt/article/view/13209.

[15] A. Roebel, "SuperVP software," http://anasynth.ircam.fr/home/english/software/supervp, 2015.

[16] ——, "A shape-invariant phase vocoder for speech transformation," in *Proc. Digital Audio Effects (DAFx)*, 2010.

[17] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2012.

[18] S. Huber and A. Roebel, "On the use of voice descriptors for glottal source shape parameter estimation," *Computer Speech and Language*, vol. 28, no. 5, pp. 1170–1194, 2014.

[19] ——, "Voice quality transformation using an extended source-filter speech model," in *12th Sound and Music Computing Conference (SMC)*, 2015, pp. 69–76.

[20] L. Ardaillon, C. Chabot-Canet, and A. Roebel, "Expressive control of singing voice synthesis using musical contexts and a parametric f0 model," in *submitted for Interspeech 2016 conference*, 2016.

[21] "Max," http://cycling74.com/.

[22] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[23] S. Le Beux, B. Doval, and C. d'Alessandro, "Issues and solutions related to real-time td-psola implementation," in *Audio Engineering Society*, 2010.

[24] L. Feugère and C. d'Alessandro, "Contrôle gestuel de la synthèse vocale. les instruments cantor digitalis et digitartic (gestural control of voice synthesis: the cantor digitalis and digitartic instruments," *Traitement du Signal*, vol. 32, no. 4, pp. 417–442, 2015.