INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

SERAPHIM *Live!* Singing Synthesis for the Performer, the Composer, and the 3D Game Developer

Paul Yaozhu Chan[1], *Minghui Dong*[1], *Grace Xue Hui Ho*[2], *Haizhou Li*[1]

[1]Human Language Technology Department, Institute for Infocomm Research, A*Star, Singapore
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

{ychan, mhdong, hli}@i2r.a-star.edu.sg, gho004@e.ntu.edu.sg

Abstract

The human singing voice is highly expressive instrument capable of producing a variety of complex timbres. Singing synthesis today is popular amongst composers and studio musicians accessing the technology by means of offline sequencing platforms. Only a couple of singing synthesizers are known to be equipped with both the real-time capability and the user interface to successfully target live performances. These are LIMSI's Cantor Digitalis and Yamaha's VOCALOID Keyboard. However, both systems have their own shortcomings. The former is limited to vowels and does not synthesize complete words or syllables. The latter is only real-time to the syllable level and thus requires specifications of the entire syllable before it commences in the performance. A demand remains for a singing synthesis system that truly solves the problem of real-time synthesis - a system capable of synthesizing both vowels and consonants to form entire words while being capable of synthesizing in real-time to the sub-frame level. Such a system has to be versatile enough to exhaustively present all acoustic options possible to the user for maximal control while being intelligent enough to fill in acoustic details that are too fine for human reflexes to control.

SERAPHIM is a real-time singing synthesizer developed in answer to this demand. This paper presents the implementation of SERAPHIM for performing musicians and studio musicians, together with how 3D game developers may use Seraphim to deploy singing in their games.

Index Terms: SERAPHIM; SERAPHIM Live; SERAPHIM Unity package; singing synthesis; real-time TTS; real-time singing synthesis; talking head; lip synchronization

1. Introduction

Music is traditionally a performing art form, with real-time requirements, where the only offline aspects of the art are composition and arrangement. In addition to the introduction of the offline aspects mixing and sound engineering with recording technologies, synthesis introduced the offline aspect of sequencing.

Table 1 lists offline and real-time examples of artificial sound reproduction systems. As shown in the table, most music synthesizers may be used both offline and in real-time[1, 2]. A number of text-to-speech synthesizers work offline, or at best at word level in runtime[3]. A number of early works feature real-time performance of artificial speech[4]. Even though synthesized singing combines music and speech synthesis, there has yet to be a satisfactory answer to real-time singing synthesis.

Table 1: Offline versus Real-time Implementations of Artificial Sound Reproduction

	Offline *Sequenced*	**Real time** *Performed*
Synthesized Music	Theremin[1], Moog[1], MOTIF[2]	
Artificial Speech	TTS[3]	Kempelen[4], Voder[4], Euphonia[4]
Synthesized Singing	VOCALOID[5, 6], UTAU[7], SinSy[8]	*?*

2. Comparison with Existing Work

2.1. Constraints with Coverage and Real-time Capabilities

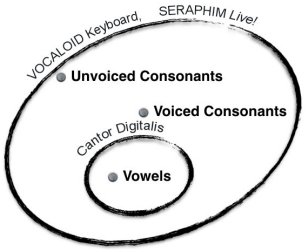

Figure 1: Coverage of Speech Particles by Cantor Digitalis[9], VOCALOID Keyboard[5, 6] and SERAPHIM

The existing real-time singing synthesis systems today are Yamaha's VOCALOID Keyboard[5, 6] and LIMSI's Cantor Digitalis[9]. Figure 1 illustrates the speech particle coverage by the two against SERAPHIM while Figure 2 illustrates the real-time capabilities of the two against SERAPHIM. In Figure 1 it can be seen that Cantor Digitalis does not cover consonants and in Figure 2 it can be seen that VOCALOID Keyboard is only real-time to the syllable level. SERAPHIM is the only system that covers all speech particles while being real-time to the

sub-frame level.

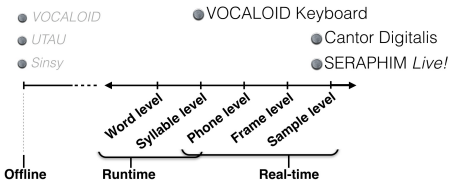

Figure 2: Real-time Capabilities of Singing Synthesis Systems

2.2. Quality of Synthesized Singing Voice

Table 2 lists the mean scores of each utterance in the experiment from [10], where 4 utterances from each existing system[11, 12] were mimicked by SERAPHIM *Live!*, and the total of 16 resultant utterances were randomized and presented to 12 listeners who were tasked to score each utterance for pleasantness of synthesized voice on a scale of 1 to 10, with 1 being worst sounding and 10 being best sounding. The results are normalized and presented as a percentage in the table. According to the results, SERAPHIM *Live!* sounds 26.04% better than Cantor Digitalis and 18.54% better than VOCALOID Keyboard.

Table 2: Subjective Listening Test Results

Utterance Pair	Systems		Utterance Pair	Systems	
	Cantor Digitalis[11]	SERAPHIM *Live!*		VOCALOID Keyboard[12]	SERAPHIM *Live!*
A	20.00%	50.83%	**E**	60.00%	72.50%
B	25.83%	60.83%	**F**	53.33%	78.33%
C	34.17%	55.00%	**G**	64.17%	78.33%
D	36.67%	54.17%	**H**	58.33%	80.83%
Mean	29.17%	55.21%	**Mean**	58.96%	77.50%

3. Deployability

While compiled versions of SERAPHIM Studio and SERAPHIM *Live!* are almost ready for release, SERAPHIM will also be made available in the form of a Unity 3D package, complete with lip synchronization feature, for 3D game developers to deploy singing synthesis in their games.

4. Conclusion

In this show-and-tell paper, we presented how our work in SERAPHIM Live! is an answer to the demand of real-time singing synthesis, how it compares in terms of capability and performance to existing systems, and how it may easily be deployed in 3D games.

5. Demo

We plan to demonstrate the following during the conference:

- **SERAPHIM *Live!* Mandarin** Real-time Mandarin Singing Synthesis System
- **SERAPHIM *Live!* Japanese** Real-time Japanese Singing Synthesis System
- **Lip-Synchronization in SERAPHIM *Live!*** Automatic Real-time Lip-Synchronization with SERAPHIM Singing Synthesis System

A demo 3D video of singing synthesized by SERAPHIM, with automatically lip animation is submitted together with this paper.

6. References

[1] A. Smirnov, "Synthesized Voices of the Revolutionary Utopia: Early Attempts to Synthesize Speaking and Singing Voice," in *Electrified Voices: Medial, Socio-Historical and Cultural Aspects of Voice Transfer*, D. Zakharine and N. Meise, Eds. Vandenhoeck & Ruprecht, 2012, pp. 163–185. [Online]. Available: https://books.google.com.sg/books?hl=en\&lr=\&id=3pWeC9070HcC\&oi=fnd\&pg=PA163\&dq=moog+theremin+motif\&ots=gcDRyFdb-a\&sig=8NphvvnGss-fkMBzffSunk5jkYU\&redir_esc=y\#v=onepage\&q\&f=false

[2] F. Holm, "Believe in Music NAMM Show 2004," *Computer Music Journal*, vol. 28, no. 3, pp. 79–81, 2004. [Online]. Available: http://www.jstor.org/stable/3681511

[3] M. Dong, P. Y. Chan, L. Cen, B. Ma, and H. Li, "I2R text-to-speech system for Blizzard Challenge 2010," in *Blizzard Challenge Workshop*, 2010.

[4] J. O. i. Font, "Musical and phonetic controls in a singing voice synthesizer," Ph.D. dissertation, Polytechnics University of Valencia, 2001. [Online]. Available: files/publications/pfc2001-jortola.pdf

[5] M. Umbert, J. Bonada, and M. Blaauw, "Generating Singing Voice Expression Contours Based On Unit Selection," p. 19, 2009.

[6] H. Kenmochi, "Singing Synthesis as a New Musical Instrument," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 5385–5388.

[7] Ameya Purojekuto. Ameya/Shobu, "UTAU-Synth," http://utau-synth.com, 2008.

[8] K. Oura, A. Mase, Y. Tomohiko, S. Muto, Y. Nankaku, and K. Tokuda, "Recent Development of the HMM-based Singing Voice Synthesis System - Sinsy," in *7th ISCA Workshop on Speech Synthesis*, 2010, pp. 211–216. [Online]. Available: http://20.210-193-52.unknown.qala.com.sg/archive/ssw7/papers/ssw7_211.pdf

[9] L. Feugère, S. L. Beux, and C. D'Alessandro, "Chorus Digitalis Polyphonic Gestural Singing," in *INTERSPEECH 2011*, 2011.

[10] P. Y. Chan, M. Dong, G. X. H. Ho, and H. Li, "SERAPHIM - A Wavetable Synthesis System with 3D Lip Animation for Real-time Speech and Singing Applications on Mobile Platforms," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, San Francisco, 2016.

[11] Audio Acoustique LIMSI CNRS, "Cantor Digitalis - JDEV 2013, Palaiseau - Cantate 2.0," 2013. [Online]. Available: https://www.youtube.com/watch?v=d4TV-IcK8c8

[12] DigInfo TV, "Yamaha Vocaloid Keyboard - Play Miku Songs Live! #DigInfo," 2012. [Online]. Available: https://www.youtube.com/watch?v=d9e87KLMrng