SOUND SYNTHESIS SYSTEM BASED ON SINUSOID SUPERPOSITION WITH APPLICATION TO MULTILINGUAL VOWEL SYNTHESIS

Shuichi ITAHASHI and Hiroaki KOJIMA

National Institute of Advanced Industrial Science and Technology (AIST)

ABSTRACT

Vowel synthesis by superposing harmonics of a fundamental frequency is easy to understand for beginners. This paper introduces a speech synthesis system based on this method utilizing a GUI running on MATLAB. This system can superpose not only sinusoids of constant amplitude, but also those with amplitudes inversely proportional to frequency. It can also output the harmonics by superposing them sequentially starting from some lowest frequency harmonic. Furthermore, the system can synthesize vowel-like sounds by controlling the amplitudes of the harmonics 1 to 5 in the first three formant regions of the vowel. It can also synthesize the stationary vowels of seven languages, as well as a subset of the cardinal vowels. We think that the described system is useful for understanding the fundamentals of vowel synthesis.

***Index Terms*—**GUI, Fundamental frequency, Harmonics, Formant, Spectrum

1. INTRODUCTION

There are several methods for synthesizing speech currently in use, including vocal-tract analogue, spectrum analogue, and speech wave concatenation. Vowel synthesis by superposing harmonics of the fundamental frequency is easy to understand for beginners. In this article, we introduce a speech synthesis system based on this method that utilizes a GUI running on MATLAB. Graphical sound synthesis systems such as those by UCL [1] and KTH [2], exist, but the proposed system has such features that it can superpose not only sinusoids of constant amplitude as the start point, but also sounds that are similar to those produced with vocal chords by superposing sinusoids with amplitudes that are inversely proportional to frequency or the square of frequency. It can also output the harmonics by superposing them sequentially starting from some lowest frequency harmonic. Furthermore, the system can synthesize vowel-like sounds by controlling the amplitudes of the first to fifth harmonics in the first three formant regions of the vowel [3][4]. Vowels of various languages can be synthesized, and the current implementation of the proposed system can synthesize the stationary vowels of seven languages, as well as a subset of the cardinal vowels. We feel that this system will be useful for understanding the fundamentals of vowel synthesis.

2. SPEECH SPECTRUM

In the frequency domain, stationary sounds, such as those of vowels, can be represented by a line spectrum containing multiple harmonics of the fundamental frequency (F_0), as shown in Fig. 1. The system proposed in this paper synthesizes vowels on the basis of this principle.

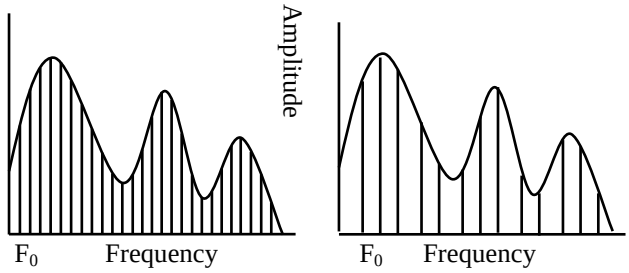

Fig. 1 Harmonics of the fundamental frequency (F_0) component and the spectrum envelope. Left: low F_0. Right: high F_0.

The basic equation of the method is as follows:

$$S(t) = \sum_{k=1}^{K} A_k \sin(k\omega_0 t + \theta_k), \quad (1)$$

where $\omega_0=2\pi F_0$, k denotes the harmonic number, K denotes the highest harmonic number, and A_k indicates the amplitude of the k-th harmonic. A vowel can be synthesized by using suitable A_k corresponding to the spectral envelope of the desired vowel. Here, θ_k represents the phase of the k-th harmonic. The proposed system using eq. (1) runs on MATLAB.

3. SOUND SYNTHESIS BY SUPERPOSING SINUSOIDS

3.1. Constant amplitude, 1/f, and 1/f^2 amplitude spectra

The proposed system is suitable for a sampling rate of 8–16 kHz and a fundamental frequency (F_0) of 100–800 Hz. This range of F_0 corresponds roughly to the fundamental

frequencies of adult males (100–200 Hz), adult females (200–400 Hz), and children (400–800 Hz).

The initial harmonic frequency and the number of harmonics of speech are variable. For example, we can start the harmonic components from 200 Hz or higher, or we can reduce the number of harmonic components, although the total number of harmonic components of F_0 = 100 Hz with a 10 kHz sampling rate is 50. We can choose a constant, 1/f (inversely proportional to the frequency), or $1/f^2$ (inversely proportional to the square of the frequency) amplitude. Figure 2 shows a constant amplitude spectrum and Fig. 3 shows the corresponding speech wave. The speech wave is close to a (two-sided) pulse to the degree that the spectrum amplitude is flat.

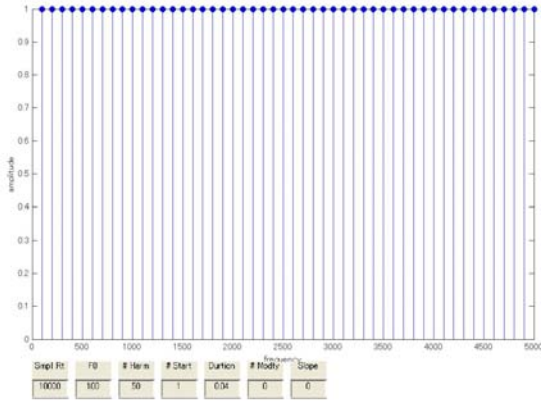

Fig. 2 Spectrum of sinusoidal harmonic superposition: constant amplitude, 10 kHz sampling, F_0 = 100 Hz.

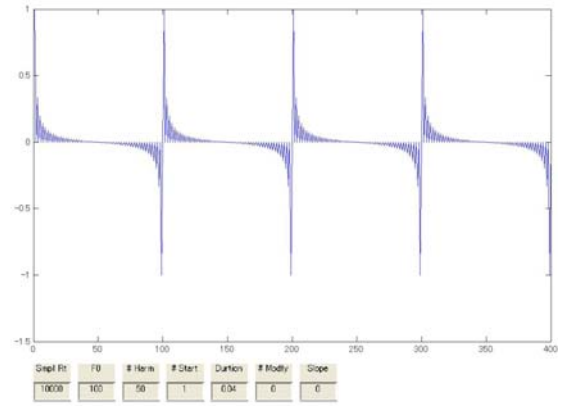

Fig. 3 Speech wave based on sinusoidal harmonic superposition: constant amplitude, 10 kHz sampling, F_0 = 100 Hz.

Figure 4 shows a spectrum of 1/f amplitude, and Fig. 5 shows the corresponding speech wave, which has a saw-tooth shape. Figure 6 shows a spectrum of $1/f^2$ amplitude, and Fig. 7 shows the corresponding speech wave, which has a sinusoid-like shape.

3.2. Simultaneous and sequential output

It is, of course, possible to output simultaneously any specified sequence of harmonics; or harmonics starting from some initial harmonic can be output one at a time. For example, in the latter case, we can choose to output five harmonics starting at 500 Hz. Changes in the voice quality can be perceived as the number of harmonic components is increased or decreased.

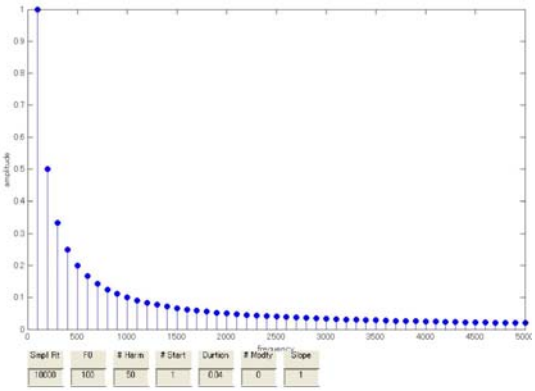

Fig. 4 Spectrum of sinusoidal harmonic superposition: 1/f amplitude, 10 kHz sampling, F_0 = 100 Hz.

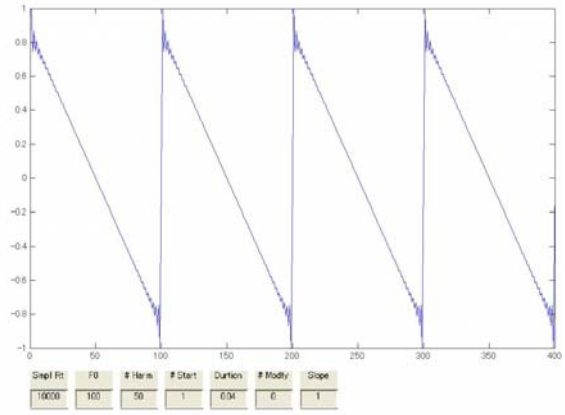

Fig. 5 Speech wave based on sinusoidal harmonic superposition: 1/f amplitude, 10 kHz sampling, F_0 = 100 Hz.

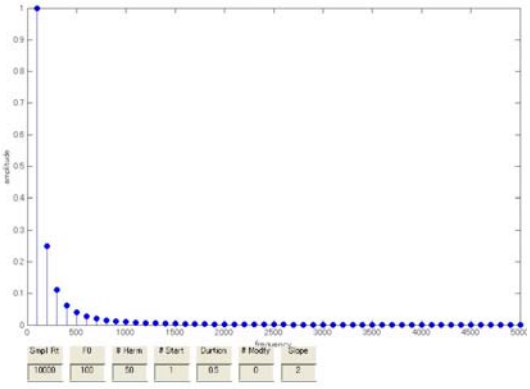

Fig. 6 Spectrum of sinusoidal harmonic superposition: $1/f^2$ amplitude, 10 kHz sampling, F_0 = 100 Hz.

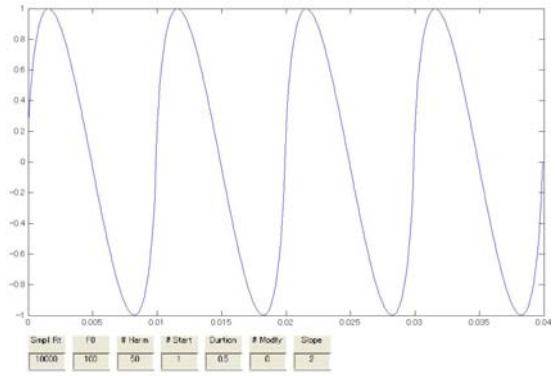

Fig. 7 Speech wave based on sinusoidal harmonic superposition: $1/f^2$ amplitude, 10 kHz sampling, F_0 = 100 Hz.

3.3. Specifying formant regions

The first three formants F1, F2, and F3 of the Japanese vowel sound /a/ when spoken by an adult male lie around 800, 1300, and 2500 Hz, respectively. When F_0 is 100 Hz, a sound similar to /a/ can be synthesized by selecting the harmonics at 700, 800, and 900 Hz, 1200, 1300, and 1400 Hz, and 2400, 2500, and 2600 Hz. It is possible to obtain a vowel sound of better quality by setting the formants beyond F3, but here we limit the formants to those up to F3 for simplicity. Changes in voice quality can be controlled by varying the amplitudes of these harmonics. The number of frequency components in each region can be changed within a range of one through five. Figure 8 shows the line spectrum of a sound similar to /a/ with three components in each formant region.

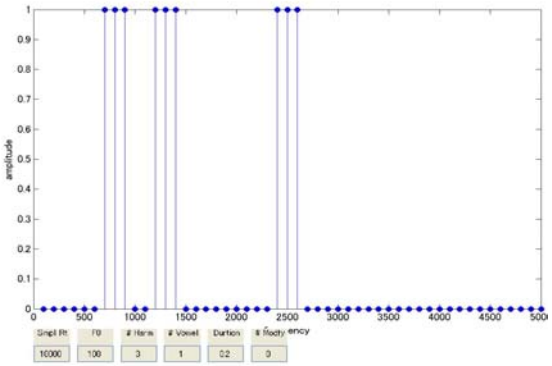

Fig. 8 Line spectrum in the region of formants of the Japanese vowel /a/. 10 kHz sampling rate, F_0 = 100 Hz.

3.4. Arbitrary selection of harmonics

An arbitrary set of harmonic components can be selected. For example, we can select the 300, 500, and 1000 Hz components. In this way, various perceptible changes in sound quality can be generated by varying which harmonic components are used.

3.5. Phase control

In the proposed system, the phase of each harmonic can also be controlled. The phase of each harmonic component is typically set to zero, but we can set the phases randomly within the range zero to $\pi/2$, π and 2π.

4. VOWEL SYNTHESIS OF VARIOUS LANGUAGES

Using the proposed system, we can synthesize a vowel by calculating the spectrum envelope from the formant frequencies. Figure 9 shows an example screenshot for a Japanese male voice /a/. The upper right of the figure shows the selected language, and the upper left table lists the F1 to F3 values of the five Japanese vowels. The speech spectrum is illustrated in the center of the figure. The lower left shows a F1–F2 diagram and the lower right shows the speech wave. At the bottom of the figure, various parameter values are shown, such as sampling rate and fundamental frequency.
The transfer function of the *n*-th formant is represented by the following equations:

$$H_n(z) = 1/\{(1 - z_n z^{-1})(1 - z_n^* z^{-1})\}, \quad (2)$$

$$z = \exp(sT) = \exp(-\sigma T)\exp(j\omega T), \quad (3)$$

$$\sigma_n = \pi B_n, \omega_n = 2\pi F_n, \quad (4)$$

where F_n and B_n designate the frequency and bandwidth of the *n*-th formant, respectively. The transfer function of the first *N* formants is represented by eq. (5):

$$\begin{aligned}|H(z)|^2 = \Pi_{n=1}^{N} |H_n(z)|^2 = \Pi_{n=1}^{N} 1/\{1 - 4\exp(-\sigma_n T)(\cos\omega_n T)\cos\omega T \\ + 2\exp(-2\sigma_n T)\cos 2\omega T + 4\exp(-2\sigma_n T)(\cos\omega_n T)^2 \\ - 4\exp(-3\sigma_n T)(\cos\omega_n T)\cos\omega T + \exp(-4\sigma_n T)\}\end{aligned} \quad (5)$$

We referred to References [5]–[12] for the first three formants of each language. For F4 and higher formants, we calculated the resonance frequency of a uniform vocal tract tube with a length of 17 cm for males, 14 cm for females, and 12 cm for children. The formants of the males, females, and children are chosen automatically by specifying the fundamental frequency. Table 1 shows the formant frequencies for a Japanese adult male voice. Bandwidth is given as a function of formant frequency in the following equation:

.

$$B_n = 50\{1 + F_n^2/(6\times 10^6)\}. \quad (6)$$

Figure 10 shows an F1–F2 diagram for the first 8 cardinal vowels. Clicking the button for a vowel causes the vowel to be synthesized so that it can be heard. At present, the system can synthesize the stationary vowels of seven languages including Japanese, Chinese, Korean, Thai, Hindi, American English and German, as well as a subset of the cardinal vowels.

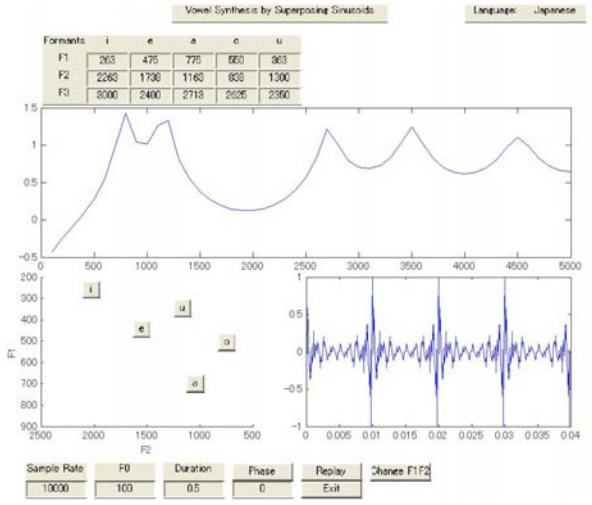

Fig. 9 Screenshot for a Japanese male voice /a/.

Table 1 Examples of formant frequencies for the five Japanese vowels (male voice, in Hz).

Male	F1	F2	F3
/i/	263	2263	3000
/e/	475	1738	2400
/a/	775	1163	2713
/o/	550	838	2625
/u/	363	1300	2350
F4=3500		F5=4500	
F6=5500		F7=6500	

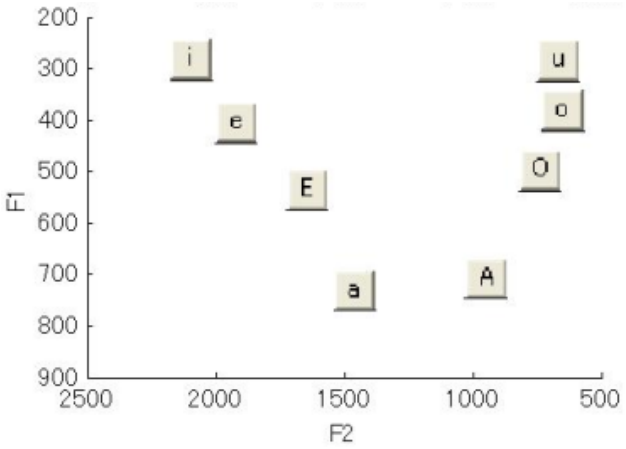

Fig. 10 F1–F2 diagram of 8 cardinal vowels.

5. CONCLUSION

We have introduced a stationary sound synthesis GUI system that superposes sinusoidal harmonics. With this system, we can synthesize vowel-like sounds and stationary vowels of seven languages by changing the amplitude of the harmonics. Increasing the number of languages and improving the present system to make it more easy to use remain as future work.

Acknowledgements

We express our sincere gratitude for advising us on the use of MATLAB to Dr. Y. Ishimoto of the NII Speech Media Group and Dr. M. Yoshikawa, former member of the Speech Processing Group at AIST. We also thank Dr. Y. Arimoto from RIKEN for her suggestions for improvements to the display of spectra and speech waves.

6. REFERENCES

[1] http://www.phon.ucl.ac.uk/resource/software.php
[2] http://www.speech.kth.se/wavesurfer/formant/
[3] S. Itahashi and H. Kojima, “Vowel synthesis by superposing sinusoids with application to education in acoustics,” (in Japanese) Preprints, Fall Meeting of ASJ, pp. 231–232 (Sep. 2010).
[4] S. Itahashi and H. Kojima, “Vowel synthesis system by superposing sinusoids with application to education in acoustics,” (in Japanese) IEICE Technical Report SP2011-39, pp. 51-56 (Jun. 2011).
[5] "Vowel Chart with Sound Files" http://www.linguistics.ucla.edu/people/hayes/103/charts/VChart/
[6] H. Kasuya, H. Suzuki and K. Kido, “Changes in pitch and first three formant frequencies of five Japanese vowels with age and sex of speakers,” (in Japanese) J. Acous. Soc. Jpn. 24 (6), 355–364 (1968).
[7] G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of Vowels," JASA. 24 (2), 175-184 (Mar. 1952).
[8] Matthias Paetzold, Adrian P. Simpson, "Acoustic analysis of German vowels in the Kiel Corpus of Read Speech," in A.P.Simpson, K.J.Kohler, T. Rettstadt, eds., The Kiel Corpus of Read/Spontaneous Speech - Acoustic Database, Processing Tools and Analysis Results, Arbeitsberichte des Instituts fuer phonetik und digitale Sprachverarbeitung der Universitaet Kiel (AIPUK), pp. 215-247 (1997).
[9] I. Khan, S. K. Gupta, S. H. S. Rizvi, "Formant frequencies of Hindi vowels in /hVd/ and C1VC2 contexts" JASA. 96 (4), pp. 2580-2582 (Oct. 1994).
[10] Vaishna Narang, Deepshikha Misra, "Acoustic Space, Duration and Formant Patterns in vowels of Bangkok Thai" International Journal on Asian Language Processing 20 (3):123-140 (2010).
[11] Chong Chang, S. Makino, M. Kimura, K. Kido, "Analysis and recognition of Chinese isolated vowels using formants" (in Japanese) J. Acous. Soc. of Jpn, 47 (4), 281-288 (1991).
[12] Byunggon Yang, "A comparative study of American English and Korean vowels produced by male and female speakers," Journal of Phonetics 24, 245-261 (1996).