# Synthesizing Vowel-Like Tones with Pitch Circularity

Kaori Hashimoto*, Takao Kawamura*, and Nobutaka Ono*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: hashimoto-kaori@ed.tmu.ac.jp, kawamura-takao@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this paper, we present a method for synthesizing vowel-like tones that exhibit pitch circularity, an auditory illusion in which a cyclic sequence of tones appears to endlessly ascend or descend in pitch. Such illusions have traditionally been achieved using synthetic sinusoidal tones, such as the Shepard tone. However, the timbral characteristics of these tones and their relationship to speech have not yet been thoroughly investigated. Aiming at potential applications in music and auditory research, we attempt to synthesize sounds that simultaneously evoke vowel perception and pitch circularity. In our method, we construct a 12-tone scale from a given vowel by varying its fundamental frequency (F0) while preserving its spectral envelope. To induce pitch circularity, we apply pitch-dependent attenuation to the odd harmonics relative to the even harmonics. Subjective listening experiments were conducted to evaluate the pitch perception and vowel identification accuracy of the synthesized tones. Using five Japanese vowels, "A," "I," "U," "E," "O," the proposed method demonstrated stronger pitch circularity while maintaining high vowel identification accuracy (average F1-score: 99.7%). Audio demos of the proposed method are available online.**

## I. Introduction

Auditory illusions have long served as a window into the mechanisms of auditory perception and cognition, while also offering practical utility in various creative domains. A notable example is the Shepard tone, which creates the auditory illusion of a continuously ascending or descending pitch, even though it consists of a finite sequence of tones [1]. This illusion has not only contributed to theoretical explorations of pitch perception, but has also been widely adopted in musical composition, film scoring, and video game sound design to enhance tension, atmosphere, and perceptual continuity. Additional research has shown that it is possible to apply the Shepard tone to melody [2].

Such pitch circularity is often achieved using tone complexes composed of partials spaced at octave intervals. Fig. 1 shows the spectrogram of the tones with pitch circularity generated by the method proposed in [1]. In the Shepard tone, for instance, pitch proximity plays a crucial role: when the pitch interval between consecutive tones is small, listeners tend to judge the direction of pitch change based on the closest perceived pitch height [3]–[5]. Additional research has expanded on this approach by demonstrating pitch circularity using stimuli with non-octave intervals [6], [7]. However, these octave-related designs inherently limit the types of tones that can be generated. If pitch circularity were constrained only to such tone complexes, its theoretical implications and practical applications would remain relatively narrow.

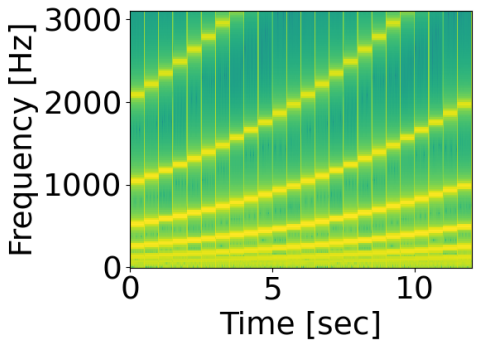


Fig. 1. The spectrogram of the tones with pitch circularity generated by the method proposed in [1]. These spectrograms show a scale consisting of 12 tones repeated twice.

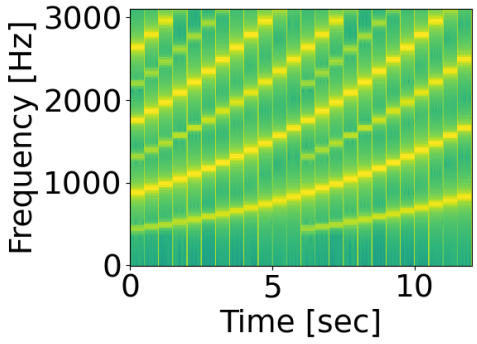


Fig. 2. The spectrogram of the tones with pitch circularity generated by the method proposed in [8]. These spectrograms show a scale consisting of 12 tones repeated twice.

Recent studies suggest that pitch circularity can also be achieved using sine wave tones that contain a full harmonic series. For instance, [8] demonstrated that pitch circularity can be created by manipulating the relative amplitudes of odd and even harmonics. Specifically, suppressing odd harmonics enhances the perception of pitch doubling, enabling the illusion of pitch circularity even with harmonically rich tones. Fig. 2 shows the spectrogram of the tones with pitch circularity generated by the method proposed in [8]. Related work has shown that a good flute player can modulate harmonic balances to control pitch perception [9], and pitch judgments of tone sequences with reduced odd harmonics have also been ex-

979-8-3315-7206-8/25/$31.00 

plored in this context [10], [11]. These approaches broaden the expressive and analytical potential of pitch circularity beyond octave-spaced tones.

In this study, we propose a method for simultaneously achieving pitch circularity and vowel perception. While it is well known that vowel identification can be performed robustly across different fundamental frequencies in natural speech, such an illusion of endlessly ascending or descending pitch cannot be realized with ordinary vowel sounds. In our method, we first construct a 12-tone scale from a single vowel by preserving its spectral envelope and adjusting the fundamental frequency (F0) in semitone steps. To induce pitch circularity, we apply pitch-dependent attenuation to the odd harmonics relative to the even harmonics, following the principle proposed in [8]. We also conduct subjective listening experiments to evaluate both perceived pitch circularity and vowel identification accuracy in the resulting tones.

The rest of this paper is organized as follows. In Section II, we describe the preliminary investigation on two approaches and the proposed method. In Section III, we present a subjective evaluation experiment and provide music demos of audio samples that are available online. Finally, we conclude the paper in Section IV.

## II. Synthesis of Pitch-Circular Vowel Tones

### A. Preliminary investigation on two approaches

The purpose of this study is to synthesize tones that realize both pitch circularity and vowel perception. In the following, we refer to such tones as "pitch-circular vowel tones."

As a preliminary investigation toward this purpose, we examined two methods: the method known for generating the Shepard tone [1], which uses octave-spaced sinusoids combined with a fixed spectral envelope; and the method proposed by Deutsch et al. [8], which applies pitch-dependent attenuation to the odd harmonics relative to the even ones. The spectrograms of these two tone types are illustrated in Fig. 1 and Fig. 2, respectively.

The second method is based on the principle that suppressing odd harmonics enhances the perception of pitch doubling [8]. For example, when the odd harmonics of a tone with a fundamental frequency of 110 Hz are sufficiently attenuated, listeners tend to perceive a pitch near 220 Hz, which is one octave higher. In this method, as the pitch is gradually increased in semitone steps, the attenuation of the odd harmonics is progressively reduced. When the pitch reaches one semitone below the octave (e.g., 208 Hz in this case), the attenuation becomes zero. Because the perceived pitch at the beginning is already close to the octave (due to pitch doubling), each pitch step is perceived as a continuation of the upward motion. Repeating this pattern creates the illusion of endlessly ascending pitch.

We applied a spectral envelope filter, estimated from a vowel utterance using linear predictive analysis, to these two types of tone complexes. The tones based on the first method were perceived to exhibit pitch circularity, but vowel identity was not perceived. We presume this is because the harmonics exist only at octave intervals in this method, making the spectral components too sparse to evoke clear formant structures. In contrast, the second method led to better perception of vowel quality along with pitch circularity compared to the first method. Based on these observations, we propose a method that extends the second approach by applying it to synthesized vowel sounds.

### B. Proposed method

The proposed method aims to synthesize tones that exhibit both pitch circularity and vowel perception, referred to as *pitch-circular vowel tones*. It consists of the following four steps:

1) Select a single vowel sound at a fixed pitch.
2) Extract its spectral envelope and synthesize a 12-tone scale by varying the fundamental frequency (F0) in semitone steps.
3) Apply pitch-dependent attenuation to the odd harmonics relative to the even harmonics.
4) Normalize the amplitude of each tone to maintain consistent power across the scale.

In the following, we describe each step in detail.

**Step 1: Vowel Selection and Spectral Envelope Extraction**

We begin by selecting a single vowel sound at a fixed pitch. Its spectral envelope is extracted using an analysis and synthesis tool such as WORLD [12].

**Step 2: Synthesis of a 12-Tone Scale**

Using the extracted spectral envelope, we synthesize twelve tones by varying the fundamental frequency (F0) in semitone steps. This results in a 12-tone scale spanning one octave, in which all tones share the same vowel identity. We refer to this set of tones as the *normal vowels*. The tone with the lowest fundamental frequency is referred to as the tonic, and the index of each tone is denoted by $i$, representing the number of semitone steps from the tonic. The fundamental frequency of each tone is denoted as $f_i$.

**Step 3: Pitch-Dependent Harmonic Manipulation**

To induce pitch circularity while preserving vowel perception, we manipulate the relative amplitudes of the odd and even harmonic bands in the normal vowels. Following the method proposed in [8], amplitude manipulation is performed by applying weights to the spectrogram of the normal vowel such that the relative sound pressure level of odd harmonics decreases by $L$ dB as the pitch decreases by a semitone. This is implemented by the following:

$$Y_i(f,\tau) = \begin{cases} R_i X_i(f,\tau) & \text{if } \left\lfloor \dfrac{f}{f_i} + \dfrac{1}{2} \right\rfloor \equiv 1 \pmod 2, \\ X_i(f,\tau) & \text{otherwise,} \end{cases} \tag{1}$$

$$R_i = 10^{\frac{(i-11)L}{20}}, \tag{2}$$

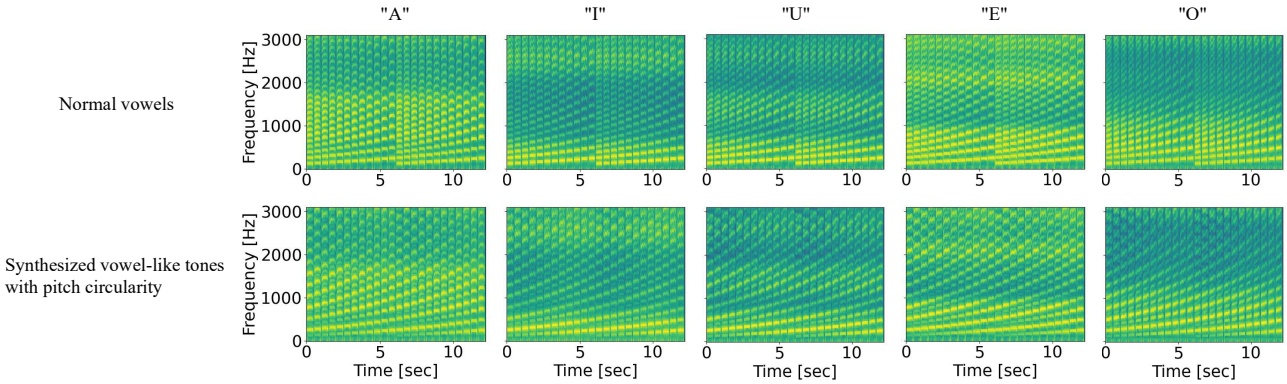


Fig. 3. Spectrograms of the normal vowels and the synthesized vowel-like tones with pitch circularity based on the proposed method.

where $X_i(f, \tau)$ is the short-time Fourier transform of the normal vowel at index $i$, and $f$ and $\tau$ are the frequency bin and time frame indices, respectively. $f_i$ is the fundamental frequency of tone $i$, and $R_i$ is the weighting factor used to attenuate the odd harmonics. We adopted the same value for $L$ ($L = 3.5$) as reported in [8].

**Step 4: Power Normalization**

The amplitude manipulation introduces variations in power across tones. To ensure uniform power levels, we normalize the amplitude of each tone accordingly. The inverse short-time Fourier transform is then applied to obtain the final *pitch-circular vowel tones*.

These processes are expected to produce tones that exhibit both pitch circularity and vowel perception, while preserving the spectral envelope and systematically controlling harmonic balance. When arranged and repeated cyclically, the twelve pitch-circular vowel tones give the auditory impression of an endlessly ascending or descending pitch.

Figure 3 shows spectrogram examples of the normal vowels and pitch-circular vowel tones for the Japanese vowels "A," "I," "U," "E," and "O," arranged so that the tones ascend by a semitone. These spectrograms show a scale consisting of 12 tones repeated twice. The display range of the spectrograms is from 0 to about 3100 Hz. The spectrograms with the normal vowels clearly show the repeated parts of the tones. On the other hand, in the spectrogram of the pitch-circular vowel tones, the amplitude of odd harmonics gradually increases. As a result, even in the repeated part, it can be seen that the sound naturally shifts to the next repetition. It can also be observed that the spectral envelope of the vowel is visually preserved across all tones.

## III. Experiments

### A. Experimental conditions

A subjective evaluation experiment was conducted to evaluate the judgments based on proximity and vowel identification using the normal vowels and pitch-circular vowel tones. A single fourth note of vowel sound was first synthesized as a base vowel using Sinsy [13], [14] for each of the five Japanese vowels ("A," "I," "U," "E," "O") at a tempo of 120. For each base vowel, twelve tones were then synthesized by adjusting the fundamental frequency in WORLD [12] while preserving the spectral envelope, resulting in twelve semitone-spaced tones. These twelve tones for each vowel served as a total of 60 normal vowels (12 tones $\times$ 5 vowels). Amplitude manipulation based on the proposed method was applied to each of these tones, yielding a total of 60 pitch-circular vowel tones. The parameters of the short-time Fourier transform and inverse short-time Fourier transform in the amplitude manipulation were set to a frame length 2048 and a frame shift of $\frac{1}{8}$ of the frame length. The tonic of the pitch-circular vowel tones was set to C3 (130.8 Hz), while that of the normal vowels was set to F#3 (185.0 Hz) to ensure a comparable perceived pitch height across the two sets.

Subjects were presented with tone pairs and instructed to identify the vowel category of each tone and to judge whether the pitch of the pair was ascending or descending. Each pair consisted of two tones at different pitches, and all permutations of two tones selected from the 12 tones were used, resulting in 132 pairs per tone type (normal vowels or pitch-circular vowel tones). Thus, a total of 264 tone pairs were created and presented to the subjects. For the normal vowels and pitch-circular vowel tones, pairs of the same pitch were set to the same vowel. The combination of the two vowels was adjusted so that they appeared equally. The order of presentation of the pairs and the placement of the vowels was random. All tone pairs were divided into 24 blocks of 11 pairs each. Each tone in a pair was presented continuously

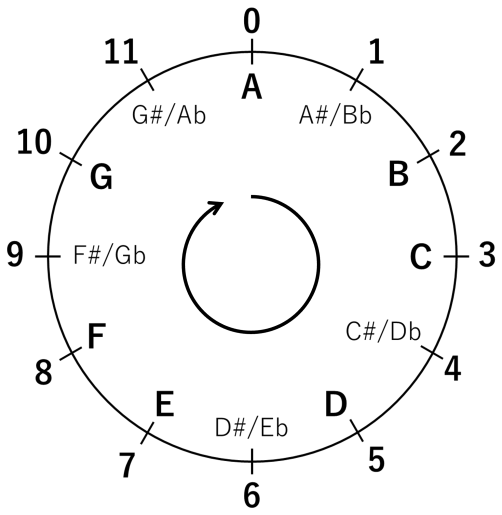


Fig. 4. Pitch class circle.

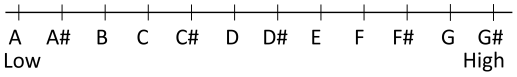


Fig. 5. Pitch class arranged in a straight line.

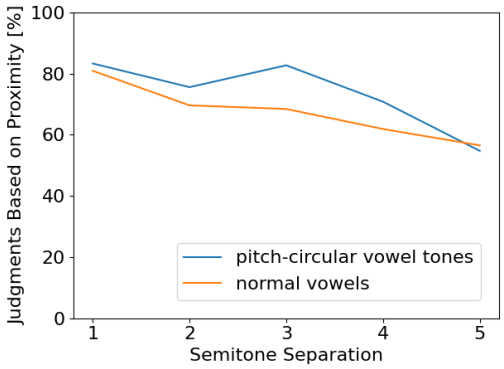


Fig. 6. Percentages of judgments Based on Proximity as a function of semitone separation along the pitch class circle between the tones within a pair.

without intervals, allowing the subject to play each pair. A break of approximately 1 minute was set up between blocks. Subjects performed the experiment twice, making judgments on four randomly selected blocks.

A total of 21 participants were recruited from Tokyo Metropolitan University students and related persons. These were 16 males and 5 females, and all participants had normal hearing as verified by self-report. Their average age was 22.9 years (range 20-29 years), and the average musical experience was 5.0 years (range 0-17 years), with 14 participants having musical experiences and 7 having none. The tones were presented to participants diotically in the soundproof room via Sony MDR-CD900ST headphones. Participants played the test sound and adjusted the volume to a comfortable level before beginning the experiment. The study was approved by the Research Ethics Committee of Tokyo Metropolitan University (Approval No. R7-013) .

For vowel judgments, Precision, Recall, and F1-score were calculated for “A,” “I,” “U,” “E,” and “O,” respectively. For pitch judgments, we calculated the percentage of judgments based on pitch class proximity [8]. The percentage of judgments is expressed by the following equation:

$$P = \frac{N_{\text{proximity}}}{N_{\text{sum}}}, \tag{3}$$

where $N_{\text{sum}}$ is the total number of judgments made by subjects for each pair, and $N_{\text{proximity}}$ is the total number of judgments that are based on pitch class proximity. Let $i$ and $j$ denote the indices corresponding to the fundamental frequencies of the first and second tones in the pair, respectively. For judgments based on pitch class proximity, the correct answer was “ascending” if $1 \leqq j - i \leqq 5$ or $-11 \leqq j - i \leqq -7$. In contrast, “descending” was correct when $-5 \leqq j - i \leqq -1$ or $7 \leqq j - i \leqq 11$. Here is an example of the judgments based on the proximity of pitch classes. For example, when tones A and G are presented, the difference in pitch between them can be considered to be either 2 semitones or 10 semitones, but if the respondent perceives the pitch of tone A to be higher because of the closer proximity in the Fig. 4, in this case 2 semitones, the judgment is based on the proximity of the pitch class. On the other hand, tone pitch can be represented as a straight line with respect to the normal tones as shown in the Fig. 5, and when comparing tones A and G, the pitch of tone G is perceived as higher. In other words, the closer the percentage of judgments based on pitch class proximity to 100%, the more likely the pitch circularity is perceived.

### *B. Results*

The percentage of judgments based on pitch class proximity is shown in Fig. 6 as a function of the semitone separation along the pitch class circle between the tones within a pair. For 1 to 4 semitone separations, the percentage of judgments for the pitch-circular vowel tones based on the proposed method was higher than the percentage of judgments for the normal vowels; for 5 semitone separations, there was little difference in the percentage between the two. In most cases, it was suggested that the pitch-circular vowel tones promoted pitch judgments based on pitch proximity more effectively than the normal vowels.

A $5\times2$ mixed analysis of variance (ANOVA) was performed, with the value of semitone separation along the pitch class circle (1-5) and sound used (normal vowels, pitch-circular vowel tones manipulated based on the proposed method) as factors. The value of 6 semitones was omitted from the analysis since at this value the same distance between the tones along the pitch class circle is traversed in either direction, so that proximity cannot be used as a cue. The overall effect of semitone separation was highly significant ($p < 0.001$). The overall effect of sound used was also significant ($p < 0.01$),

TABLE I
PRECISION, RECALL AND F1-SCORE OF VOWEL DISTINCTIONS FOR THE NORMAL VOWELS AND THE PITCH-CIRCULAR VOWEL TONES BASED ON THE PROPOSED METHOD.

| | vowel | Precision [%] | Recall [%] | F1-score [%] |
|---|---|---|---|---|
| normal | "A" | 99.5 | 99.7 | 99.6 |
| | "I" | 100.0 | 99.7 | 99.9 |
| | "U" | 99.7 | 100.0 | 99.9 |
| | "E" | 99.7 | 99.7 | 99.7 |
| | "O" | 99.7 | 99.5 | 99.6 |
| pitch-circular | "A" | 99.5 | 100.0 | 99.7 |
| | "I" | 99.7 | 100.0 | 99.9 |
| | "U" | 99.7 | 99.7 | 99.7 |
| | "E" | 100.0 | 99.7 | 99.9 |
| | "O" | 100.0 | 99.5 | 99.7 |

and the interaction between the value of semitone separation and sound used was nonsignificant ($p > 0.1$).

TABLE I shows the Precision, Recall, and F1-score of vowel distinctions for the normal vowels and the pitch-circular vowel tones based on the proposed method. The experimental results show that the F1-score for the distinction of vowels in the normal vowels is 99.73% on average. The F1-score of the distinction of vowels for the pitch-circular vowel tones based on the proposed method is 99.78% on average. High classification accuracy was observed for all vowels, suggesting that the amplitude manipulation introduced by the proposed method did not adversely affect vowel classification performance.

### C. Demo

As an example of the application of the proposed method, we created music demos using only 12 semitone-spaced tones within one octave. The corresponding audio samples are available online[1].

In conventional settings, it is difficult to express melodies spanning a wider pitch range using such a limited tone set. Even if tones share the same pitch class, differences in pitch height cannot be expressed, often resulting in unnatural pitch transitions. In contrast, the proposed method enables pitch circularity based on pitch class proximity, allowing listeners to perceive pitch more naturally, even when the melody extends beyond one octave.

To examine this, we used parts of "Nanatsunoko" (a Japanese folk song with lyrics by Ujo Noguchi and music by Nagayo Motori) and "Boléro" (a classical piece by Maurice Ravel). They span 18 and 14 semitones respectively, but the semitone separation between adjacent tones is no more than five. We prepared sounds of the lyrics for "Nanatsunoko" and "La-la-la" vocalizations for "Boléro" respectively using Sinsy [13], [14], with all tones initially at the same pitch. Then, WORLD [12] was used to shift each tone to the corresponding pitch class within the range of C3 (130.8 Hz) to B3 (246.9 Hz). Amplitude manipulation based on the proposed method was applied only to the voiced segments of the normal tones to produce the pitch-circular vowel tones. The parameters of the short-time Fourier transform and inverse short-time Fourier transform were set to a frame length of 2048 and a frame shift of $\frac{1}{8}$.

The difference between the normal vowels and pitch-circular vowel tones can be clearly heard in the audio samples. In the normal vowels, unnatural pitch transitions can be perceived when tones outside one octave range are replaced by same pitch class tones within the limited range. In contrast, the pitch-circular vowel tones provide more natural pitch transitions based on pitch class proximity, demonstrating the expressive potential of the proposed method.

## IV. CONCLUSIONS

In this study, we proposed a method for synthesizing tones that simultaneously produce vowel perception and pitch circularity. In the proposed method, 12 semitones were created by adjusting the fundamental frequency while preserving the spectral envelope of the vowel sound, and these tones were used as the normal vowels. The perceived pitch was manipulated by adjusting the relative amplitudes of the bands of odd and even harmonics, and vowel tones with a cyclical pitch perception were synthesized. Subjective evaluation experiments showed that for most of the semitone separations, the percentage of judgments for the pitch-circular vowel tones based on the proposed method was higher than that for the normal vowels. The F1-score for vowel distinction based on the proposed method was 99.7% on average. In the future, we would like to apply the proposed method to various types of speech sounds to investigate new musical expressions and continue to study the synthesizing method with the aim of further improving the accuracy of pitch circularity and vowel perception.

## REFERENCES


[1] R. N. Shepard, "Circularity in judgments of relative pitch," *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, Dec. 1964. DOI: 10.1121/1.1919362.

[2] P. Patrício, "From the Shepard tone to the perpetual melody auditory illusion," in *Proc. Sound and Music Computing Conference (SMC)*, 2012. DOI: 10.5281/zenodo.850104.

[3] K. Ueda and K. Ohgushi, "Perceptual components of pitch: Spatial representation using a multidimensional scaling technique," *The Journal of the Acoustical Society of America*, vol. 82, no. 4, pp. 1193–1200, Oct. 1987. DOI: 10.1121/1.395255.

[4] J. Allik, E. N. Dzhafarov, A. J. M. Houtsma, J. Ross, and N. J. Versfeld, "Pitch motion with random chord sequences," *Perception & Psychophysics*, vol. 46, no. 6, pp. 513–527, Nov. 1989. DOI: 10.3758/BF03208148.

[5] D. Deutsch, "Pitch proximity in the grouping of simultaneous tones," *Music Perception*, vol. 9, no. 2, pp. 185–198, Dec. 1991. DOI: 10.2307/40285528.


[1] https://153hashimoto.github.io/synthesizing-vowel-like-tones-with-pitch-circularity/

[6] E. M. Burns, "Circularity in relative pitch judgments for inharmonic complex tones: The Shepard demonstration revisited, again," *Perception & Psychophysics*, vol. 30, no. 5, pp. 467–472, Sep. 1981. DOI: 10.3758/BF03204843.

[7] Y. Nakajima, T. Tsumura, S. Matsuura, H. Minami, and R. Teranishi, "Dynamic pitch perception for complex tones derived from major triads," *Music Perception*, vol. 6, no. 1, pp. 1–20, Oct. 1988. DOI: 10.2307/40285413.

[8] D. Deutsch, K. Dooley, and T. Henthorn, "Pitch circularity from tones comprising full harmonic series," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 589–597, Jul. 2008. DOI: 10.1121/1.2931957.

[9] A. H. Benade, *Fundamentals of Musical Acoustics*. New York: Dover Publications Inc., 1990.

[10] R. D. Patterson, R. Milroy, and M. Allerhand, "What is the octave of a harmonically rich note?" *Contemporary Music Review*, vol. 9, no. 1-2, pp. 69–81, 1993. DOI: 10.1080/07494469300640351.

[11] J. D. Warren, S. Uppenkamp, R. D. Patterson, and T. D. Griffiths, "Separating pitch chroma and pitch height in the human brain," *Proceedings of the National Academy of Sciences*, vol. 100, no. 17, pp. 10 038–10 042, 2003. DOI: 10.1073/pnas.1730682100.

[12] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016. DOI: 10.1587/transinf.2015EDP7457.

[13] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Sinsy: A deep neural network-based singing voice synthesis system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2803–2815, 2021. DOI: 10.1109/TASLP.2021.3104165.

[14] *Sinsy - HMM/DNN-based Singing Voice Synthesis System*. [Online]. Available: https://www.sinsy.jp/.