Synthesis of Choir Songs Using MBROLA with Multiple Voices

Yohanes Suyanto, *Member, IAENG*

***Abstract*—Previous research has explored the synthesis of singing using the input of numbered musical notation and song lyrics, with a primarily focus on solo singers. This study takes a different approach by synthesizing songs that are simultaneously sung by 2 – 4 singers, essentially forming a choir performance. Each synthetic song a soprano (S), alto (A), tenor (T), and a bass (B), collectively known as SATB voices. Each voice type is separately synthesized using the MBROLA speech synthesis software. The resulting individual voice components are then combined into a single audio sound using the cross-platform audio editing software SOX. Findings reveal that the fundamental frequency produced when combining these four voice types differs from the average, thereby highlighting a distinctive outcome.**

***Index Terms*—numbered musical notation, singing synthesis, SATB, choir, SOX.**

I. Introduction

IN [1], a humanoid robotic head was developed and is capable of performing songs based on a script written in numbered notations. This robotic system was designed to mimic human characteristics. The image data of the script are converted into text and then transformed into a sound format using the Harmonic plus Noise Model (HNM).

The sound synthesis was performed by syllables, allowing the system to adapt to the Chinese script, which follows the CxVCn structure. Here, "Cx" can represent null (empty), sound consonant phonemes, or unsound consonant phonemes; "Cn" can mean null or nasal phonemes, such as /n/ or /ng/; and "V" represent vowel, diphthong, or triphthong phonemes.

If the "Cx" portion corresponded to a long silent consonant, such as /s/ or /p/, then the synthesis signal was transformed into a noise signal using the HNM. If the "Cx" section short such as /b/ or /d/, the signal synthesis was directly copied to other related parts within the same syllable. However, if the "Cx" part represented a consonant sound, such as /m/ or /r/, and was considered necessary for other phonemes, the synthesis signal was generated as a combination of partial harmonics and noise using HNM.

This speech-to-singing conversion system (STS), was designed to correct the pitch inaccuracies often from non-professional singers. Various components were considered, including the fundamental frequency (F_0) and duration per phoneme.

In essence, this study discusses the development of a humanoid robot capable of singing songs based on a script, with a focus on the techniques used to convert text into sound, especially in the context of the Chinese language's phonetic structure.

Manuscript received June 5, 2023; revised November 27, 2023.

Y. Suyanto is an Assistant Professor of Departemen of Computer and Electronics, Universitas Gadjah Mada, Indonesia. (email: yanto@ugm.ac.id)

Several components within this system — including the fundamental frequency (F_0), phoneme duration, and spectrum — undergo control and adjustment based on templates derived from voice recordings of one song by several professional singers. A typical STS system comprises three key stages: learning, transformation, and synthesis.

During the learning stage, template voices and speech sounds undergo analysis to extract essential features such as Mel-frequency cepstral coefficients (MFCC), short-time energy, and information about voiced and unvoiced segments (VUV). Following this analysis, phoneme alignment is performed using a two-stage transformation that utilizes the dynamic time warping algorithm.

In the transformation stage, discrepancies in the duration of individual phonemes and the fundamental frequency contours are addressed. These aspects are refined and improved to increase the suitability of the resulting phonemes for singing.

Ultimately, in the synthesis stage, the refined phonemes are prepared for transformation into actual sounds, helping ensure that the output closely aligns with the desired singing performance [2].

The synthesis of singing voices in the Chinese language has also been explored [3]. Unit selection synthesis was used and focused on the smaller segments, specifically half-syllables, to prevent the need for handling an extensive array of phoneme combinations. Notably, this approach steers clear of using diphone-sized units to prevent the blending of adjacent units, which is a challenge often encountered in Chinese language synthesis. The research involved the creation of a library by recording the singing voices of women performing 44 Chinese songs, resulting in a collection of 5882 syllables, each sampled at a frequency of 22.050 Hz.

ByteSing is a Chinese singing voice synthesis (SVS) system that relies on acoustic models, specifically using the Tacotron algorithm for duration and the WaveRNN neural vocoder [4]. Its innovative approach sets ByteSing apart from the traditional SVS model. An encoder–decoder structure reminiscent of Tacotron is adopted for its acoustic model. Within this structure, the encoder utilizes the CBHG model while the decoder uses repetitive neural networks. Furthermore, ByteSing incorporates an additional phoneme duration prediction model, which serves to extend the input sequence. This inclusion leads to the potential to enhance the model's controllability, stability, and accuracy in predicting tempo.

In machine learning, several studies carry out singing syntheses. In the STS system, convolutional neural networks can be used to model the long-term dependencies of singing voices, which allows for the generation of natural-sounding singing voices even with complex melodies and harmonies [5]. Then, the system proposes a new one that is both fast and high-quality, and has the potential to be used in various

applications.

A new approach to SVS uses a generative adversarial network (GAN) called HiFi-WaveGAN [6], which can generate high-fidelity singing voices that are indistinguishable from human ones. HiFi- WaveGAN consists of two components: a generator and a discriminator. For its input, the generator takes a sequence of mel-spectrograms that represent the pitch and timbre of a singing voice, and then generates a corresponding waveform. The discriminator takes a pair of waveforms, one real and one generated, and attempts to distinguish them.

The generator is trained using a technique called adversarial training, where the generator and discriminator are trained simultaneously. The generator attempts to fool the discriminator into considering that its output is real, while the discriminator attempts to distinguish between real and generated waveforms. Thus, the system can generate high-fidelity singing voices that are indistinguishable from those of humans. The system is also significantly faster than previous ones for SVS.

MelGAN is an architecture for conditional waveform synthesis to address the challenge of generating coherent raw audio waveforms using GANs [7]. The results show that MelGAN achieves superior performance in audio quality and similarity to the target waveform. This system contributes to the field of audio generation and provides insights into the development of high-quality audio synthesis models.

An SVS system based on Hidden Markov Models is also proposed and focuses on Mandarin Chinese songs [8]. The system incorporates a Mandarin Chinese singing voice corpus and includes well-designed musical contextual features for training. Experimental results demonstrate that the proposed system outperforms the baseline in terms of generating a more natural F_0 contour.

Numeric music notation is an alternative form of musical notation that uses numbers 1–7 to represent musical notes. Each of these numbers corresponds to a specific note: 1 for 'do', 2 for 're', 3 for 'mi', 4 for 'fa', 5 for 'so', 6 for 'la', and 7 for 'ti.' Different octaves can be indicated by placing a dot (.) either below or above these number symbols.

Notably, the frequency of the 'do' tone, for instance, may not always be consistent across different songs. Therefore, a reference for pitch must be established by specifying that 'do' equals 'C' [9].

In accordance with [10], the synthesis of singing voices using numeric music notation and lyrics for Indonesian songs was also explored. The findings suggest the feasibility of effectively synthesizing singing voices using the current speech synthesis technology. However, the study was restricted to a single or solo singing voice. As such, the findings allow for the potential synthesis of multiple voices or a choir.

The genesis of numeric musical notation in Indonesia can be traced back to the period of Dutch colonial rule [9]. At that time, the locals faced challenges in deciphering traditional staff notation. In consideration of the needs of missionaries who had arrived in Indonesia to facilitate hymn singing during religious gatherings, the Dutch authorities introduced a numerical notation system for reading music. This system was designed to simplify the teaching of music to the local population and is still in use today.

A. MBROLA

MBROLA is a speech synthesis system that carries out concatenation of phonemes provided as input in phonetic transcription. MBROLA is equipped with a database of recorded phone sounds encompassing multiple languages sourced from 35 different countries. The MBROLA Project, initiated by the TCTS Lab at the Faculte´ Polytechnique de Mons in Belgium in 1996, is the driving force behind this application. The overarching goal of the project is to offer the free MBROLA application for non-commercial use, and thus enable the creation of speech synthesizers in as many languages as possible [11].

B. SATB numbered music notation

A choir refers to a collective of singers with diverse vocal qualities, brought together and guided by a conductor. The conductor assumes leads the group of vocalists, instrumentalists, or both. Essentially, a choir is an ensemble of vocalists organized into distinct voice categories, including soprano (S), alto (A), tenor (T), and bass (B), (SATB) [12]. Notably, while children may not yet have the capacity to perform according to the SATB arrangement commonly seen in choirs, they can still be organized into at least two vocal sections.

Choral singing, often referred to as a choir, represents one of the most prevalent forms of vocal performance. The European Choir Association serves as a prime example, encompassing over 2.5 million singers, conductors, composers, and managers across over 40 European nations. This collective singing community reaches out to more than 37 million active participants in Europe [13].

A choir is a group of individuals with varying vocal ranges who sing together. The commonly utilized voice types include SATB. While several people consider vocal groups as choirs, different perspectives may occur. Given the numerous voices within a choir, this musical notation comprises multiple parts [14].

Human voices can be categorized into four primary types: 1) Soprano, characterized by a high-pitched range of female voices; 2) Alto, with a low pitch of female voices; 3) Tenor, featuring a high pitch in male voices; and 4) Bass, with its low pitch male voices.

Further subcategories of vocal types exist, such as mezzo-soprano between soprano and alto in the female vocal range, and baritone lower than tenor but higher than bass in the male vocal range. For the effective arrangement of choirs, knowledge of voice classification and its associated vocal ranges is essential. Arrangements must align with the range of each voice type.

In addition to singing harmoniously, choir songs can also involve unison singing. The intricacies of unison singing within choirs involve the examination of deviations in pitch and timing among individual singers. This understanding serves as a foundation for creating a system that can synthesize performances resembling unison singing [15]. Meanwhile, subjective listening tests play a vital role in assessing how well the synthesis replicates the essence of unison singing.

Each vocal part in a choir arrangement includes musical and lyrical notations. In Indonesia, numerical notation is

predominantly used for choral music. However, in various other countries, particularly Western nations, standard staff notation is nearly universal. Musical notation allows others to interpret and perform a composer's work. Specific symbols represent each note within the composer's composition. For a mutual understanding of these symbols among musicians, the standardization of musical notation symbols is necessary.

C. Relevant previous research

Notation serves as a system for composing music with basic elements referred to as notes [16]. These notes represent specific tones, each defined by its frequency or number of vibrations [17]. The high or low aspect of short notes is the pitch, which is conveyed through symbols known as notes or pitches such as the seven primary pitches: C-D-E-F-G-A-B. Traditional concepts related to note duration, time signatures, tempo, key signatures, and base tones have been in use for centuries and are also applicable here. The notation system incorporates numbers 1

–7 and chromatic and octave markings. A forward slash denotes a musical note, as in 5/, while a backward slash represents a rest, as in 7\. Dots are used to indicate high or low octaves, with a dot above signifying a high octave.

Research efforts, as exemplified in [18], aid teachers and students in composing notations. However, the challenge of costly music books has also been reported [19]. Thus, the resulting product assists users in generating their own musical scores in beam notation from existing MIDI files.

TThe conversion of musical notes from the MusicXML structure to numbered notation simplifies composition for musicians who are more accustomed to using the latter than beam notation [20]. Letter–number notation results in chord representations in the form of numbered notation fonts [21]. The method involves direct generation, albeit with certain restrictions on chord arrangement variations, and entails the use of lexical and syntax analyzers.

To prepare music notations from audio files, onset detection is employed, resulting in a high accuracy rate of 99.62%. The application outputs data in the form of a MusicXML file [22].

The development of a speech synthesis system in the Indonesian language utilizes text as input and generates sound files as output [23]. This system leverages pho files from MBROLA. Tone settings within voice synthesis serve as the foundation for defining high and low singing tones. The research involves processing plain text input and producing choral singing voice files as output.

II. Research Methods

In the present study, MBROLA tools were used for sound synthesis, and the free open-source software octave was utilized to analyze the combination of multiple sound files and integrate into a single audio file. Figure 1 illustrates the system design. Initially, the content of the song.txt file was extracted using a text reader. Subsequently, the text was divided into individual sections corresponding to SATB segments and then subjected to synthesis. The results were four distinct audio components of the SATB sounds, which were integrated by merging them into a unified wav chant voice file.

A. Text splitter

The text was divided into four distinct sections, each representing the SATB components. Each of these sections comprises lines of numerical notation and song lyrics. The SATB components were identified by their initials as markers, such as 'S' for soprano and so forth. However, this labeling is not absolute; in cases where such markers are absent, the first section is considered as voice 1, the second as voice 2, and so forth. The lines that contain lyrics are labeled with 'L:'.

The text is processed by a splitting mechanism to organize it into rows of numerical notations and their corresponding lyrics. Consequently, the outcome includes the voice text, which contains soprano notation and lyrics, for each of the SATB. The algorithm responsible for this splitting mechanism is detailed in Figure 2.

B. Synthesis

Each section of text, which includes notation and lyrics as derived from the text splitting, is directed to the synthesis phase. Consequently, four distinct parts correspond to the SATB components and are synthesized using song synthesis technology powered by the MBROLA engine. The lyrics lines are transformed into audio with pitch adjustments based on the numerical notation lines. Typically, each syllable in the lyric line is matched with one notation on the notation line. However, in the case of notations marked as "legatura," a single syllable can be matched with two or more notations. The synthesis algorithm is detailed in Figure 3.

C. The symbol used

In this study, symbols were used to represent the following: tones from 1 to 7; note durations such as half, quarter, and eighth notes; half-tone increments or decrements; changing the octave of a tone; and legatura lines. The symbols are illustrated in Table I. The aim is to use symbols that can be easily typed using a standard computer keyboard without the need for *Ctrl*, *Alt*, or *Fn* keys.

D. Intermediate data format

The intermediate data format is read by the pho reader application and must adhere to the requirements of the pho file format. The lines are as follows:

```
<pho> : phoneme duration <frac-freq>
<frac-freq> : frac freq <frac-freq>
```

Each line includes a phoneme name, duration (measured in milliseconds), and a series (which may be empty) of pitch targets. These pitch targets consist of pairs of float numbers; that is, the position of the pitch target within the phoneme (expressed as a percentage of its total duration) and the pitch value in Hertz at such specific position.

Rather than the 45 phonemes in previous literature [24], the present study used 29 to match the database of Indonesian voice in MBROLA: 18 single consonants of /b/, /c/, /d/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/, /w/, /y/, /z/; 2 double consonants (compounds) /ng/ and /ny/; and nine vowel phonemes of /a/, /e/, /@/, /i/, /u/, /o/, /au/, /oi/, and /ai/.

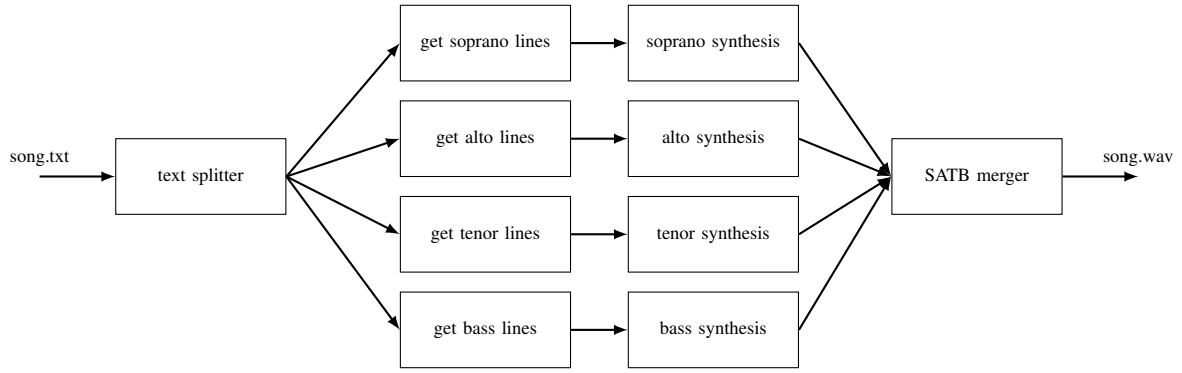

Fig. 1. System design

TABLE I
NUMERICAL MUSIC NOTATION SYMBOLS

Symbols	Function	Examples
12345670.	Tone notation	1 1 2 3 1 1 . 1 1 4 3 2 .
|	Bar	1 1 2 3 1 1 . |1 1 4 3 2 . |
’	One octave above	5’ 1’ 1’ 3’ 1’ 1’ 3’ 2’ 3’ 2’ 1’
,	One octave below	5, 1 1 6, 7, 1 3 2 3 2 1
/	Sharp	5 4/ 4/ 5
\	Flat	5 7\ 7\ 6
( )	Beginning and end of the half-tone group	(5 4) 3 3 (3 4) (5 6) (5 4) 2
{ }	The beginning and end of the sluric arch	(5 4) 3 3 {(3 4)} {(5 6)} {(5 4)} 2

Require: lines of text
Ensure: array of text pairs

while not end of data **do**
  read line
  if line is soprano **then**
    sopranoText[0] ← line
  else if line is alto **then**
    altoText[0] ← line
  else if line is tenor **then**
    tenorText[0] ← line
  else if line is bass **then**
    bassText[0] ← line
  else if line is lyric **then**
    sopranoText[1] ← line
    altoText[1] ← line
    tenorText[1] ← line
    bassText[1] ← line
  else
    next group of notations
  end if
end while

Fig. 2. Notations splitting algorithm

Require:
  notations : array of notations
  lyrics : array of syllabels
Ensure: array of text pairs

$n \leftarrow$ length of notations
for $i \leftarrow 0$, $n-1$ **do**
  note ← notations[i]
  tone ← GetTone(note)
  noteDuration ← GetDuration(note)
  for all ph in lyrics[i] **do**
    if ph if consonant **then**
      duration ← $0.1\times$ noteDuration
    else
      duration ← $0.9\times$ noteDuration
    end if
    print ph tone duration
  end for
end for

Fig. 3. Tone synthesis algorithm

Notably, one syllable corresponds to one note. As a single syllable comprises several phonemes, the duration must be distributed among each phoneme. Phonemes themselves are categorized into two types: vowels and consonants. Vowel phonemes generally have longer durations compared with consonant phonemes, with a ratio of 9:1. The implication is that the vowel part constitutes 90% of the syllable sound duration, while the consonant part accounts for only 10%.

For example, consider the syllable ‘bin’ in the song ‘Bintang Kecil’, which has a pitch of 5. The corresponding SAMPA representation for these syllables is ‘b’, ‘I’, and ‘n’, consisting of two consonants and one vowel. The duration for each phoneme is as follows: ‘b’ (0.05 parts), ‘I’ (0.9 parts), and ‘n’ (0.05 parts).

The frequency values are based on the 4th octave, a tone frequency, which is 440 Hz. For the tone ‘do=C’ or, more precisely, ‘do=C4’ (4th octave C), the fifth tone corresponds

to the 3rd octave G tone, which has a frequency of 98 Hz. Tone calculations are determined using Formula 1 [25]:.

$$f_k = f_d \times A^n \quad (1)$$

where

f_d = fundamental frequency, e.g., A4 = 440 Hz
A = $2^{\frac{1}{12}}$
n = the distance between tone k and tone d in increments of 0.5

E. Merge Audios

Two or more audio channels can be combined into a single channel, and is referred to as audio mixing. A straightforward mixing approach involves adding each corresponding sample, as demonstrated in Figure 4.

Require: channels : array of audio channels
Ensure: audio: a single channel audio

$nc \leftarrow$ length of channels
$n \leftarrow$ length of channels[0]
for $i \leftarrow 0$, $n-1$ **do**
  audio[i] = 0
  for $j \leftarrow 0$, $nc-1$ **do**
    audio[i] $\leftarrow$ audio[i] + channels[j][i]
  end for
end for

Fig. 4. Merge algorithm

F. An input example

An example of an input taken from the first line of the song “Indonesia Pusaka” in Figure 10 can be seen in Figure 5. This conversion refers to Table I. Lines 1–4 consist of numerical music notations representing SATB, while the last line include lyrics. The SATB lines are filtered and read by the corresponding sound section. Lines of lyrics are used as utterances in speech synthesis for each type of sound.

Figure 5 shows an illustrative input, extracted from the initial line of the song “Indonesia Pusaka” depicted in Figure 10. This conversion adheres to the specifications listed in Table I. In Figure 5, the first four lines contain numeric musical notations representing the SATB parts, respectively, while the last line comprises the song lyrics. The SATB lines are filtered and processed by their respective sound sections. Additionally, the lyrics lines are used as utterances in the speech synthesis for each sound type.

```
S.#(5, 1) | 3 (. 1) (5, 1) (3 6) | 5 . (3 0)
A.#(5, 5,) | 1 (. 1) (5, 5,) (1 2) | 3 . (1 0)
T.#(1 2) | 5 (. 3) (1 2) (5 6) | 5 . (5 0)
B.#(5, 5,) | 1 (. 1) (1 1) (1 2) | 1 . (1 0)
L : In-do-ne-sia ta-nah a-ir be-ta
```

Fig. 5. The begining of *Indonesia Pusaka* song as input

The soprano (S) line is associated with the lyric line; that is, each note in the S line corresponds to a syllable in the lyric line. For instance, the pairs are (5, ’In’), (1, ’do’), (3, ’ne’), (1, ’sia’), (5, ’ta’), (1, ’nah’), (3, ’a’), (6, ’ir’), (5, ’be’), and (3, ’ta’). Each of these pairs is then converted into MBROLA codes.

Take, for example, the pair (5, ’In’) that uses ‘do’ as C4, and note 5 (sol) is equivalent to G3, which is 196 Hz. When using a full note with a duration of 400 ms, the resulting pho format is as follows:

```
I 360 100 196.00
n 40 100 196.00
```

For the remaining pairs, such as (1, ’do’), (3, ’ne’), (1, ’sia’), (5, ’ta’), (1, ’nah’), (3, ’a’), (6, ’ir’), (5, ’be’), and (3, ’ta’), the following MBROLA codes are as follows:

```
d 40 100 261.63
Q 360 100 261.63
n 80 100 329.63
e 720 100 329.63
s 40 100 261.63
i 40 100 261.63
V 320 100 261.63
t 40 100 196.00
V 360 100 196.00
n 40 100 261.63
V 320 100 261.63
h 40 100 261.63
_ 20
V 400 100 329.63
I 360 100 440.00
r 40 100 440.00
b 160 100 392.00
e 1440 100 392.00
t 40 100 329.63
V 360 100 329.63
```

G. Example of an output

The result generated by this model is an audio file in WAV format. Figure 6 shows its visual representation.

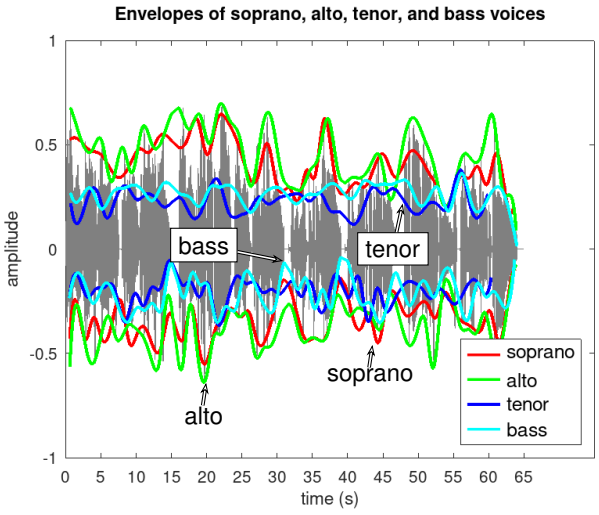

Fig. 6. The envelopes of the soprano, alto, tenor, and bass voice components in *Indonesia Pusaka*

Figure 6 displays the amplitude envelopes of the SATB voices.

Comparing these envelopes, evidently each SATB vocal component possesses a unique amplitude envelope, reflecting

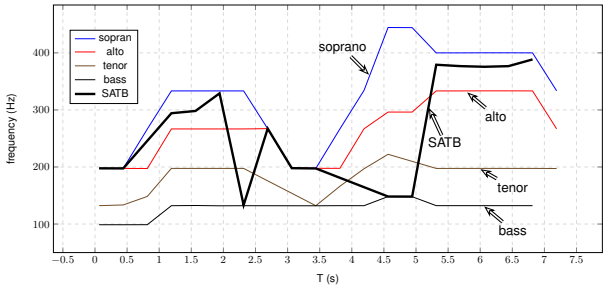

Fig. 7. A fundamental frequency the first line of *Indonesia Pusaka*

its specific pitch range and role within the choral composition. These amplitude envelopes play a crucial role in understanding the choir's dynamics and harmonies.

In Figure 7, the soprano tone exhibits a higher pitch than the other tones, including the combined SATB voice. This characteristic is particularly noticeable at the beginning of the song *Indonesia Pusaka*.

III. Result and Discussion

The study examined multiple songs, including *Indonesia Pusaka* and *Si Patokaan*. *Indonesia Pusaka* was composed by Ismail Marzuki, and the SATB arrangement was performed by N. Simanungkalit. The musical notation for this song is referenced from [26], and it is presented in Figure 10. A reworked version of the song is shown in Figure 12.

This particular song is structured into four lines, with the shortest note duration being half of a quarter or sixteenth note. Notably, there is a single legatura arch in the second line within the tenor voice. The audio analysis results confirm the four distinct voices within the song.

The fundamental frequency value chart (F_0) for *Indonesia Pusaka* is depicted in Figure 8 and 9 show the fundamental frequency value charts for *Indonesia Pusaka* and *Si Patokaan*, respectively.

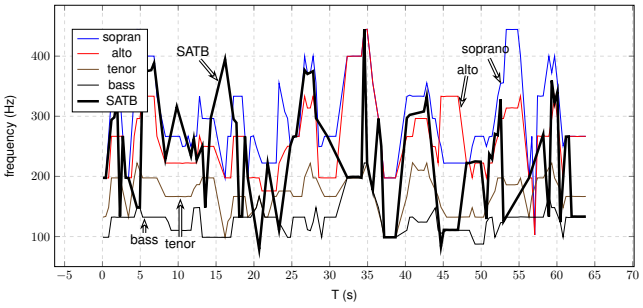

Fig. 8. Fundamental frequency of *Indonesia Pusaka*

Comparing the average values of the individual SATB voices to the combined one shows that they do not haveperfect alignment. This disparity is noticeable when examining the SATB score and average values in Table II. Hence, merging the individual voices does not inherently yield an F_0 that is the simple average of the individual frequencies or count of the combined components.

This discrepancy in the F_0 values can be attributed to the calculation or determination of the fundamental frequency (F_0), which is influenced by the width of the analysis window. The term "window width" refers to the duration of

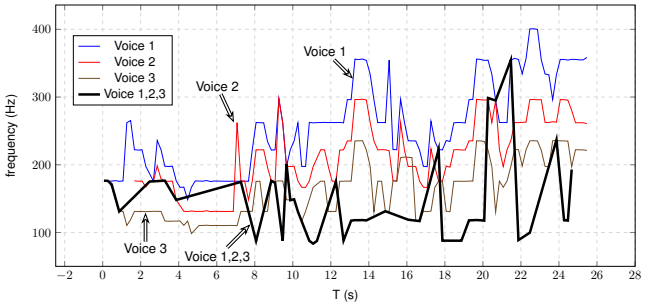

Fig. 9. Fundamental frequency of *Sipatokaan*

the audio signal segment being analyzed at any given time. The choice of window width can significantly affect how F_0 is estimated or computed, leading to variations between the merged SATB sound and its average value compared with those of the individual voices.

Indonesia Pusaka comprises four lines with diverse note durations, including the smallest value being quarter of a quarter or one-sixteenth. An interesting feature is the presence of a legato arch in the tenor voice during the second line. The audio analysis revealed the four distinct voices in the composition. Examination of the fundamental frequency value chart for *Indonesia Pusaka* indicates that the average values of the individual SATB voices do not precisely match the combined voice. This discrepancy is attributed to the calculation of F_0 as influenced by the window width.

TABLE II
Comparison between SATB voice and averages of S, A, T, and B voices

	S	A	T	B	SATB	Averages
1	197.53	197.53	132.23	98.8	197.64	156.52
2	197.49	197.48	133.47	98.76	197.52	156.8
3	333.34	266.68	197.53	132.16	294.22	232.43
4	333.33	266.67	197.53	132.49	298.18	232.51
5	333.33	266.67	197.53	131.97	329.04	232.37
6	333.33	266.67	197.53	132.15	133.22	232.42
7	197.59	197.63	132.12	132.01	197.46	164.84

Another song that was examined was *Si Patokaan*, which originated from Minahasa and was arranged by Paul Widyawan. This composition comprises three distinct voice types, referred to as voices 1–3, respectively. The lyrics of these three voices do not consistently synchronize because of the application of canon voice technique by Paul Widyawan. In the first line, voice 1 initiates, followed by voice 3, with voice 2 joining later. In the second line, all three voices starts simultaneously, but voice 1 diverges by the 4th measure. In the final two lines, the third voice is sung in unison.

The smallest note duration in this composition is also a quarter of a quarter note (1/4) or one-sixteenth (1/16). The second line of the second measure has two legato passages and the second line of the 4th measure has one triplet rhythm. The combined auditory result consists of three distinct and discernible sounds. Figure 8 shows the graph representing the fundamental frequency values for this song.

In brief, this composition comprises three voice types (voices 1, 2, and 3) with lyrics that do not consistently synchronize due to the use of canon voice technique. The song's structure involves different voice entries in the first line, simultaneous starts in the second line, and subsequent

divergences. The composition uses quarter and sixteenth notes, features legato passages, and includes a triplet rhythm. The analysis yields three distinct and discernible sounds.

Overall, this study provides insights into the musical intricacies of *Indonesia Pusaka* and *Si Patokaan*, shedding light on the presence of multiple voices, note durations, and techniques that affect the perception of harmony and melody in these compositions. Additionally, the findings emphasize the importance of considering window width, which significantly influences the interpretation of audio data, in the analysis of fundamental frequency.

IV. Conclusion

Choir singing voices are created by blending the individual elements of each voice type. The fundamental frequency observed in the combined voices does not align with the average fundamental frequency of each individual voice component. However, the outcome of merging these voices remains distinct and clear, allowing for the distinction of each component's voice.

Appendix

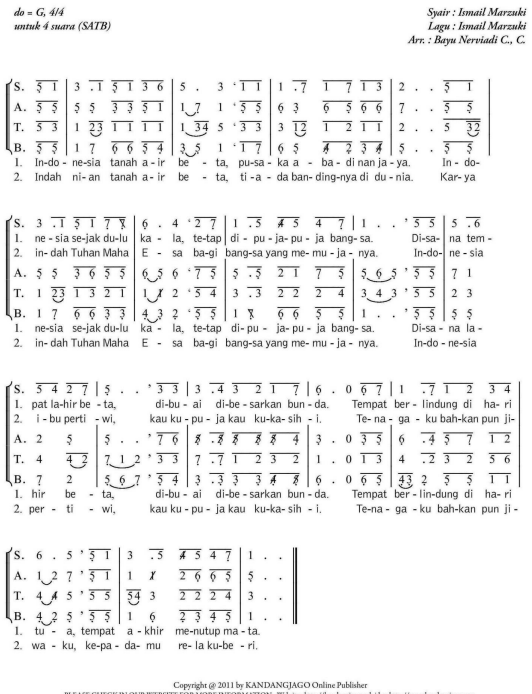

Fig. 10. *Indonesia Pusaka*

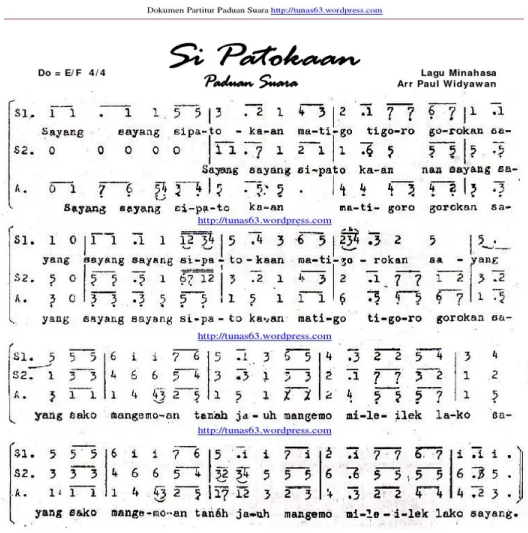

Fig. 11. *Sipatokaan*

```
S.#(5, 1)| 3 (. 1) (5, 1) (3 6)|5 . (3 0) (1 1)| 1 (. 7,) (1 7,) (1 3)| 2 . 0
A.#(5, 5,)| 1 (. 1)(5, 5,)(1 2)|3 . (1 0) (6, 6,)|6, (. 6,) (6, 6,)(6, 1)| 7, . 0
T.#(1 2)| 5 (. 3) (1 3) (5 6)|5 . (5 0) (3 3)| 3 (. 3) (3 3) (4/ 4/)| 5 . 0
B.#(5, 5,)| 1 (. 1) (1 1) (1 2)|1 . (1 0) (1 7,)| 6, (. 6,) (6, 6,) (2 2)| 5, . 0
L: In-do-ne-sia ta-nah a-ir be-ta pu-sa-ka a-ba-di nan ja-ya

S.#(5, 1)|3 (. 1) (5, 1) (7, 6/,)| 6, . (4 0) (2 7,)|1 (. 5) (4/ 5) (4 7,)| 1 . 0
A.#(5, 5,)|1 (. 1)(5, 5,) (5, 5,)|4, . (6, 0)(7, 7,)| 1 (. 3)(2/ 3)(2 5,)| 5, . 0
T.#(5, 1) | 3 (. 3) (1 3)(2 3)|{1 (2 3)} (4 0) (4 4)| 3 (. 5)(6 5) (5 4)| 3 . 0
B.#(5, 5,) | 1 (. 1) (1 1) (2 3)| 1 . (1, 0)(5, 5,)| 1 (. 1)(6, 7,) (2 5,)| 1 . 0
L: In-do-ne-sia seˆ-jak du-lu ka-la sla-lu di pu-ja pu-ja bang-sa

S.#(5 5)|5 (. 6) (5 4) (2 7,)| 5, . 0 (3 3)| 3 (. 4) (3 2) (1 7,)|6, . 0
A.#(5 5)|5 (. 6) (5 4) (2 7,)| 5, . 0 (1 1)| 2 (. 2) (1 7,) (6, 3)|3 . 0
T.#(5 5)|5 (. 6) (5 4) (2 7,)| 5, . 0 (3 6)| 5/ (. 6) (5 4) (3 2)|1 . 0
B.#(5 5)|5 (. 6) (5 4) (2 7,)| 5, . 0 (6, 1)| 7, (. 7,) (1 2) (6, 5,/)| 6, . 0
L: Di sa-na teˆ-mat la-hir be-ta Di-bu-a di beˆ-sar-kan bun-da

S.#(6, 7,)| 1 (. 7,) (1 2)(3 4)| 6 . (5 0)(5, 1)| 3 (. 5)(4/ 5)(4 7,) | 1 . . 0|
A.#(6, 5,)| 6, (. 5,)(6, 7,)(1 2)| 2/ . (3 0)(5, 5,)|1(. 1)(2 3)(2 5,)| 1 . . 0|
T.#(1 2)| 1 (. 2) (3 4) (5 5)| 4/ . (5 0) (1 3) | 5 (. 5) (6 5) (4 2) | 3 . . 0|
B.#(6, 5,)|4, (. 7,)(1 5,)(6, 7,)| 1 . (1 0)(1 1)| 1 (. 3)(2 2)(7, 5,)| 1 . . 0|
L: Teˆm-pat beˆr-lin-dung di ha-ri tu-a Sam-pai a-khir meˆ-nu-tup ma-ta
```

Fig. 12. *Indonesia Pusaka* song was rewritten

```
S.#(1 1) (. 1) 1 (5 5) | 3 (. 2) 1 (4 3) | 2 (. 1) (7, 7,) (6, 7,) | 1 (. 1)
L: Sa-yang sa-yang si pa-to-ka-an ma-ti-go ti-go-ro go-rok-an sa-
A.#0 0 0 0 | (1 1) (. 7,) 1 (2 1) | 1 (. 6,) 5, (5, 5,) | 5, (. 5,)
L: Sa-yang sa-yang si pa-to-ka-an nan sa-yang sa-
T.#(0 1) (7, 6,) [(5, 5,)] (3, 4,) | 5, (. 5,) 5, . | 4, 4, (4, 3,) (4, 2,) |
   3, (. 3,)
L: Sa-yang sa-yang si pa-to-ka-an ma-ti-go-ro go-rok-an sa-

S.#1 0 | (1 1) (. 1) 1 ([(1 2)] [(3 4)]) | 5 (. 4) 3 (6 5) | [(2 (3 4))] (. 5)
   2 5 | 5 .
L: yang sa-yang sa-yang si pa-to-ka-an ma-ti-go-rok-an sa-yang
A.#5, 0 | (5, 5,) (. 5) 1 ([(6, 7,)] [(1 2)]) | 3 (. 2) 1 (6 5) | 2 (. 1)
  (7, 7,) (1 2) | 3 (. 2)
L: yang sa-yang sa-yang si pa-to-ka-an ma-ti-go ti-go-ro go-ro-an sa-
T.#3, 0 | (3, 3,) (. 3,) 5, (5, 5,) | 1 5, 1 (1 1) | 6, (. 5,) (4, 5,) (6, 7,)
   | 1 (. 5,)
L: yang sa-yang sa-yang si pa-to-ka-an ma-ti-go ti-go-ro go-ro-an sa-

S.#. (5 5) | 6 1’ 1’ (7 6 ) | 5 (. 1’) 3 (6 5) | 4 (. 3) (2 2) (5 4) | 3 4
L: sa-ko ma-nge-ro-an ta-nah ja-uh ma-nge-ro si-le-i-lek le-ko sa-
A.#1 (3 3) | 4 6 6 (5 4) | 3 (. 3) 1 (5 3) | 2 (. 1) (7, 7,) (3 2) | 1 2
L: yang sa-ko ma-nge-ro-an ta-nah ja-uh ma-nge-mo si-le-i-lek le-ko sa-
T.#5, (1 1) | 1 4 [(4 3)] (2 5,) | 1 5, 1 (2/ 2/) | 2/ 4, (5, 5,) (5, 7,) | 1 5,
L: yang sa-ko ma-nge-ro-an ta-nah ja-uh ma-nge-mo si-le-i-lek le-ko sa-

S.#5 (5 5) | 6 1’ 1’ (7 6) | 5 (. 1’) 1’ (7 1’) | 2’ (. 1’) (7 7) (6 7) | 1’
   (. 1’) 1’ .
L: yang sa-ke ma-nge-so-an ta-nah ja-uh ma-nge-mo si-le-i-lek la-ko sa-yang
A.#3 (3 3) | 4 6 6 (5 6) | [(3 2)] [(3 4)] 5 (5 5) | 6 (. 6) (5 5) (5 5) | 6
   (. 5/) 5 .
L: yang sa-ke ma-nge-so-an ta-nah ja-uh ma-nge-mo si-le-i-lek la-ko sa-yang
T.#1 (1 1) | 1 4 [(4 3)] (2 5,) | [(1 7,)] [(1 2)] 3 (2 3) | 4 (. 3) (2 2) (4 4)
  | 4 (. 2) 3 .
L: yang sa-ke ma-nge-so-an ta-nah ja-uh ma-nge-mo si-le-i-lek la-ko sa-yang
```

Fig. 13. *Sipatokaan* song was rewritten

REFERENCES

[1] C. H. Lu and Y. L. Hu, "Speech Control System for Robot Based on Raspberry Pi," *Advanced Materials Research*, vol. 791-793, pp. 663–667, Sep. 2013.

[2] L. Cen, M. Dong, and P. Chan, "Template-based personalized singing voice synthesis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4509–4512.

[3] H.-Y. Gu and J. He, "Singing-voice synthesis using demi-syllable unit selection," *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, pp. 654–659, 2016.

[4] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, "ByteSing: A Chinese Singing Voice Synthesis System Using Duration Allocated Encoder-Decoder Acoustic Models and WaveRNN Vocoders," *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.

[5] K. Nakamura, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Fast and High-Quality Singing Voice Synthesis System Based on Convolutional Neural Networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7239–7243.

[6] C. Wang, C. Zeng, and X. He, "HiFi-WaveGAN: Generative Adversarial Network with Auxiliary Spectrogram-Phase Loss for High-Fidelity Singing Voice Generation," *ArXiv*, vol. abs/2210.12740, 2022.

[7] K. Kumar, R. Kumar, T. Boissiere, L. Gestin, W. Teoh, J. Sotelo, A. Brebisson, Y. Bengio, and A. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Oct. 2019.

[8] X. Li and Z. Wang, "A HMM-based mandarin chinese singing voice synthesis system," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 192–202, 2016.

[9] J. Zobanaky, *Galin-paris-chevé Method ...: Easy Popular Sight-singing Manual*. Creative Media Partners, LLC, 2015.

[10] J. Jonathan and Y. Suyanto, "Sintesis Suara Bernyanyi Dengan Teknologi Text-To-Speech untuk Notasi Musik Angka dan Lirik Lagu Berbahasa Indonesia," *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 10, no. 1, pp. 1–10, 2020.

[11] T. Dutoit and H. Leich, "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, no. 3, pp. 435–440, 1993.

[12] P. Banoe, *Kamus musik*. Penerbit Kanisius, 2003.

[13] J. Sundberg, *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[14] O. Tobing, P. Silitonga, and N. L. F. Gulo, "The Utilization of Candle Media to Improve Breathing Engineering on Solfeggio Choir Universitas Negeri Medan," *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, vol. 3, no. 4, pp. 2048–2055, 2020.

[15] P. Chandna, H. Cuesta, and E. G'omez, "A Deep Learning Based Analysis-Synthesis Framework For Unison Singing," *ArXiv*, vol. abs/2009.09875, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221818810

[16] M. Suharto, *Kamus Musik Indonesia*. Jakarta: Balai Pustaka, 1992.

[17] A. Sukohardi, *Teori Musik Umum*. Yogyakarta: Pusat Musik Liturgi, 1990.

[18] A. Arief, "Pembuatan Perangkat Lunak Transkrip Notasi Balok ke Notasi Angka," Skripsi, Program Studi Pendidikan Seni Musik, FBS Universitas Negeri Yogyakarta, Yogyakarta, 2009.

[19] R. Alamveta, "Pembuatan dan Aplikasi, Penulisan Notasi Balok dari File Midi," Skripsi, Universitas Kristen Petra, Surabaya, 2007.

[20] L. Chrisantyo, K. Wijana, and M. T. Restyandito, "Program Konversi Not Balok dengan Struktur Musicxml ke Not Angka," in *Seminar Nasional Teknologi*. Yogyakarta: Universitas Kristen Duta Wacana, 2007.

[21] E. Sediyono, Y. M. Susanto, and T. H. W., "Aplikasi Generator Akord dengan Menggunakan Font Notangka.ttf dan Mengadaptasi Logika Direct ProductPada Notasi Musikal Angka," *Jurnal Informatika*, vol. 10, no. 1, 2010.

[22] A. Suryarasmi and R. Pulungan, "Penyusunan Notasi Musik dengan Menggunakan Onset Detection pada Sinyal Audio," *Indonesian Computer, Electronics, and Instrumentation Support Society (IndoCEISS)*, vol. 7, no. 2, pp. 167–176, 2013.

[23] Y. Suyanto, Subanar, A. Harjoko, and S. Hartati, "An Intonation Speech Synthesis Model for Indonesian Using Pitch Pattern and Phrase Identification," *Journal of Signal and Information Processing*, vol. 05, no. 03, pp. 80–88, 2014.

[24] E. Setyati, S. Sumpeno, M. H. Purnomo, K. Mikami, M. Kakimoto, and K. Kondo, "Phoneme-Viseme Mapping for Indonesian Language Based on Blend Shape Animation," *IAENG International Journal of Computer Science*, vol. 42, no. 3, pp. 233–244, 2015.

[25] G. De Poli and N. Orio, "Music information processing," *Algorithms for Sound and*, 2006.

[26] N. Simanungkalit, *Lagu-Lagu Daerah Dan Nasional*. Gramedia Pustaka Utama, 2013.

Yohanes Suyanto (M'2015) currently serves as a faculty member and holds the position of associate professor at the University of Gadjah Mada. His primary research areas encompass text-to-speech technology, multimedia studies, and geographic information systems (GIS). He earned his undergraduate degree from the University of Gadjah Mada in 1987, followed by a master's degree from the University of Indonesia. Additionally, he completed his doctoral studies at the University of Gadjah Mada in 2014.