音声波形に含まれる周期成分に注目した母音の客観的音声解析

Objective Voice Analysis of Vowels
Focused on Periodic Components in Speech Waveform

情報工学専攻　リン ジンズゥ
Information and System Engineering　Lim Jing Zhi

Abstract: In this paper, a vowel analysis method focused on the periodic components in speech waveform and a few efforts on identifying the key elements in voices is presented. Instead of conducting analysis on fixed-size window from parts of a waveform, the proposed method extracts and analyze the vowel-like waveform, the periodic components, in the audio data. The proposed method consists of three parts that are the extraction of the cyclic waveform, normalization of the target data, and estimation of the power spectral density (PSD) of the target data. From the result, it can be seen that, although affected by the speakers' pitches, the soundwave in each cycle have significant correlation with the explicit message, i.e. the spoken vowel.
Keywords: objective voice analysis, periodic soundwave, vowel detection.

1 Introduction

Human voice is the sound caused by the vibration of air pushed from the lungs, vocal tract, and mouth by a human. Although is limited by the law of physics, it can be anything intelligible or unintelligible while a sequence of intelligible voices, means having continuous units of phonemes, is considered a speech. Phonemes consist of two categories that are voiced sound, which usually is a vowel; and unvoiced sound, which usually is a consonant. In this paper, a three-steps method focus on the voiced sound, which vibrates in a periodically pattern, with objective to objectively analyze the samples in hope to find useful information that can be effectively applied on robust automatic speech recognition (ASR) or speech synthesis (TTS) in the future is presented. To specifically identify the elements needed to effectively recognize or reconstruct human voice, the distribution and concentration of the active elements in the extracted cyclic waveform were analyzed. Following that, a unique visualization method focusing on the amplitude-frequency relation was devised and presented. The proposed method and visualization were applied on the Tohoku-Matsushita Speech Database (TMW), a Japanese word corpus. it can be seen from the visualized results that, although affected by the speakers' pitches, the spectrum has significant correlation with the spoken phoneme.

2 Theory

2.1 Objective Voice Analysis

Voice analysis (or speech analysis) is the process of analyzing the voice of human. While many research and studies have been done in the field of voice analysis, the results of the analysis often vary from case to case and from person to person.

The basic idea of the concept was gotten by exploring the nature of human voice and what is the shape of a sound. Compared to the existing analysis methods, that usually creates fixed size windows and extract various information from a soundwave, the proposed method focuses only the important part of a soundwave, where periodic component exists.

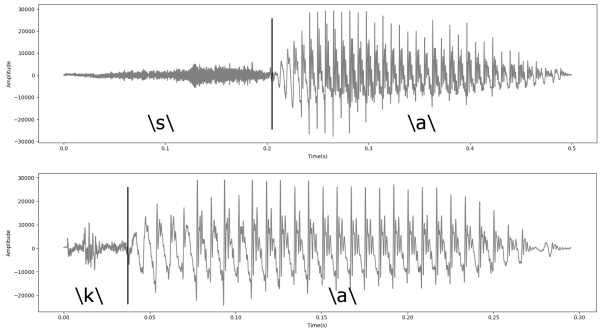

Figure 1: Waveform of Japanese phoneme. Top: \sa\ Bottom: \ka\

2.2 Nature of Human Voice

Human voice is the sound caused by the vibration of air pushed from the lungs, vocal tract, and mouth by a human. Although limited by the law of physics, it can be anything intelligible or unintelligible while a sequence of intelligible voices, means having continuous units of phonemes, is considered a speech.

Human voice, despite being a single dimension information, is still a difficult topic after all of the research and studies done in the field due to the complexity in linguistics and non-consistence definition on phonemes. For example, a listener can usually only catch and comprehends a spoken speech when the listener understands the language; a listener can comprehend and understand a spoken speech even if the speech consists a small number of inconsistent phonemes. This is because human have the ability to learn from their experience and to predict the content based on information such as the scenario of a speech, i.e. what was spoken before and after an unsure part. In the same fashion, different listeners could perceive a physically same sound as a different phoneme as long as it makes sense to the listener. In a way, without the information of what was intended by the speaker, we have no way to know what it really is. This makes automation of speech recognition a very difficult topic, where our currently best bet is by using a classifier constructed utilizing machine learning.

Human languages' phoneme mainly consists of a mixture of vowel and consonant. Vowels usually have periodic components while consonants do not have periodic components most of the time[4]. In linguistics, they are defined based on the state of the speaker's vocal tract[4]. However, because of the difficulties in objectively defining voices, it should also be defined based on their physical state at the same time to reduce the ambiguity in various studies. In this research, the focus was put on the periodic components in the speech, which is mostly the vowels, because it shows strong correlation to the spoken phonemes.

2.2.1 Periodic and aperiodic components

A sound can be categorized, at any instance, into two types of sound, i.e. sound with cyclic feature and sound without cyclic feature. Sounds with cyclic fea-

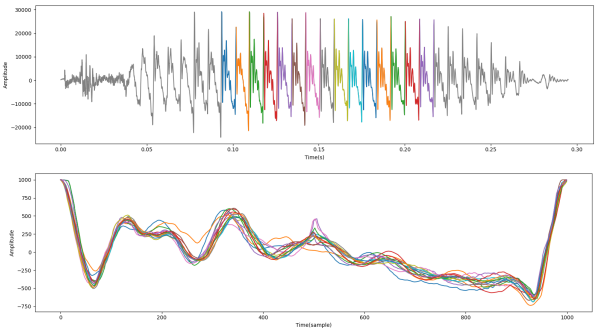

Figure 2: Waveform of the Japanese phoneme \ka\. Top: Overall waveform and periodic components representing \a\. Bottom: Normalized periodic components representing \a\ overlaid.

ture are sounds that last for a longer time instance and sounds without cyclic feature are sounds that only last for a single instance.

Figure 1 shows the raw waveform of a Japanese phoneme \sa\ and \ka\ extracted directly from two 44.1KHz, 16bit PCM, WAV data[5]. As shown in figure 1, the consonants \s\ and \k\ have close to none cyclic feature, and the vowel \a\ in both of the graphs hold strong cyclic feature as the similar pattern repeat itself until the amplitude drops.

2.2.2 Characteristics of vowels

To further understand the characteristic and nature of vowel, utilizing the three-steps method presented that automatically extract and normalize the periodic components for further analysis, it was confirmed that, in a sound instance, a vowel does indeed shows strong cyclic feature and shows repetition of a similar waveform as long as it last. However, the first and last few cycles shows inconsistent compared to the rest. It was assumed that they are unintended, but natural, mid-result from the process of conversion from a phoneme to another.

Figure 2 show waveforms of the Japanese phoneme \ka\ and the components of the vowel \a\ within. It can be seen that in the same sound instance, the vowel is shown as a repetition of a similar pattern for more than 15 times in a row. However, as shown in figure 6 the waveforms of same vowel spoken by a same speaker does not always share the same shape.

2.2.3 Human's listening ability

Compared to image processing, knowledge on the human brain in processing sound and understand language are less known. A few ad-hoc casual experiment were done to test what really is important in voice for human to comprehends voiced sounds. In the field of voice analysis, it is widely known that the important feature of a sound for human to comprehends is in their frequency distribution. To check the validity of the information, a synthesized WAV file with comprehensible contents, in a general meaning, was made and then further processed to discharge either its' dynamic amplitude, or frequency distribution.

Figure 3 shows waveform of a 4-seconds synthesized speech that was made by the software VOICEROID2[6], and waveforms of two digitally processed speech. The first graph shows the raw waveform of the synthesized speech while the second graph shows the waveform when the amplitude is digitally equalized. The third graph shows the waveform when the complex waveform is digitally substituted, and the frequency distribution is substituted by a simple sine waveform.

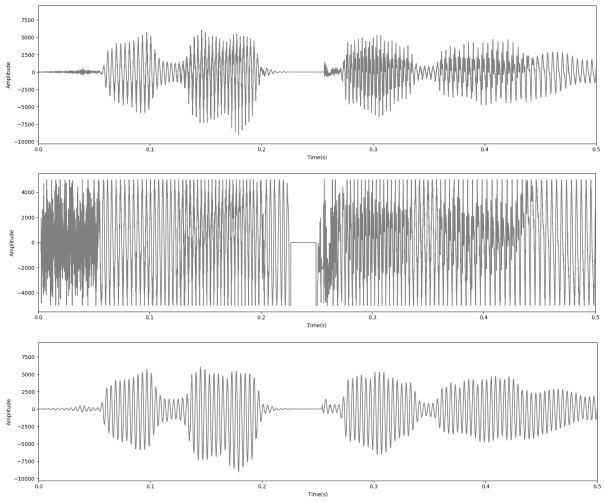

Figure 3: Waveform of synthesized speech. Top: The original waveform. Middle: Waveform after digitally equalizing the amplitude of each sample. Bottom: Waveform after substituting the complex frequencies with a simple sine wave.

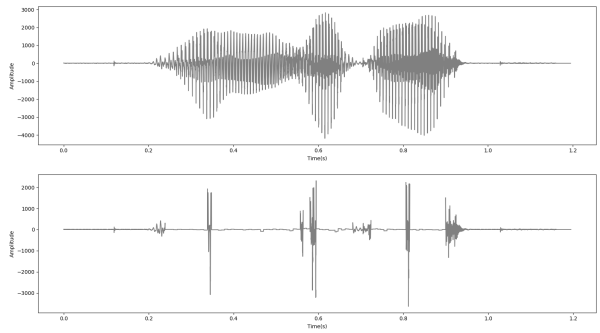

Figure 4: Waveform of human spoken speech. Top: Waveform of the speech. Bottom: Waveform of the speech after neutralizing the periodic components.

When comparing the second and third speeches, by instinct when looking at the visualized waveform, the second sound, although contain spikes, is comprehensible while the third sound is not. This shows that the information on the importance of the frequency distribution is valid and allows this research focus on the periodic components knowing it is indeed an essential element in voice.

Figure 4 shows waveforms of a human spoken speech and its' digitally processed speech where most of the periodic components in the original speech was zero-ed out leaving the aperiodic components. When listened, the processed speech was incomprehensible. This again shows that the periodic component in the speech is indeed one of the most important elements for comprehensible speech and human voice.

3 Methods

In this paper, a robust three-steps method to extract and analyze the vowel-like part of an audio data is presented. The proposed method focuses on the periodic component of sounds, which is one of the most important elements for comprehensible speeches. The proposed method consists of three major parts, each with an objective of detecting and extracting periodic component; normalizing extracted component; and analyzing normalized data

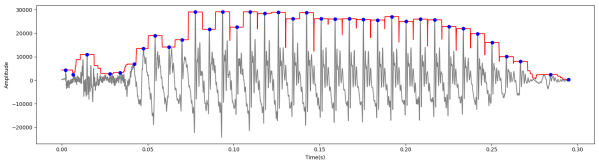

Figure 5: Waveform of Japanese phoneme \ka\ ; on detecting the periodic components.

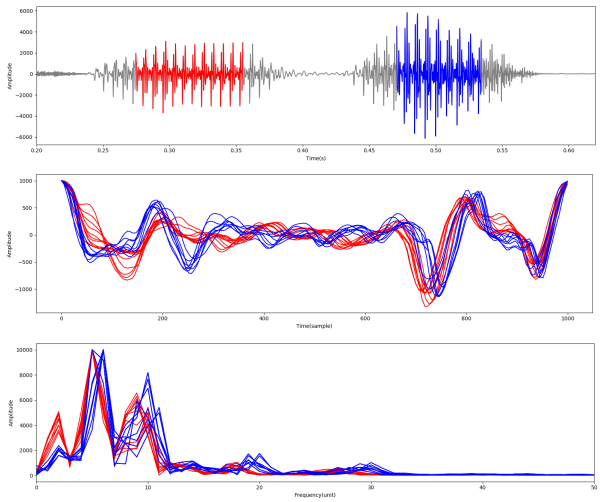

Figure 6: Waveform of human spoken Japanese word ESABA, red and blue being the first and second \a\ in the word. Top: Waveform of the speech and extracted periodic components. Middle: Normalized periodic components. Bottom: Power density spectral of the normalized periodic components.

3.1 Detecting and extracting periodic component

First, a function to extract the maximum value in a small window, of about 3 – 4 milliseconds, around each instance was prepared. Example of the result is showed as the red line in figure 5. Then, the points where the red line intersect the gray line, i.e. the sound waveform; shown as dots in the figure, is chosen as candidates of the said peaks in each cycle. Lastly, the points that share similar intervals before and after are assumed to be a periodic component of the sound and extracted for the next step. The example of the result is shown in figure 2.

3.2 Normalizing extracted component

Then, in order to make comparison of the waveforms possible, the information need to be normalized into a standardized structure. In this research, the data is normalized into an array of data consisting of 1000 points, with the highest peak valued 1000 in amplitude, for each cycle. In this research, a simple linear interpolation was utilized in normalizing the data because of the simplicity of the method. The example of the result of this step can be seen in figure 2.

3.3 Analyzing normalized data

Figure 6 shows the waveform of human spoken Japanese word ESABA zoomed in at around 0.2 second – 0.6 second. It can be seen that the same vowel, \a\ spoken at different time, at 0.25 second and at 0.45 second, has similar but not the same shape. For it to be able to be recognized as a same vowel, it has to share a set of same parameters. In this step, an analysis on the normalized data was conducted, in search of any useful information that is needed to categorize the periodic components.

In this step, the most basic power spectral density (PSD) estimation, periodogram[2] utilizing fast Fourier transform (FFT) was adopted, because of the simplicity of the method, and applied to the normalized data. The example of the result is showed in the last graph in figure 6. It can be seen that the red waveforms and the blue waveforms shows more similarities when their PSD is compared.

Table 1: Active elements needed to contain N information (Male)

Information contained (%)	Mean (Freq. unit)	SD (Freq. unit)
95	51.825	60.03
90	29.016	38.67
85	20.837	27.27

Table 2: Active elements needed to contain N information (Female)

Information contained (%)	Mean (Freq. unit)	SD (Freq. unit)
95	45.609	63.88
90	23.372	41.09
85	16.014	28.63

4 Analysis and Results

4.1 Concentration of elements

When analyzed with the proposed method, it can be observed that the frequency elements appear mostly in a low frequency range. A cropped graph focused on 0 to 50 frequency units are shown in figure 6. According to human observation, the elements in most of the waveforms extracted concentrate at the lower frequency range with exception when a consonant, such as \s\ , are extracted.

To verify, an analysis on the number of elements needed to represent partial of the original waveforms was conducted. Table 1 and table 2 are partial of the statistics of the amount of active elements needed to contain partial of the total information calculated from 5088 male and 5724 female spoken words data from the TMW database respectively.

From the result, the mean number of active elements needed to contain, and to reproduce, over 95 percent of the human voice are about 50 units. Based on this assumption, the visualization presented in the next sections focus on the first 50 frequency elements.

4.2 Visualization: Original-view

With the objective to identify the essential elements and their corresponding values in different vowels, two types of visualization were applied. The first visualization adopted the technique of spectrogram that is conventionally applied to visualize frequency histogram of audio data. The proposed visualization method processes the amplitude spectrum data computed with FFT. In order to be able to observe slight difference in the power spectrum, color map “Vega 20” is adopted.

By utilizing the “Vega 20” color map, the elements with less amplitude, such as those in the range of 20 – 30, can be highlighted and is easily noticeable. A few words containing repetition of the same vowels are chosen and shown in the main paper.

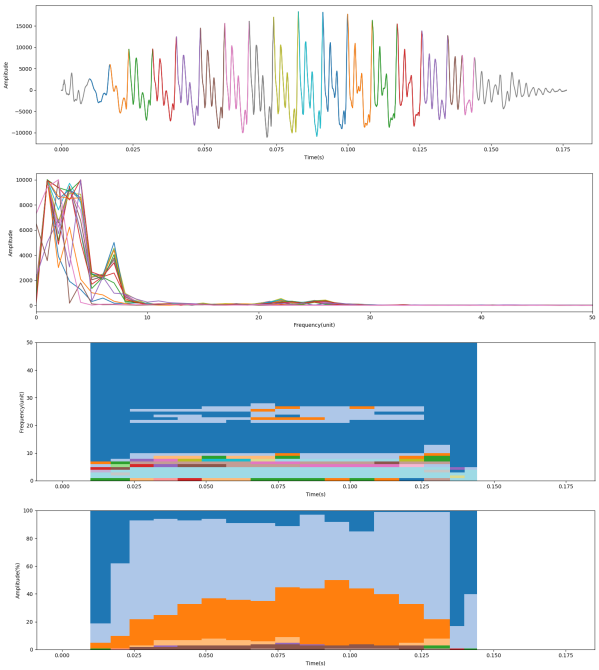

Figure 7: The Japanese phoneme \O\ Top: the original (gray) and extracted (colored) soundwaves. Second: Amplitude Spectrum of each extracted cycle. Third: Original-view. Bottom: Side-view.

4.3 Visualization: Side-view

The second visualization focuses on a different view point that can be easily overlooked with the first visualization method. The second visualization method emphasize the highest frequency on each amplitude. It can be thought as a display by contour when the spectrum is observed from the right side of the graph. Overall, the result of the side-view method shows mountain-like shaped layers. It is assumed that this is caused by the transformation of voices when spoken. The results and observation are discussed in the main paper.

5 Discussion

In this paper, a robust three-steps method with an objective to objectively analyze vowels are proposed. The proposed method extracts the vowel-like part, the periodic components, of an audio data and transform it to shape that shows more relevance to the human's listening ability for easier understanding. However, this paper focuses on the concept of paying attention to not the overall nor mean values, but the exact values of the periodic components, knowing the nature and mechanism of human voice.

Utilizing the proposed method, the essential elements in voiced sound were examined and the number of elements presumably required to recognize and reproduce recognizable sound is discussed. However, to clarify the effect of the number of active elements needed to successfully reproduce or recognize voiced sound, further experiment should be done with enough test subjects. Naturally, experiment on multiple different languages is ideal, and could be essential.

Despite the fact that the length, or pitches, of each periodic component wasn't paid much attention, it should be put on consideration in future research. Still, it can be seen that although not being perfect consistence, naturally, because of it being a phenomenon of natural actions, both of the visualization proposed shows noticeable trends useful to classify the vowels.

The proposed method gives voice analysis a new sight and proved that the points of interest focused in this paper show significant values when classifying of vowels are to be done. The proposed method is rudimentary and should be polished for best results. Other than that, overall the proposed method focuses on the periodic components opens up new research chances on voice and speech analysis. By utilizing the proposed method, voice can be synthesized, or processed in a real-time fashion.

6 Conclusion

In this paper, the nature of voiced sounds was explained. Utilizing the nature of human voice, a method focuses on the periodic components in speech waveform were proposed. The proposed method consists of three major parts respectively extract the periodic component, normalize the extracted component, and analyze the normalized data.

The proposed method was applied on a Japanese isolated spoken word corpus – Tohoku-Matsushita Speech Database, and partial of the results were selected as representative to show the trends of each vowels when analyzed with the proposed methods in the main paper. As a result, although not being perfect consistence, both of the visualization methods proposed was able to pick up noticeable trends useful to classify the vowels as they show significant correlation with the explicit message.

The proposed method is, generally, rudimentary and need to be polished before applying to actual situation. However, the concepts adapted a new point of view and shows great potential for understanding the fundamental structure and essential elements of voiced sound.

References

[1] J. Bradbury, "Linear Predictive Coding," *Online Scholarly Article,* Florida Institute of Technology, Dec. 2000. Available: http://my.fit.edu/~vkepuska/ece5525/lpc_paper.pdf [Aug. 1, 2018].

[2] S. M. Kay, S. L. Marple, "Spectrum analysis - a modern perspective," *Proc. of the IEEE,* Vol. 69, No. 11, Nov. 1981.

[3] M. Forsberg, "Why is speech recognition difficult?," Chalmers University of Technology, Feb. 2003. Available: http://www.speech.kth.se/~rolf/gslt_papers/MarkusForsberg.pdf [Aug. 6, 2018].

[4] "International Phonetic Association," Internet: https://www.internationalphoneticassociation.org/ [Jan. 21, 2019].

[5] "音声波形のサンプル-日本語の５０音の単音の波形データの例." Internet: https://wsignal.sakura.ne.jp/onsei2007/wav_data51/wav_data51.html [Jan. 21, 2019].

[6] "VOICEROID2 結月ゆかり製品情報." Internet: http://www.ah-soft.com/voiceroid/yukari/index.html [Jan. 11, 2018].

[7] "Tohoku-matsushita Speech Database," R.C.A.I.S., TOHOKU UNIVERSITY, 1989, 1990, 1991.