STUDIA PHONOLOGICA IV (1966)

Pitch Extraction by Peak Detection Method with Multi-Channels

Toshiyuki SAKAI, Shuji DOSHITA and Kou-ichi TABATA

Summary

According to the periodicity theory of hearing, every section of the basilar membrane, independent of its place, is able to perform a frequency analysis and to determine pitch by means of the periodicity of the vibrations of that section. Similarly for speech signal the outputs of a filter bank may have the periodicity corresponding to the pitch period. Pitch was extracted by combining logically the output voltages, proportional to pitch, which were obtained from each channel of the filter bank by peak detection method.

Introduction

Pitch is the subjective perception of the tone height. A sound consisting of a single objective frequency is perceived as a subjective pitch corresponding to that frequency. However, if a sound consists of many frequency components, it is not easy to make the subjective pitch correspond to its objective attribute. In case of the sound consisting of harmonics of one fundamental frequency the fundamental frequency is regarded as the "objective" pitch, but we perceive pitch in the sound without fundamental frequency (Schouten's residue).

Especially it is very difficult to extract precisely pitch frequency of "speech signal" because of the following reasons: It is not periodic, but quasi-periodic in the steady portion: It has rapid variations of the spectral constitution: Its fundamental frequency exists over a frequency range of almost three octaves: Its dynamic range is also large: Some harmonics are emphasized by the transmission characteristics of the vocal tract. Sometimes the fundamental component is very weak, while the second harmonic is stronger in the baseband.

It is said that the pitch of speech sound is decided by the periodicity of the glottal waveform of speaker, but even that exciting glottal waveform varies in shape as well as in period and amplitude.

Recently the subject of the pitch of speech signal has received increased attention in the field of speech bandwidth compression or Vocoder. The channel Vocoder which occupies a bandwidth of less than 300 cps (tenth of the ordinary telephone bandwidth) demands very precise pitch extraction and voiced-unvoiced discrimination. The naturalness of synthesized speech depends upon the small irregularities in the pitch period, although it is not easy to detect such irregularities.

Toshiyuki SAKAI, Ph.D. (坂井利之): Professor of Electrical Engineering, Kyoto University.
Shuji DOSHITA (堂下修司): Assistant Professor of Electrical Engineering, Kyoto University.
Kou-ichi TABATA (田畑孝一): Graduate School of Electrical Engineering, Kyoto University.

In spite of many efforts at automatic or mechanical pitch extractors, little commercial use has yet been reported. To overcome these difficulties, the voice-excited Vocoder appeared which transmits the unprocessed baseband (250-950 cps) of the original speech, and uses it as the excitation information at the synthesizer. This type of Vocoder is superior to the channel Vocoder in the quality or the naturalness, though its bandwidth compression ratio is only about three to one. For these reasons we must try to seek more excellent pitch extractor.

This is a report on a new type of pitch extractor which was tried as one of the efforts in such a direction.

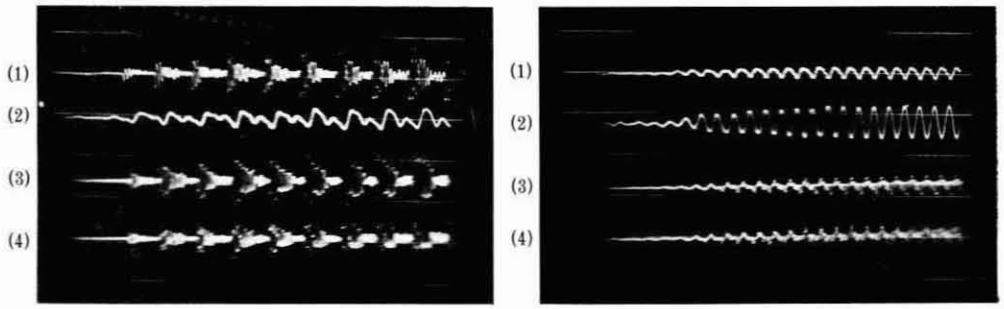
SYSTEM DESIGN

In general there are two types of automatic or mechanical pitch extractor: One attempts to measure the frequency of the fundamental component if it is present, and the other attempts to measure the time interval corresponding to the fundamental period. In the former case it is necessary for the speech signal to have the fundamental not weaker than the second harmonic in their intensities. If not, the second harmonic is extracted as pitch. This method cannot extract the minute deviation of pitch period at the steady state, and extracts the wrong pitch at the part of rapid variation such as at the head of sound, because the spectral distribution is spread and the fundamental frequency loses its meaning.

In the latter case the period is found out from the envelope of waveform. Generally the waveform of speech signal has the peaks corresponding to the glottal excitations which period is considered as pitch period. Therefore this method is called peak detection method. If the waveform of speech signal has clear peaks corresponding to the glottal excitation, this type of extractor is able not only to find the minute deviation of the pitch frequency, but also to respond instantaneously to the rapid variation. It, however, admits of improvements by reasons that peaks do not always represent the pitch especially at the tail of sound, and that this type of extractor also occassionally shows the double pitch, that is, the second harmonic frequency.

Then, what is the good extractor? The peak detection method should be adopted in order to follow the rapid variation and the fine deviation of pitch period, but such a type of exractor to date processed the raw speech signal or the baseband of speech signal in only one channel. Therefore it extracts the double pitch if the fundamental is weak and the second harmonic is strong. In Fig. 1 the waveform (2) shows the baseband waveform of the raw speech signal which is shown at the top. In this example the second harmonic is not weak in its intensity.

On the other hand, we recognize clearly the periodicity corresponding to pitch period in the waveforms shown in the rest of Fig. 1 (a) or (b). The waveform (3) and (4) are the output waveforms of the band-pass filters whose center frequ-



(a) /a/ ; male (b) /u/ ; female

(1) Original speech sound ;
(2) Output of low-pass filter with cut-off frequency 450 cps ;
(3) Output of band-pass filter with center frequency 1.25 kc ;
(4) Output of band-pass filter with center frequency 2.5 kc.

Fig. 1 Examples of outputs of filters.

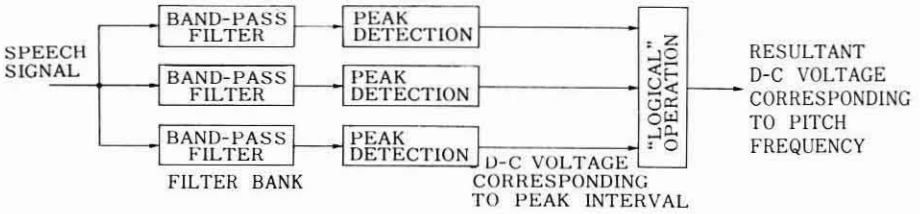
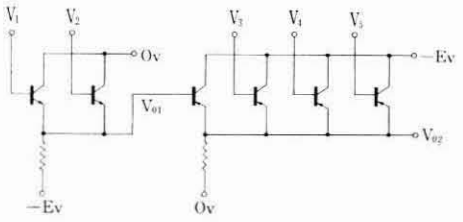


Fig. 2 Schematic diagram of the pitch extracter.



$V_{01}=\max\ (V_1, V_2)$

$V_{02}=\min\ [\max\ (V_1, V_2), V_3, V_4, V_5]$

$-E<V_1, V_2, V_3, V_4, V_5<0$

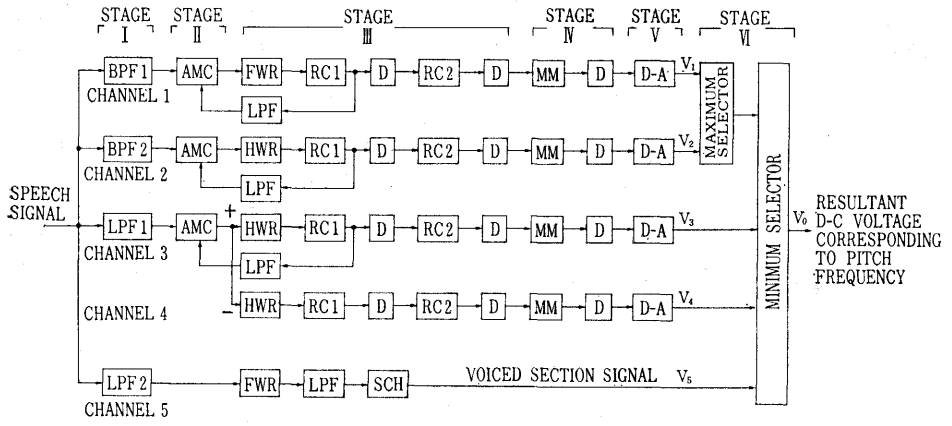
Fig. 3 An example of the "logical" network.

encies are 1.4 kc and 2.5 kc, respectively. We can apply the pitch detection method for each output. By obtaining the pitch information for each output of filter bank and by combining these pitch informations, can we expect better score in the resultant pitch extraction ?

Schematic diagram of the pitch extraction system is shown in Fig. 2. Each output of filters for speech signal is processed in the peak detection device, and d-c voltages are obtained which are proportinal to the frequency corresponding to peaks. The "logical" network processing analogue signals yields one resultant d-c voltage proportional to pitch frequency from these inputs. One example of such logical networks is shown in Fig. 3. It is chosen experimentally so as to yield the best result.

Detailed Description of the System

The block diagram of this system is shown in Fig. 4. The upper four channels (CHANNEL 1~4) are set for the pitch extraction, and the lowest channel



BPF1 ; Single Tuned Band-Pass Filter, Center 2.5 kc, Width 1.0 kc.
BPF2 ; Single Tuned Band-Pass Filter, Center 1.25 kc, Width 1.4 kc.
LPF1 ; Low-Pass Filter, Cut-Off Frequency 450 cps.
LPF2 ; Low-Pass Filter, Cut-Off Frequency 450 cps.
LPF ; Low-Pass Filter.
AMC ; Amplitude Compressor, Compression Ratio 1/2.
FWR ; Full-Wave Rectifier.
HWR ; Half-Wave Rectifier.
RC1 ; The First R-C Circuit, Decay Time Constant 4 ms.
RC2 ; The Second R-C Circuit, Decy Time Constant 4 ms.
D ; Differentiation Circuit.
MM ; Monostable Maltivibrator.
D-A ; D-A Converter.
SCH ; Fixed Level Schmitt Circuit.

Fig. 4 Block diagram of the system.

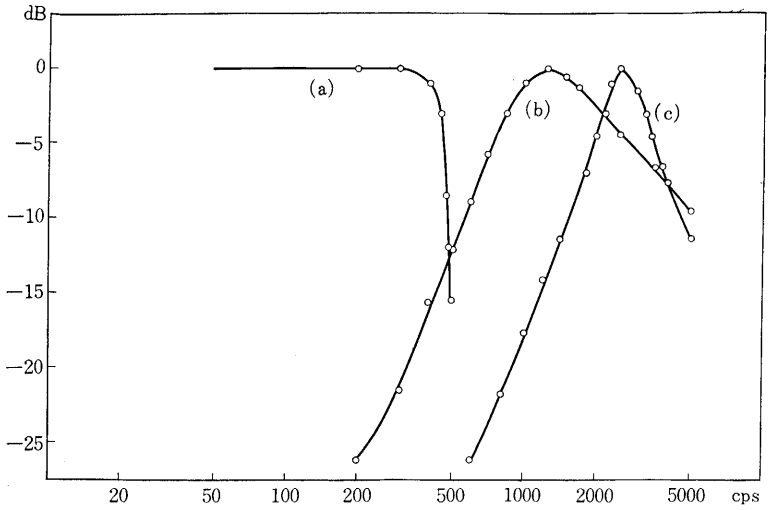
(CHANNEL 5) discriminates voiced-unvoiced sections. The voiced-unvoiced discrimination is also an important and difficult problem for the channel Vocoder, but this time it was not paid attention to so much. The time section at which the intensity of the baseband exceeds the fixed level is regarded as the voiced section.

At the stage I of Fig. 4 the filter bank was chosen experimentally so as to get a good score in the result, the frequency characteristics of which are shown in Fig. 5. Filter (a) is prepared for the baseband, filter (b) for the first or second formant, and filter (c) for higher order of formant.

At the stage II of Fig. 4 output of each filter is compressed in amplitude at the rate of two to one, which means, for example, that the change of 10 dB in output amplitude corresponds to the change of 20 dB in the input, and thereby the dynamic range is extended greatly.

At the stage III the peaks are detected by the well-known device* for peak detection. The decay time constants of the first and second R-C circuits are 4 ms. The outputs with positive and negative polarities of the low-pass filter are

*) L.O. DOLANSKY ; An Instantaneous Pitch-Period Indicator, J.A.S.A. 27, 67–72 (1955)



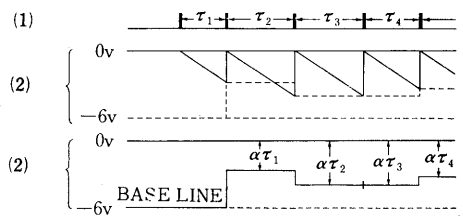
(a) Low-pass filter ; Cut-off frequency 450 cps.
(b) Single tuned BPF ; Center frequency 1.25 kc, Bandwidth 1.4 kc.
(c) Single tuned BPF ; Center frequency 2.5 kc, Bandwidth 1.0 kc.

Fig. 5 Frequency characteristics of filter bank of the system.

half-wave rectified, separately (CHANNEL 3, 4). At this stage, impulse train is obtained which shows the peak positions in the envelope.

At the stage IV the mistaken impulse which appears between the correct impulses is suppressed.

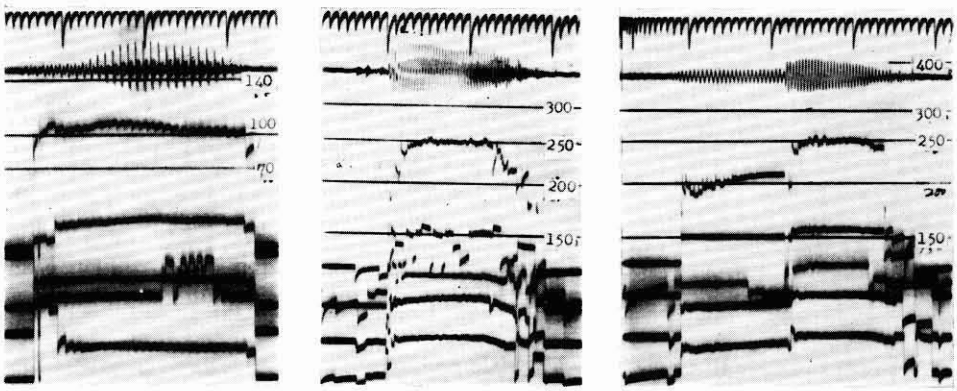
At the V the impulse train showing the peak positions is converted into the analogue d-c voltage approximately proportional to the pulse repetition frequency. The sweep generator in the D-A converter begins to sweep at the time when input impulse comes, and continues to run untill the next impulse comes (see Fig. 6). The voltage at the final point of sweep is sampled and held till the next sampling is done. Although the output voltage is negative, the higher the output voltage is algebraically, the higher the repetition frequency of input impulse train is. Allowable frequency range is from 50 to 400 cps, and if sweep waveform is chosen appropriately, the linearlity of frequency scale may be improved.



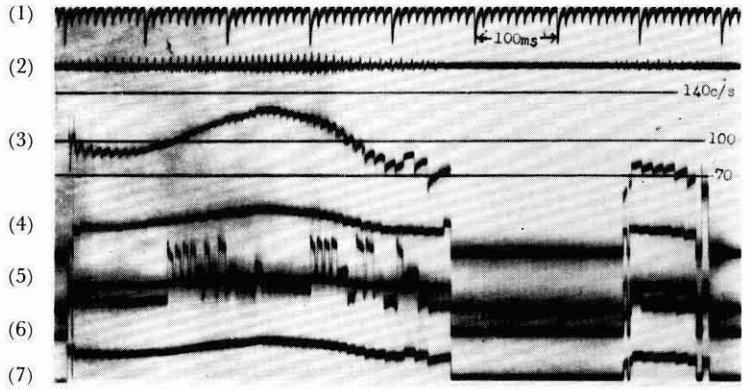
(1) Input impulse train.
(2) Sweep waveform.
(3) Output d-c voltage.

Fig. 6 Waveforms of the D-A converter.

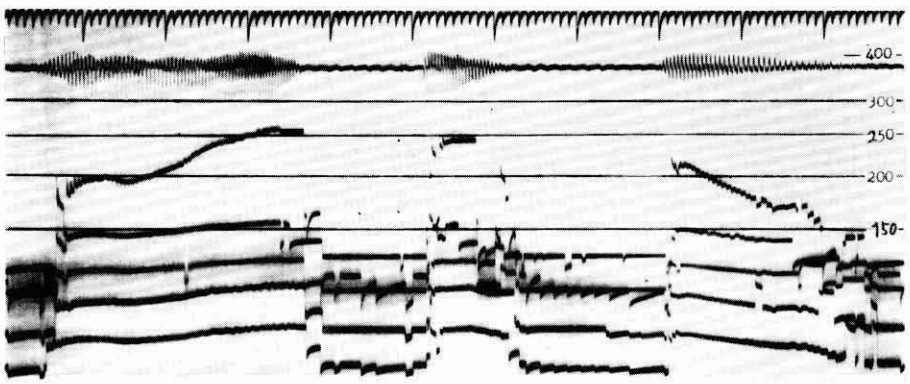
At the stage VI of Fig. 4 the "logical" operation is carried out for these voltages. First, the algebraic maximum voltage is selected in the upper two channels (CHANNEL 1, 2), and, next, the algebraic minimum voltage of this maximum and the other outputs (CHANNEL 3, 4, 5) is the resultant output corresponding to pitch frequency.



/ŋe/ ; male /ko/ ; female /bi/ ; female



/uguisu/ ; male



/jajakoʃii/ ; female

(1) Time-maker.
(2) Original speech signal.
(3) Resultant pitch frequency.
Outputs of the stage V of Fig. 4 (before processed in the logical network) from
(4) the channel with 2.5 kc BPF ; V_1 of Fig. 4 :
(5) the channel with 1.25 kc BPF ; V_2 :
(6) the channel with 450 cps LPF ; V_3 :
(7) the same as (6), but dealing with the other polarity ; V_4.

Fig. 7 Examples of pitch extraction.

Output voltage $V_0 = \min [\max (V_1, V_2), V_3, V_4, V_5]$

This logical network was chosen experimentally in order to obtain the best result.

The output of the lowest channel, voiced section indicator, fixes the resultant voltage to the base line in the sections of no sound and unvoiced sound, but it never interrupts the outputs of the other channels (CHANNEL 1~4) in the voiced sections, because the output voltage (V_5) of the lowest channel is lower than any other voltage ($V_1 \sim V_4$) for the former sections, and higher than any other voltage for the latter sections.

Pitch Extraction

Pitch frequency was extracted from the Japanese monosyllabic and conversational speech sounds uttered by male and female.

Some examples of extractions are shown in Fig. 7. These were recorded on oscillograph paper by a direct recording oscillograph device. The top is time-markers: The shorter ones show 10 ms, and the longer ones show 100 ms. The second record from above, the record (2), is the original speech signal. The record (3) is the resultant d-c voltage corresponding to pitch frequency, and this shows the final result of pitch extraction by the system. The lower four records, ((4), (5), (6) and (7)) show the d-c voltages in the four channels Fig. 4 (V_1, V_2, V_3 and V_4, respectively), from which the resultant d-c voltage is obtained, processed in the logical network. These lower four outputs, (4), (5), (6) and (7), correspond to the channels with 2.5 kc band-pass filter, 1.4 kc band-pass filter, 450 cps low-pass filter and the same 450 cps low-pass filter (but dealing with the other polarity), respectively.

The frequency scale of Fig. 7 is usuable only for the resultant d-c voltage, and for the other voltages the scales were compressed and the positions on the recorded paper were appropriately adjusted. Therefore, detailed fluctuations cannot be observed, though the extent of their contributions to the result can be evaluated. To make it easy to see, the frequency scales on the recorded paper were appropriately adjusted for the male case and for the female case, but, in fact, all the conditions in the extractor itself are the same for both cases.

Evaluation of the Results

In each example in Fig. 7 the outputs of four channels, (4), (5), (6) and (7), produced the better result, co-operating one another. In case of /bi/ the extractor could follow the rapid variation between buzz sound and vowel sound, and show the pitch precisely.

To evaluate the ability of extraction, error rate was calculated for 87 male monosyllables and 37 female ones, which is shown in Table 1. On this report, voiced section means the part in which the pitch periodicity is observed visually in the recorded waveform of the original speech signal, so that unvoiced consonants

Table 1 The error rate of pitch extraction.

	Type of sound	Total time length of voiced section (Sec)	Type of error	Error rate (Time ratio) (%)				
				Resultant pitch extraction	Output of the stage V of Fig. 4 (Before processed in the logical network)			
					BPF 2.5 kc V_1	BPF 1.25 kc V_2	LPF 450 cps V_3	LPF 450 cps (The other polarity) V_4
Male monosyllables (87)	Vowel-like sounds	15.8	Simple	8.3	5.3	9.0	18.	14.
			Burst	0.64	17.	5.6	0.64	17.
	Buzz sounds	1.3	Simple	14.	6.9	7.6	4.8	12.
			Burst	41.	76.	62.	20.	34.
	Nasal consonants	1.1	Simple	20.	15.	19.	24.	29.
			Burst	12.	20.	8.5	4.2	17.
Female monosyllables (37)	Vowel-like sounds	6.5	Simple	10.	13.	16.	5.8	4.1
			Burst	5.4	18.	5.4	0.0	0.0
	Buzz sounds	0.8	Simple	7.2	0.0	4.8	12.	7.2
			Burst	54.	100.	87.	3.0	12.
	Nasal consonants	0.4	Simple	26.	32.	20.	2.5	2.5
			Burst	16.	19.	19.	0.0	0.0

such as /k/ are not contained in the section.

The error rate is the ratio of the total time in which the extractor missed to show correct pitch to the total time of voiced sections in all examined monosyllables. But the segments of buzz sounds such as /b/, and nasal consonants such as /m/ were treated separately from vowel-like sounds.

Sometimes errors occur consecutively over several pitch periods. Such error, so to speak, burst error is treated separately at the calculation of the error rate.

The following facts are known from Table 1. (i) As for the resultant output, the error rate in vowel-like sounds (the voiced sections except for buzz sounds and nasal consonants) is about 10%. This is better than any error rate in the output of each channel before combined by the logical network, if the male and female cases are considered together. The effect of the logical operation is observed, but

the error is large in buzz sounds and nasal consonants. (ii) As for the output of each channel, the male case is good in the channel with band-pass filter, while the female case is good in the channel with low-pass filter. But in both channels the extraction in buzz sounds and nasal consonants is not good.

On the other hand, not the automatic or mechanical selection but the visual selection of the resultant pitch frequency from recorded four outputs was tried. The number of monosyllables throughout which the correct indication of pitch frequency is possible by combining the four outputs was 11 in the 87 male monosyllables, and 5 in the 37 female ones. (The ratios are about 13% in both cases).

In another system, which had only two channels with 1.7 kc low-pass filter dealing with positive and negative polarities separately, the error rate of the visual selection was about 5%. (That of the automatic selection was also about 10%.)

These facts show that the error rate of this type of pitch extractor will not become below 5% even if the best logic is used.

Conclusions

Many pitch informations were obtained from each channel of filter bank by the peak detection method, and pitch frequency was extracted from these informations, processed in the logical network.

The effect of the logical operation was observed, and the error rate was about 10%, except for buzz sounds and nasal consonants in which the error was large.

The result of visual selection of resultant pitch frequency from these informations shows that the error rate of this type of pitch extractor will not become below 5% even if the best logic is used. Moreover, none of these channels shows a good result in buzz sounds and nasal consonants.

These facts seem to suggest the limit of the extraction ability of peak detection method itself.

The perfect and precise pitch extraction is still now difficult, and it is the same with other pattern recognitions.