ROBUST PITCH DETECTION BY NARROW BAND SPECTRUM ANALYSIS

Hiroshi SHIMODAIRA and *Mitsuru NAKAI* †

School of Information Science,
Japan Advanced Institute of Science and Technology, Hokuriku,
Tatsunokuchi, Ishikawa 923-12, JAPAN
E-mail: sim@jaist-east.ac.jp
† Dept. of Information Eng., Faculty of Engineering, Tohoku University,
Sendai-shi, 980 JAPAN

ABSTRACT

This paper proposes a new technique for detecting pitch patterns which is useful for automatic speech recognition, by using a narrow band spectrum analysis. The motivation of this approach is that humans perceive some kind of pitch in whispers where no fundamental frequencies can be observed, while most of the pitch determination algorithm (PDA) fails to detect such perceptual pitch. The narrow band spectrum analysis enable us to find pitch structure distributed locally in frequency domain.

Incorporating this technique into PDA's is realized to applying the technique to the lag window based PDA.

Experimental results show that pitch detection performance could be improved by 4% for voiced sounds and 8% for voiceless sounds.

1 INTRODUCTION

Pitch is one of the most important prosodic information for automatic speech recognition and understanding. It can be used to break up speech signals into some prosodic phrases such as accent phrases [1] [2], and also used to estimate the structure of sentence [3]. Furthermore, there is some efforts of incorporating pitch information into a HMM phoneme recognizer by exploiting the correlation between pitch and spectral parameters [4].

Although many pitch determination algorithm (PDA) both in time and frequency domains were proposed, there is still no reliable and useful algorithm which can be used sufficiently for automatic speech recognition. Because many PDA's tend to fail to detect correct pitch values in low S/N sounds and unvoiced or voiceless sounds, or they are designed to discard such sounds in advance while automatic speech recognizer requires continuous pitch patterns as possible.

On the other hand, from various perceptual experiments, it is known that humans can perceive some kind of pitch even in voiceless or unvoiced sounds like in whispers, while conventional PDA's fail to detect any pitch. One of such perceptual pitch is called *periodicity pitch* or *residue pitch*. The periodicity pitch contains no energy in fundamental frequency region, but it contains high order harmonics of the fundamental frequency F_0. Using this characteristics of periodicity pitch, pitch without fundamental period can detect by analyzing high frequency regions on power spectrum.

The cepstrum method, one of the typical PDA, uses wide-band power spectrum information, it fails to detect pitch period if the harmonics of F_0 are distributed only over a particular frequency region. Because such local micro structure will be smoothed and becomes difficult to be found by wide-band analysis. This is the reason why the narrow band spectrum analysis is required. A similar study but not the same to ours is reported in [6]. The difference is that their approach is based on auditory cochlear model and correlogram analysis while ours is based on narrow band FFT spectrum analysis.

Table 1: Pitch Analysis Condition

sampling	16bit, 12kHz
time window	Blackman-Harris (32ms)
FFT	512 points (42.7ms)
frame-shift	120 points (10ms)
pitch extraction (on database)	cepstrum method & obsevation
pitch extraction (proposed)	lag-window method by narrow band spectrum analysis

This paper is organized as follows: In section 2 speech database used in our experiments and conditions for performance evaluation of pitch detection are discussed. In section 3 new pitch analysis method based on narrow band spectrum analysis is proposed and its performance experiments are reported. Section 4 gives an algorithm of integrating pitch candidates into a continuous pitch pattern and its experimental results. Conclusions are given in section 5.

2 EXPERIMENTAL CONDITIONS

2.1 Speech Database

Speech database used in this study is an ATR continuous speech database consisted of phoneme-balanced 503 Japanese sentences uttered by a single male speaker 'MYI'. The database provides phonemic transcriptions and pitch frequency information.

In the speech database, pitch is extracted by using the cepstrum method and then illegal pitch values caused from such as 'octave errors' are modified by observation. The original frame-shift of pitch analysis is 2.5 ms, but we changed to 10 ms for this study. Details of the analysis condition is shown in Table 1.

2.2 Performance Evaluation

Although pitch in voiceless and unvoiced sounds is treated 'non-pitch' and its value is set to zero in the database, correct pitch values for such sounds are required if available to evaluate our new PDA. Pitch estimation in those sounds is a kind of incon-

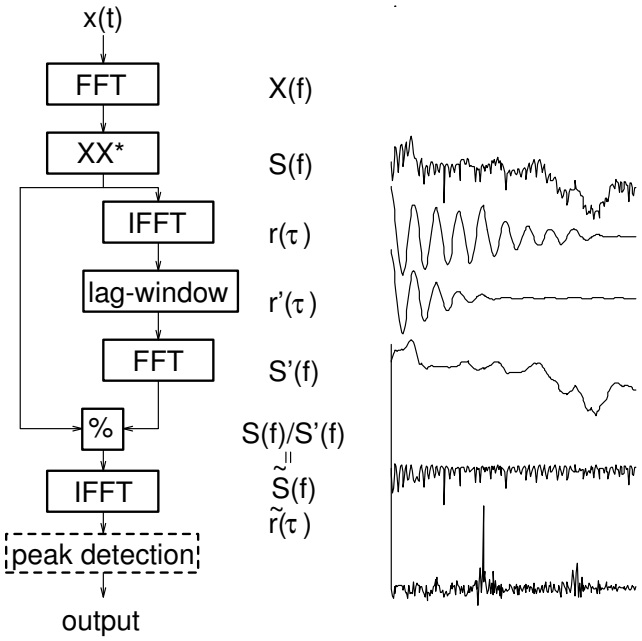

Figure 1: Block diagram of a basic pitch analysis with lag-window method

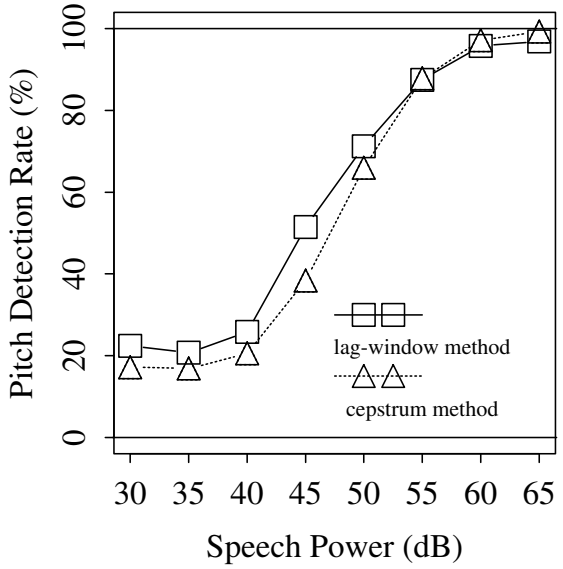

Figure 2: Performance comparison between the lag-window method and the cepstrum method

sistent problem, but we forced to estimate pitch values by interpolating pitch contours with the cubic splines.

In the following experiments in this paper, logarithm of the pitch are used, and in the performance evaluation of pitch detection, we treat a pitch detection correct if the error from the correct pitch value is within 10 % tolerance.

3 PITCH ANALYSIS

3.1 Basic Pitch Analysis

Before describing the new pitch analysis method based on the narrow band spectrum analysis, we must at first explain *the lag-window pitch extraction method* which is a basic technique for our pitch extraction.

Fig.1 shows the algorithm of the lag-window method. As is shown in the figure, we at first calculate power spectrum $S(f)$ of speech signal and then calculate its smoothed power spectrum $S'(f)$ by applying a lag-window to $S(f)$ in time domain. This is because convolution operation between $S(f)$ and spectrum smoothing window can be realized by multiplication operation in time domain between auto-correlation function derived from $S(f)$ and the lag-window obtained by IFFT of the smoothing window. Next $S(f)$ is numerically divided by the smoothed power spectrum $S'(f)$. This division produces flat power spectrum $\tilde{S}(f)$ consists mainly of harmonics of fundamental frequency F_0. Therefore inverse fourier transform of this flat power spectrum gives a clear auto-correlation function $\tilde{r}(\tau)$ of F_0. Then a pitch interval is determined simply by peak picking for $\tilde{r}(\tau)$.

The meaning of spectrum division $S(f)/S'(f)$ is to remove vocal tract effects, which are represented in the envelope of the power spectrum pattern, from observed power spectrum that is the composite of glottal information and vocal tract characteristics.

Fig.2 shows the comparison of pitch frequency detection performance between the lag-window method and the conventional cepstrum method, where the horizontal axis shows relative power of speech signal. It can be said from the figure, the lag-window method is superior to the cepstrum method if the power of speech signal is less than 50dB, while it is inferior to the cepstrum method if the power is greater than 50dB. This inferiority under high S/N environment mostly comes from what we call 'octave errors'. But it is supposed to be overcome easily by modifying illegal pitch values on the basis of continuity of pitch patterns. This is why we determined to use the lag-window method as the basic pitch determination algorithm.

3.2 Narrow Band Spectrum Analysis

3.2.1 Algorithm

As we discussed previously, it is important to consider the harmonics of fundamental frequency to realize a good perceptual pitch detector. Our approach for that is to introduce a technique of narrow band spectrum analysis which is usually used to detect some unstational periods under low S/N environments.

Fig.3 shows the block diagram of our pitch detector. The pitch detector is realized by incorporating the narrow band spectrum analysis method into the lag-window method. Therefore the algorithm is very similar to the lag-window method except to the latter half of the algorithm. In the proposed method, the flat power spectrum $\tilde{S}(f)$ obtained from dividing $S(f)$ by $S'(f)$ is split into narrow-band power spectra $\tilde{S}_k, k = 1, \cdots, K$ by applying spectrum narrow windows $W_k(f), k = 1, \cdots, K$. Here the spectrum windows have the same effective band width and they are not overlapped each other. We call these narrow bands as *01-band*, *02-band*, $\cdots$ *K-band* for short. Pitch extraction algorithm for each band is same as the lag-window method and then we have K pitch frequency candidates $P_1, \cdots, P_K$.

3.2.2 Pitch Extraction Experiment

A pitch extraction experiment was carried out to determine the number of narrow bands K. Fig.4 displays relations be-

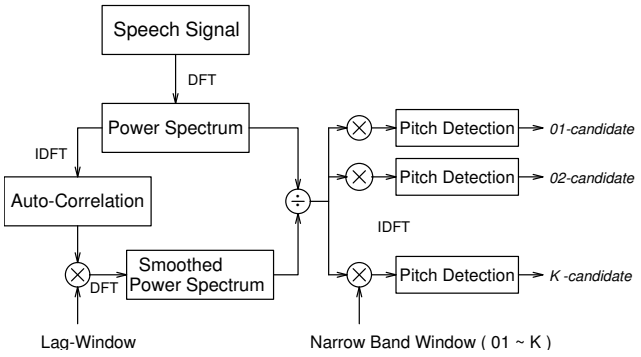

Figure 3: Block diagram of pitch extraction by narrow band spectrum analysis

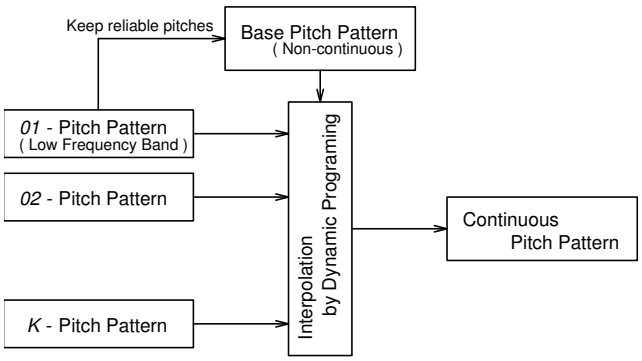

Figure 5: Integration of pitch candidates

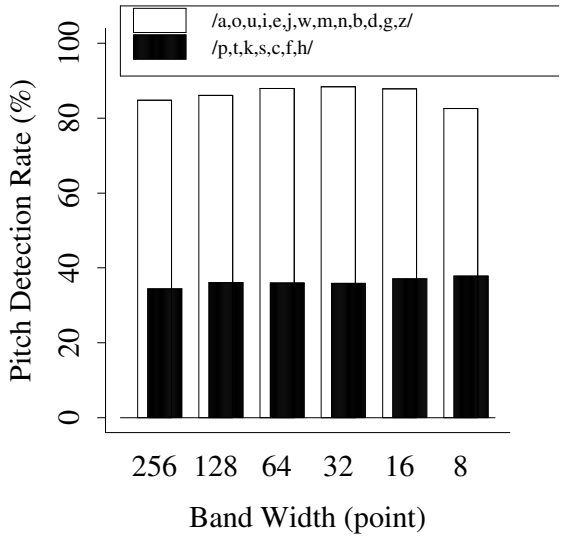

Figure 4: Difference of pitch detection rate by band widths

tween the band width [points] and pitch extraction performance for the lowest band $k = 1$. Here the 256 points means no band division ($K = 1$) and 6 kHz band width. The white bar-graphs denote the performance for voiced sounds such as /a,o,u,i,e,j,w,m,n,b,d,g,z/ and the black bar-graphs denote the performance for voiceless sounds such as /p,t,k,s,c,f,h/. It can be seen from the figure, the band width of 32-points (750 Hz wide) shows the best performance for voiced sounds while the band width of 8-points (187.5 Hz wide) shows the highest performance for voiceless sounds.

From this result, we determined to use the band width of 32 points (750 Hz wide), *ie.* $K = 8$ for the narrow band analysis. Therefore there are 8 narrow bands and 8 pitch candidates for every analyzing frame.

4 PITCH PATTERN ESTIMATION

As the result of applying the PDA to all frames of input speech, we have K pitch patterns corresponding to $01, 02, \cdots, K$ band respectively. Next we have to do is to estimate a single pitch pattern by integrating these patterns across all bands.

One possible approach is to average those values simply. But from our preliminary experiments, we found that pitch detection accuracy is so different across the bands, especially the band '01' shows the highest pitch detection accuracy, that simple arithmetic average is not supposed to work well.

The approach we took is to use a pitch pattern of the most reliable band as the basic pitch pattern, and then modify the basic pattern by using pitch candidates from other bands on the basis of pattern continuity. The 01-band is chosen as the most reliable band in our experiment. The block diagram of our pitch integration method is shown in Fig.5, and the details are described in the following four steps.

Step 1 Make the basic pitch pattern.

Step 2 Eliminate pause region from the basic pattern. Pause regions, which would arise between words or sentences, are determined by using following two parameters for energy function.

- E_m : energy threshold
- T_m : duration threshold where energy is below E_m

Step 3 Eliminate unreliable pitch candidates from the basic pattern on the basis of unstability of the candidate.

Step 4 Interpolate the pitch pattern by Dynamic Programming. Regions where the pitch pattern is not continuous in step 3 are interpolated by using other pitch candidates of the other bands 02 - 08. Dynamic programming approach is used to select the most possible candidate sequence which satisfies the minimum value of accumulated amount of pitch changes.

An example of pitch pattern estimation for a Japanese sentence "Bukkano heNdowo kouryoshite kyufusuijuNwo kimeru hitsuyouga aru" is shown in Fig.6.

Fig.7 shows pitch estimation accuracies. From left to right in the figure, each $01, \cdots, 08$ shows detection rate obtained when only a single pitch pattern of the corresponding band is used. 'MAX' shows expected maximum detection rate under the assumption that candidates closest to the correct pitch values were chosen, and the 'MP' (the right most) shows the result by the proposed method.

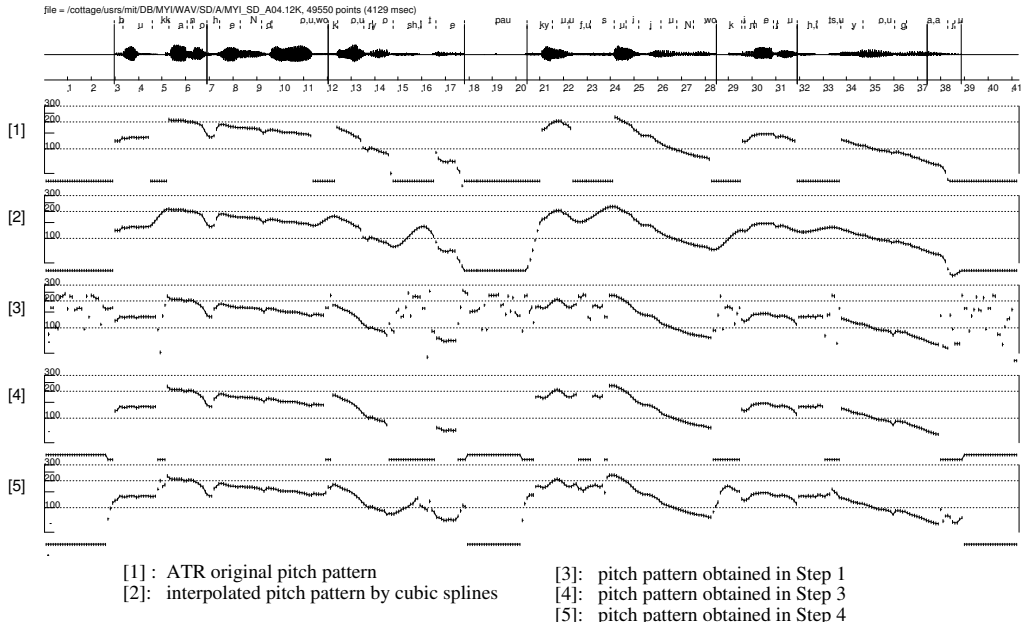

Figure 6: An example of pitch pattern estimation

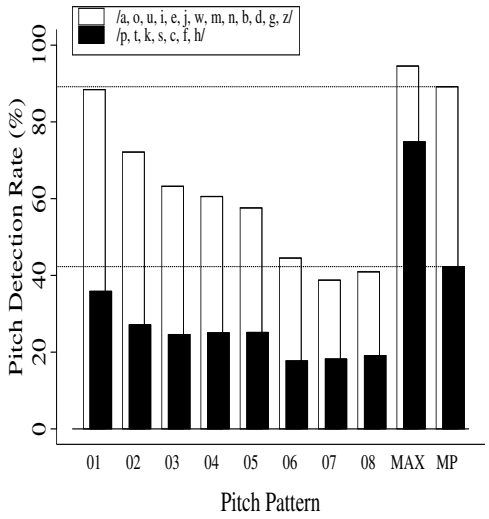

Figure 7: Performance comparison for pitch pattern estimation

It can be found from the figure that 'MP' shows improvement in 1 % point for voiced sounds and 7 % points for voiceless sounds compared to the 01-band. Comparing with the wide-band lag-window method, MP shows improvements in 4 % points for voiced sounds and 8 % points for voiceless sounds.

Furthermore the result for voiceless sounds by 'MAX' encourages us because further quite big improvement is expected if we could use more sophisticated algorithm for integrating the pitch candidates.

5 CONCLUSIONS

We have developed a novel pitch determination algorithm by incorporating a narrow band spectrum analysis technique into the lag-window methods. Evaluation experiments by using a continuous speech database showed that the proposed method is superior to the conventional algorithm based on wide-band spectrum analysis, especially for voiceless sounds. For further study, it is important to prepare more correct pitch database of perceptual pitch for accurate evaluation of the method.

The authors would like to thank Dr. H.Kanai of Tohoku Univ. for giving us useful hints on spectrum analysis.

REFERENCES

[1] W.A.Lea, M.F.Medress and T.E.Skinner: "A Prosodically Guided Speech Understanding Strategy", IEEE ASSP-23,1, pp.30–37 (1975-02)

[2] H.Shimodaira and M.Kimura, "Accent Phrase Segmentation Using Pitch Pattern Clustering," ICASSP-92, 25.17, I217-I220 (1992-03)

[3] A.Komatsu, E.Oohira and A.Ichikawa: "Conversational Speech Understanding Based on Sentence Structure Inference Using Prosodics, and Word Spotting", Trans. IEICE, J71-D,7, pp.1218–1228 (1988-07) (in Japanese)

[4] H.Singer and S.Sagayama, "Pitch Dependent Phone Modelling For HMM Based Speech Reocgnition," ICASSP-92, 36.1 (1992-03)

[5] C.W.Wightman and M.Ostendorf, "Automatic Recognition of Prosodic Phrases," ICASSP-91, S5.1, pp.321–324 (1991-03)

[6] M.Slaney and R.F.Lyon, "A Perceptual Pitch Detector," ICASSP-90, S6b.3 (1990-03)

[7] S.Sagayama and S.Furui, "A technique for pitch extraction by the lag-window method," Proc. Conf. IEICE, 1235 (1978-03) (in Japanese)