ピッチ同期 (PS)処理とピーク振幅抽出 (PA) の効果比較

ムハッマド グラム [†], 堀川順生 [‡], 新田恒雄 [‡]

豊橋技術科学大学　大学院工学研究科

〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: [†] ghulam@vox.tutkie.tut.ac.jp, [‡] {horikawa, nitta}@tutkie.tut.ac.jp

あらまし　本論分では，音声分析用帯域フィルタを FIR デジタルフィルタで構成するとともに，各出力波形からピッチに同期した（pitch-synchronous: PS）ピーク振幅（peak-amplitude: PA）を得る特徴抽出方式（PS - PA）について述べる。これまで PS – PA 方式により騒音に頑健な音声認識を実現できることを報告したが，性能に対する PS と PA の貢献の度合いについては不明であった。本文では，この解明のために行った比較実験結果を報告する。実験では，フレーム長を固定とピッチ同期で求めた場合，および振幅を rms 値とピーク値で求めた場合について比較した。実験の結果，PS と PA を同時に適用することで初めて大きな性能改善が得られること，また PS と PA の比較では PS の貢献度が大きいことが明らかになった。

キーワード　ピッチ同期分析, ピーク振幅, 特徴抽出, 頑健な音声認識

Study on contributions of pitch-synchronization and peak-amplitudes on robust ASR

Muhammad GHULAM[†], Junsei Horikawa[‡], and Tsuneo Nitta[‡]

Graduate School of Engineering, Toyohashi University of Technology,
1-1 Hibarigaoka, Tenpaku cho, Toyohashi shi, 441-8580, Aichi.

E-mail: [†] ghulam@vox.tutkie.tut.ac.jp, [‡] {horikawa, nitta}@tutkie.tut.ac.jp

Abstract　We proposed previously a novel pitch-synchronous peak-amplitude (PS-PA) based feature extraction method, which achieved significant recognition accuracy for robust ASR. It is well-known that an auditory neuron has pitch detection mechanism that can be useful for speech detection, and also peak-amplitudes in temporal pattern are robust to noise. In this paper, we conduct several experiments to find out relative contributions of pitch-synchronization (PS) and peak-amplitudes (PA) on recognition accuracy of robust ASR. Experiments include methods with fixed and pitch-synchronous frame lengths, and that with traditional peak-amplitudes and pitch-synchronous peak-amplitudes. The experimental results show that both PS and PA have strong contributions towards robust ASR and the effect of PS is higher than that of PA.

Keyword　Pitch-Synchronous Analysis, Peak-Amplitude, Feature Extraction, Noise-Robust Speech Recognition

1. Introduction

The performance of automatic speech recognition (ASR) degrades highly with increasing noise, while human beings are able to recognize even in presence of high background noise. One of the main reasons behind this difference is that an auditory system incorporates several features, which make it robust to noise. Therefore, the use of auditory-based feature extraction methods for ASR has been increased in recent years for their robustness in presence of noise. It is well known that an auditory neuron system has a pitch-synchronous mechanism [2], which can be useful for speech detection, and also peak-amplitudes are robust to noise.

We proposed earlier a pitch-synchronous peak-amplitude (PS-PA)-based feature extraction method to make a feature extractor of ASR auditory-like that uses pitch and peak-amplitude information [1]. In the proposed method, first, speech signal is passed through a bank of band-pass filters (BPFs). Second, a pitch detection algorithm (PDA) [3] detects pitch periods and determines voiced and unvoiced/silent segments. Then features are computed by extracting the highest peak in each pitch interval for each sub-band signal during voiced segments. For unvoiced/silent segments, features are extracted by averaging peaks in a frame length. In the PS-PA method, frame lengths are set proportional to pitch periods for voiced segments to make the features more pitch-synchronized, and fixed and shorter for unvoiced/silent segments. We also examined the effect of Wiener filter (WF)-based noise reduction, modulation enhancement, and auditory masking into the proposed PS-PA method [1]. These noise-suppression procedures significantly increased performance of the proposed method. In the experiments, Aurora-2J database [4] was used and results were shown using clean training only. In this paper, we present results of the proposed PS-PA method with Aurora-2J database using more meaningful multi-condition training.

From the outcomes of the experiments using the PS-PA method, we know its robustness in ASR; however, we cannot know individual effect of pitch-synchronization (PS) and peak-amplitudes (PA) on robustness issue. It seems important to know their contributions individually for their proper uses in feature extraction modules of an ASR engine.

In this paper, we carry out several experiments with different methods to find out contributions of PS and PA towards robustness of ASR. The experiments include methods with fixed frame lengths and pitch-synchronized frame lengths, and that with root mean square (RMS), traditional peak-amplitudes and pitch-synchronous peak-amplitudes. The performances of the methods are evaluated using Aurora-2J database.

The paper is organized as follows. Section 2 reviews system configuration of the PS-PA method. Section 3 shows results of the PS-PA method using Aurora-2J database with multi-condition training. Section 4 details the implementation of different methods to find out contributions of PS and PA towards robust ASR. Section 5 gives the experimental results with discussion, and finally, Section 6 draws some conclusion.

2. Review of PS-PA method

Figure 1 shows a block diagram of the proposed PS-PA method. Speech signal is, at first, passed through a bank of FIR (Finite Impulse Response) BPFs. The bands of the filters are non-overlapped and the smallest band corresponds to 80 Hz. Center frequencies of the filters are uniformly spaced on the Bark scale. A PDA [3] detects pitch periods from the output of first I_p filters (I_p = 12) and determines voiced and unvoiced/silent segments. The PDA is a robust one, for example, it has only 5% gross error rate in presence of white noise with signal-to-noise ratio (SNR) = 5 dB. In addition, the PDA is designed to give less error in detection of voiced segments as unvoiced/silent segments in order to make pitch-synchronized feature extraction not vulnerable to pitch error.

After filtering, the filtered outputs (sub-band signals) are full wave rectified. Then, for voiced segment, the highest peak (P_h) in each pitch period is extracted. Frame length is set equal to three consecutive pitch periods to ensure that no information is lost particularly for female voice, whose minimum pitch period is around 3 ms, in 10 ms frame shift. Then, an average of the logarithm of the highest peaks over the frame length is taken as weight of center frequency of current sub-band signal. An example of calculating the weight is shown in Fig. 2. For unvoiced/silent segments, frame length is set equal to two consecutive 5 ms frames (so, frame length is fixed to 10 ms), and an average of the logarithm of the highest peaks in the two 5 ms frames is taken as the weight. This kind of variable frame length can be described as an adaptation to the frequency/time resolutions depending on the spectral and temporal characteristics of the signal being processed. A long frame length is suitable for input signals whose spectrum remain stationary or varies slowly with time, such as quasi-steady state voiced regions of speech. On the other hand, a shorter frame length, processing greater time resolution, is more desirable for signals that are changed rapidly in time, such as unvoiced regions or transition between unvoiced to voiced region.

The proposed PS-PA method involves pitch-synchronization,

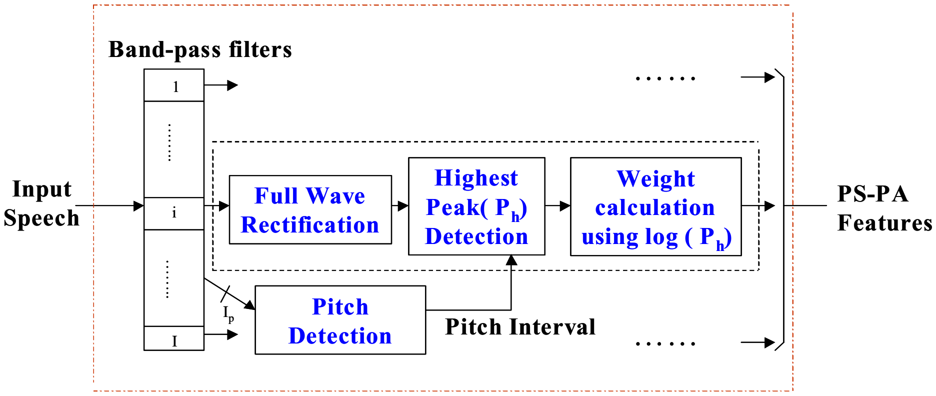


Fig.1 Block diagram of the PS-PA method.

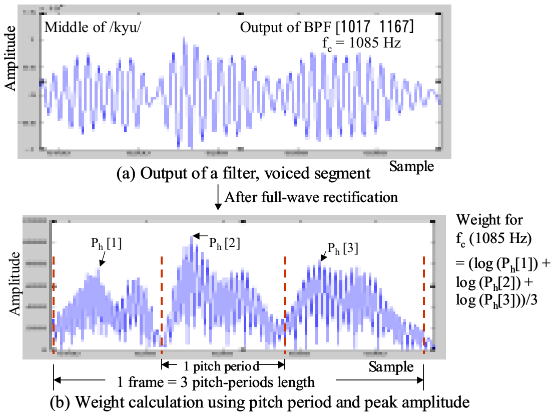


Fig. 2 Illustration of weight calculation of a sub-band signal for voiced segment.

which is an important phenomenon of auditory system. It has been reported that discharges in the auditory nerve fibers of the squirrel monkey in response to a pure tone occur at intervals which group around integral multiple of the period of the tone regardless of the best frequency of the neuron or the intensity of an effective stimulus [5]. It is also shown that discharges are locked to the cycle of component frequency of the tone [5]. These findings suggest that each period contains important information. Moreover, peaks in temporal representation are less affected by noise. The proposed PS-PA method thereby takes the highest peak in each pitch interval to extract robust features for ASR.

3. Experiments with PS-PA method

3.1. Database

The performance of the PS-PA method is evaluated using Aurora-2J database [4]. The sampling rate is 8K Hz and the utterances are connected Japanese digit strings. For the experiments in this paper, training is performed using a multi-condition dataset. For the multi-condition training dataset, four types of noise (Subway, Babble, Car, Exhibition) are added to the clean speech in five types of SNR (Signal to Noise Ratio) [SNR = clean, 20 dB, 15 dB, 10 dB, 5 dB]. The category was 0 (no change at back-end). It can be mentioned that the performance of the PS-PA method was evaluated using only clean training in [1].

3.2. Experimental setups

Twenty FIR Hamming BPFs with center frequencies uniformly spaced on the Bark scale between 150 Hz and 3.7 kHz are used. Frame length is set equal to three consecutive pitch period lengths for voiced segments, and two consecutive 5 ms frames for unvoiced/silent segments. Frame rate is 10 ms. DCT is applied to 20 dimensional features to extract 12 cepstral features. Delta and acceleration coefficients are appended to give a total of 36 dimensional features. WF-based noise reduction, modulation enhancement, and auditory masking are implemented as described in [1]. For MFCC features, frame length is fixed to 25 ms. MFCC feature vectors are of 39 dimension including power and its delta and acceleration coefficients.

3.3. Results and discussion

The experimental results are shown in Tables 1 and 2. Table 1 shows results of the PS-PA method, and Table 2 shows that using WF-based noise reduction, modulation enhancement, and masking. Table 3 gives a summarized result with multi-condition training using MFCC and MFCC with WF-based noise reduction (complete results can be found in [4]).

Table 1: Performance of the PS-PA method.

Multicondition Training (%Acc)														
	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.98	99.91	99.96	99.90	99.94	99.89	99.91	99.98	99.88	99.92	99.94	99.83	99.89	99.92
20 dB	99.75	99.79	99.78	99.78	99.78	98.89	99.81	99.26	98.23	99.05	99.79	99.63	99.71	99.47
15 dB	99.55	99.61	99.67	99.37	99.55	97.77	98.29	97.71	95.38	97.29	99.54	98.93	99.24	98.58
10 dB	98.81	98.57	98.82	98.01	98.55	88.62	91.29	92.93	86.47	89.83	97.83	94.38	96.11	94.57
5 dB	95.52	92.76	94.15	94.31	94.19	70.54	73.56	79.51	77.98	75.40	89.12	82.54	85.83	85.00
0 dB	76.87	65.87	66.79	75.66	71.30	40.13	50.12	51.04	52.78	48.52	54.45	52.89	53.67	58.66
-5 dB	32.65	39.61	28.41	30.03	32.68	25.02	28.54	28.62	26.94	27.28	25.60	22.10	23.85	28.75
Average	94.10	91.32	91.84	93.43	92.67	79.19	82.61	84.09	82.17	82.02	88.15	85.67	86.91	87.26

Table 2: Performance of the PS-PA method with noise reduction, modulation enhancement, and masking.

Multicondition Training (%Acc)														
	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.98	99.91	99.97	99.92	99.95	99.92	99.91	99.98	99.90	99.93	99.95	99.83	99.89	99.93
20 dB	99.71	99.69	99.80	99.67	99.72	99.61	99.69	99.54	99.67	99.63	99.79	99.72	99.76	99.69
15 dB	99.32	99.58	99.70	99.42	99.51	99.03	99.23	99.10	99.07	99.11	99.60	99.53	99.57	99.36
10 dB	99.11	99.02	99.23	98.35	98.93	96.81	98.02	96.92	97.08	97.21	98.85	98.22	98.54	98.16
5 dB	95.69	96.54	96.18	95.32	95.93	85.86	91.22	90.35	92.53	89.99	94.85	92.64	93.75	93.12
0 dB	83.23	73.79	84.61	81.86	80.87	58.71	74.99	72.55	75.77	70.51	73.21	68.32	70.77	74.70
-5 dB	49.06	35.76	49.70	50.37	46.22	31.08	36.72	34.93	43.21	36.49	37.71	35.11	36.41	40.37
Average	95.41	93.72	95.90	94.92	94.99	88.00	92.63	91.69	92.82	91.29	93.26	91.69	92.47	93.01

Table 3: Performance of MFCC with multi-condition training.

	MFCC	WF-noise reduction + MFCC
Overall average accuracy(%)	85.93	91.01

Table 4: Performance of different methods (MFCC and the PS-PA) with clean training.

	MFCC	WF-based noise reduction + MFCC	PS-PA	PS-PA with noise suppression procedures*
Overall average accuracy(%)	46.17	77.98	69.98	82.42

* PS-PA with WF-based noise reduction, modulation enhancement, and masking

Results using MFCC and the PS-PA method with and without noise reduction, modulation enhancement, and masking using clean training are shown in Table 4 (summarized from [1]). From Tables 1 to 4, we can see that the PS-PA method significantly improves recognition accuracy. For example, the PS-PA method with WF-based noise reduction, modulation enhancement, and masking using multi-condition training improves overall average accuracy to 93.01% from 91.01% obtained by MFCC with WF-based noise reduction, while using clean condition it improves from 77.98% to 82.42% (Table 4). These findings demonstrate robustness of the proposed PS-PA method against colorful noises.

4. Contribution of PS and PA towards robustness of ASR

From the experimental results shown in Tables 1 and 2, we can see the robustness of the PS-PA method. In this section, we carry out several experiments to find out individual contributions of PS and PA to robust ASR. In all of the experimented methods, speech signal is passed through a bank of 20 FIR Hamming BPFs. The center frequencies are spaced on the Bark scale. Noise reduction procedure, modulation enhancement and masking are not applied.

The experiments include the following methods:

(i) *RMS method with fixed frame length*: RMS power value of each filter output in each frame is calculated and is assigned as a weight of the corresponding filter. Thus we have 20 features per frame.

(ii) *Peak-amplitude method with fixed frame length*: The maximum peak-amplitude (P_{max}) of the filter output within each frame is used rather than the RMS value. Then the logarithm of P_{max} is used as weight of corresponding filter.

(iii) *Average of peak-amplitude method with fixed frame length*: An average of logarithmic value of peak-amplitudes in each frame from each filter output is calculated and assigned as weight of the filter. Figure 3 demonstrates the method with an example.

(iv) *RMS method with pitch-synchronized frame length*: Frame length is set equal to three consecutive pitch periods for voiced segments and fixed 10 ms for unvoiced segments. Then the RMS power value is calculated for each frame. The rest is as described in (i).

(v) *The PS-PA method*: Described in Section 2. Frame lengths are pitch-synchronized, as do the peak-amplitudes.

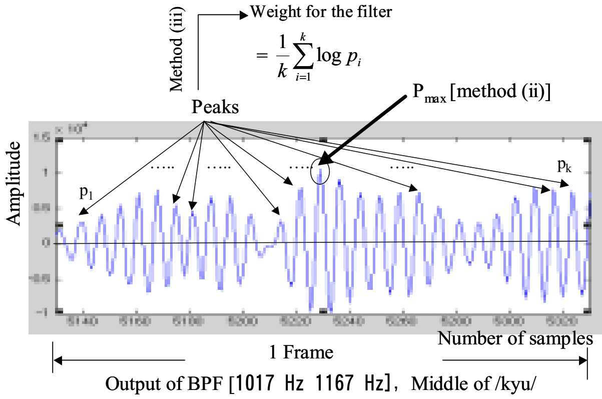


Fig.3 Demonstration of average of peak-amplitude method [method (iii) in the text]. Method (iii) uses all the peaks in a frame, while method (ii) uses only the maximum peak (encircled in the figure) in a frame.

5. Experiments

5.1. Database

Aurora-2J database is used and training is performed using clean data only.

5.2. Experimental setups

Frame length is set to 25 ms for both voiced and unvoiced segments in methods (i) to (iii) described in Section 4. After extracting 20 features per frame, DCT is applied to get 12 cepstrums. Delta and acceleration coefficients are appended to give a total of 36 vectors.

5.3. Experimental results and discussion

Table 5 shows overall average accuracy (%) of the methods. Method (i), which uses RMS power value with fixed frame, achieves performance close to MFCC. In our experiments, we consider method (i) as baseline. Relative improvements of the methods are shown in Table 6. Method (ii) that uses only the maximum peak in a frame has some degraded performance: 45.35% compared to 45.92% obtained by the baseline. Method (iii), which uses average peak-amplitudes in a frame, hardly improves performance (47.79%). It implies that PA alone has very little contribution to make ASR robust.

The usage of pitch-synchronized frame length with RMS has 7.77% relative improvement over using fixed frame length with RMS. This can be verified by recognition accuracy of method (iv) over method (i). This result suggests that pitch-synchronous frame length has some positive effect on robustness of ASR.

The proposed PS-PA method that uses pitch-synchronized peak-amplitudes and pitch-synchronized frame lengths performs the best with accuracy of 69.98%. It achieves 38.8% relative improvement over pitch-synchronized frame lengths with RMS method.

Fig. 4 graphically illustrates relative performances of the methods (ii), (iv), and (v) over the method (i). A total of 44.49% relative gain is obtained with pitch-synchronized peak-amplitudes and pitch-synchronous frame length (the proposed PS-PA method), while no gain is achieved using PA only. PA improves performance only when pitch-synchronization is applied together. With these findings, it can be concluded that both PA and PS have strong contribution towards robust ASR, and the effect of PS is higher than that of PA.

The outcomes of the experiments are interesting. Pitch-synchronization, which was previously not so much used in speech recognition stages, is proved to have a significant impact to improving performance of recognition accuracy in noisy environments. It can inspire the use of auditory functions in feature extraction modules of ASR.

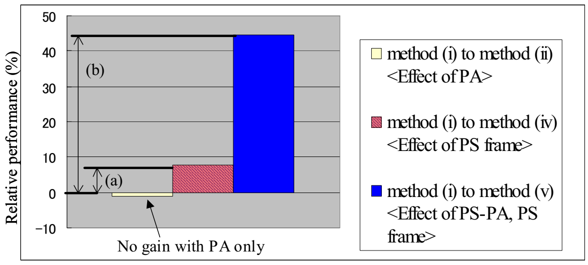
Table 5: Overall average accuracy (%) of the methods. Methods (i) to (iv) are described in Section 4, and method (v) is the proposed PS-PA method.

	MFCC	(i)	(ii)	(iii)	(iv)	(v)
accuracy (%)	46.17	45.92	45.35	47.79	50.12	69.98

Table 6: Improvements of the methods relative to method with fixed frame and RMS.

	Relative Improvement (%)
Fixed frame & PA	-1.05
PS-frame & RMS	7.77
PS-frame & PS-PA*	44.49

* The proposed PS-PA method



(a) Improvement of pitch-synchronized frame length only
(b) Improvement of pitch-synchronized frame length and pitch-synchronized peak-amplitudes

Fig.4 Graphical presentation of relative improvements between the methods.

6. Conclusion

The performance of the proposed PS-PA method was evaluated using multi-condition training data. Several experiments were conducted to find out individual effects of PA and PS to robust ASR. Experiment results indicated the dominancy of PS over PA, though both had significant positive effect towards performance.

7. References

[1] M Ghulam, et al, "A pitch-synchronous peak-amplitude based feature extraction method for noise robust ASR," In *Proc. ICASSP06*, pp. I-505-508, Toulouse, 2006.

[2] T. Hashimoto, et al, "Pitch-synchronous response of cat cochlear nerve fibers to speech sounds," *Japanese J. Physiology*, vol. 25, pp. 633-644, 1975.

[3[M Ghulam, et al, "Pitch-synchronous ZCPA (PS-ZCPA)-based feature extraction with auditory masking," In *Proc. ICASSP05*, pp. I-517-520, Philadelphia, 2005.

[4] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto and T. Endo, "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. Inf. & Sys., vol. E88-D, No.3, pp.535-544, 2005.

[5] J. E. Hind, et al, "Coding of information pertaining to paired low-frequency tones in single auditory nerve fibers of the squirrel monkey," *J Neurophysiology*, vol. 30, pp. 794-816, 1967.