A PITCH-SYNCHRONOUS PEAK-AMPLITUDE BASED FEATURE EXTRACTION METHOD FOR NOISE ROBUST ASR

Muhammad Ghulam, Junsei Horikawa, and Tsuneo Nitta

Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku-cho, Toyohashi, Japan
ghulam@vox.tutkie.tut.ac.jp

ABSTRACT

In this paper, we propose a novel pitch-synchronous auditory-based feature extraction method for robust automatic speech recognition (ASR). A pitch-synchronous zero-crossing peak-amplitude (PS-ZCPA)-based feature extraction method was proposed previously [1,2], and showed improved performance except while modulation enhancement was integrated together with Wiener filter (WF)-based noise reduction and auditory masking into it [3]. However, since zero-crossing is not an auditory event, we propose a new pitch-synchronous peak-amplitude (PS-PA)-based method to make a feature extractor of ASR more auditory-like. We also examine the effect of WF-based noise reduction, modulation enhancement, and auditory masking into the proposed PS-PA method using Aurora-2J database. The experimental results showed the superiority of the proposed method over the PS-ZCPA method, and eliminated the problem due to the reconstruction of zero-crossings from modulated envelope. The highest relative performance over MFCC was achieved as 67.33% using the PS-PA method together with WF-based noise reduction, modulation enhancement, and auditory masking.

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 [4], which can be useful for speech detection. We earlier proposed the PS-ZCPA method [1,2] that extracted pitch-synchronous features using zero-crossing and peak-amplitude information. The performance of the PS-ZCPA method was enhanced by embedding auditory masking and WF-based noise reduction into it, however, degradation in performance was observed when modulation enhancement was embedded together with WF-based noise reduction and auditory masking into the PS-ZCPA method [3], due to the problem of reconstructing zero-crossings from modulated envelope.

Moreover, because the zero-crossing is not an auditory even, we propose a new pitch-synchronous peak-amplitude (PS-PA)-based feature extraction method to make a feature extractor of ASR more auditory-like that uses pitch and peak-amplitude information and ignores zero-crossing. In the proposed method, first, speech signal is passed through a bank of band-pass filters (BPFs). Second, a pitch detection algorithm (PDA) [1,2] 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 previous method, or PS-ZCPA method, the frame lengths were proportional to inverse of center frequencies of the BPFs, and were unchanged for voiced and unvoiced/silent segments. 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.

A time-spectrum pattern processed in an auditory nerve system is influenced by various kinds of auditory effects. One of the important auditory effects is masking. From a signal processing point of view, masking enhances peaks on a time-spectrum pattern that are expected to improve the robustness in speech recognition. We implement this function of auditory masking into the proposed PS-PA method.

One of the main objectives of front-end processing in robust ASR is to preserve critical linguistic information while suppressing such irrelevant information as speaker-specific characteristics, channel characteristics, and additive noise. It has been reported that there is a strong correlation between a modulation transfer function and the intelligibility of speech [5]. Low modulation frequencies include such information as channel characteristics, speaker information, and voice quality, which are assumed not crucial for human spoken-language communication. Similarly, high modulation frequencies might be less important for spoken-language communication. In clean condition, the dominant component of the modulation spectrum of continuous speech lies between 1 Hz and 16 Hz with its peak around 4 Hz in modulation frequency [6]. To simulate this modulation into the PS-PA method, at first, envelopes are extracted from each of the BPF outputs by using peaks. Then the envelopes are filtered by a modulation filter. Finally, the modulated signals are processed using the PS-PA method.

A WF-based noise reduction procedure is also adopted in the proposed PS-PA method. In this paper, the performance of the PS-PA method with noise reduction was compared with that of the

ETSI (European Telecommunication Standards Institute) standard advanced distributed speech recognition front-end ES202 (WI008) [7] that was based on Mel-Cepstrum representation and was designed to improve recognition performance in background noise. For noise reduction step, both the PS-PA and the WI008 use a two-stage WF-based noise suppression procedure.

In this paper, Aurora-2J database [8] is used for evaluation. The performance of the proposed PS-PA method is evaluated with and without WF-based noise reduction, modulation enhancement, and auditory masking.

The paper is organized as follows. Section 2 presents system configuration of the proposed PS-PA method. Section 3 details the implementation of WF-based noise reduction, modulation enhancement, and auditory masking into the PS-PA method. Section 4 gives the experimental results with discussion, and finally, section 5 draws some conclusion.

2. SYSTEM CONFIGURATION

Figure 1 shows a block diagram of the proposed PS-PA method. Speech signal is, at first, passed through a bank of FIR 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 comparison between critical bandwidth and the bandwidth used in this paper with respect to center frequencies is shown in Fig. 2. A PDA [1,2] detects pitch periods from the output of first I_p filters (I_p = 12) and determines voiced and unvoiced/silent segments.

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. 3. 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.

Some major differences between the proposed PS-PA method and the previously proposed PS-ZCPA can be summarized as in Table 1.

3. PS-PA WITH NOISE REDUCTION, MODULATION ENHANCEMENT AND AUDITORY MASKING

3.1 PS-PA with noise reduction

The PS-PA method is subjected to be robust to noise without any kind of noise reduction procedure, however, integrating a noise reduction procedure to the method may further increase its robustness to noise. In this paper, the performance of the PS-PA method with noise reduction is compared with that of the ETSI standard Advanced DSR front-end ES202 (WI008) [7], which is based on Mel-cepstrum representation. Noise reduction of WI008

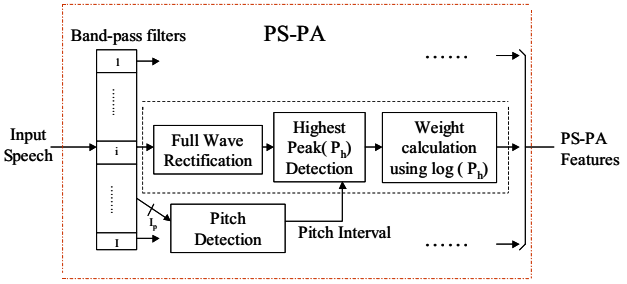

Fig. 1 Block diagram of the proposed PS-PA method

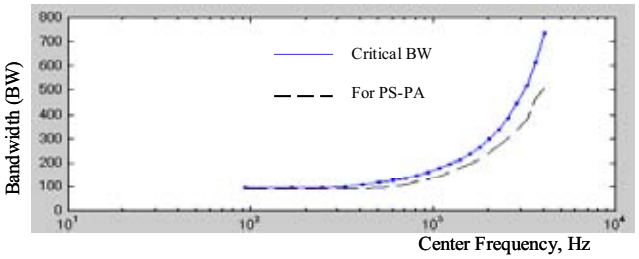

Fig. 2 Center Frequency vs. Bandwidth of the filters

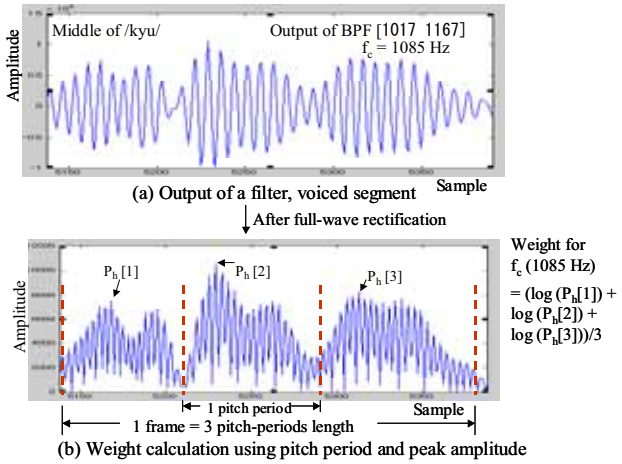

Fig. 3 Illustration of weight calculation of a sub-band signal for voiced segment

Table 1 Differences between PS-PA and PS-ZCPA methods

	PS-ZCPA	PS-PA
Frequency Calculation	Inverse of interval of zero-crossings	Center frequencies of filters
Frame Length	Inversely proportional to center frequencies of filters Same for voiced and unvoiced/silent segments	Proportional to pitch period for voiced segments Fixed and shorter for unvoiced/silent segments
Architecture Complexity	>	

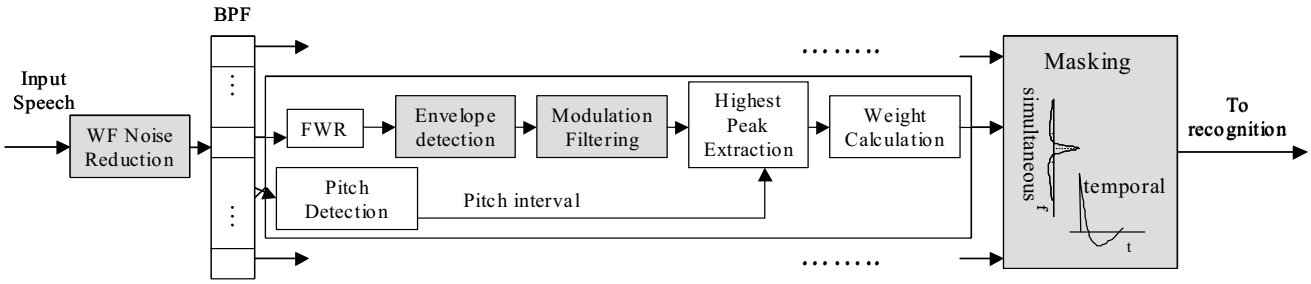

Fig. 4 Block diagram of the PS-PA method with noise reduction, modulation enhancement, and auditory masking

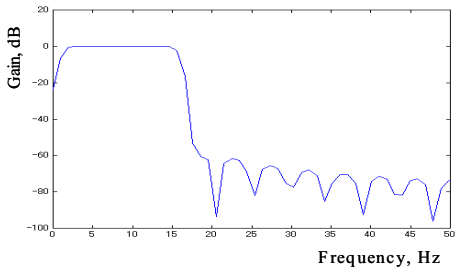

Fig. 5 Magnitude-frequency response of modulation filter

is based on Wiener filter theory and the detail algorithm is described in [7]. To implement noise reduction in the PS-PA method, the noise reduction procedure used in the WI008 is adopted without any change. First, input signal is passed through the WF-based noise reduction procedure as shown in Fig. 4. Then, de-noised signal is entered into a bank of BPFs.

3.2. PS-PA with modulation enhancement

The intelligibility of speech has a strong correlation with modulation spectrum of speech. Modulation spectrum is defined as spectral representation of temporal envelope of speech signal. One of the main objectives of front-end processing of robust ASR is to preserve critical linguistic information while suppressing some irrelevant information. Low modulation frequencies include such irrelevant information as channel characteristics, speaker characteristics, and voice quality, and high modulation frequencies might be less important for human spoken language communication. The dominant component of the modulation spectrum of continuous speech lies between 1 Hz and 16 Hz with its peak around 4 Hz in modulation frequency [6].

A process flow of modulation enhancement in the PS-PA method can be seen in Fig. 4. First, each sub-band signal is full-wave rectified. Second, an envelope is extracted by joining peaks. Then, a modulation filter is applied to the envelope. The modulation filter is designed as a 61-order FIR filter, and the magnitude characteristic of the modulation filter used in the experiment is shown in Fig. 5. The filter enhances components around 1-16 Hz and suppresses other components. After modulation filtering, PS-PA features are calculated using pitch intervals as described in section 2.

3.3. PS-PA with auditory masking

Masking is the process or amount by which the threshold of audibility of a sound is raised by the presence of another sound. There are two types of masking observed in human auditory perception: simultaneous and temporal. The equations to simulate simultaneous and temporal masking in the PS-PA method are same as used in [2]. In the experiment, both types of masking are applied; simultaneous masking is applied first to PS-PA features and followed by temporal masking.

4. EXPERIMENTS

4.1 Database

The performance of the PS-PA method is evaluated using Aurora-2J database [8]. The sampling rate is 8K Hz, and the utterances are connected Japanese digit strings. For the experiments in this paper, training is performed using clean data only, and the category was 0 (no change at back-end).

4.2. 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. The performance of the PS-PA method is evaluated with and without WF-based noise reduction, modulation enhancement (modu), and masking.

4.3. Results and discussion

The experimental results are shown in Table 2 to Table 7. Table 2 shows results of the proposed PS-PA method without noise reduction, modulation enhancement, and masking. The highest performance is achieved using the PS-PA method with WF-based noise reduction, modulation enhancement, and masking, all together, and the result is shown in Table 3. Tables 4, 5, 6 show summarized results of the PS-PA method with modulation enhancement, masking, and WF-based noise reduction, respectively. Table 7 gives relative performances of the PS-PA method with different combinations, the PS-ZCPA method, and the WI008 over MFCC.

From Table 1, we can see that the overall average accuracy obtained by the PS-PA method is 69.98% comparing to 46.17% by MFCC (not shown) and 68.98% by the PS-ZCPA [2]. The PS-PA with WF-based noise reduction achieves accuracy of 81% (Table 6), while WI008 (MFCC+WF) has accuracy of 77.98% [3],

Table 2 Performance of the PS-PA method

Clean 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	99.90	99.96	99.89	99.93	99.88	99.88	99.98	99.86	99.90	99.94	99.81	99.88	99.91
20 dB	97.91	93.20	94.02	97.32	95.61	91.50	95.71	94.53	93.36	93.78	95.01	94.14	94.58	94.67
15 dB	90.28	86.72	84.01	89.55	87.64	82.81	84.59	79.22	82.13	82.19	78.21	81.21	79.71	83.87
10 dB	81.02	74.23	76.82	78.73	77.70	79.23	77.54	75.01	74.49	76.57	73.10	70.21	71.66	76.04
5 dB	59.67	57.34	56.76	57.86	57.91	62.12	58.02	62.03	55.04	59.30	54.76	51.86	53.31	57.55
0 dB	41.39	36.88	39.28	39.20	39.19	38.85	41.21	38.78	35.20	38.51	36.07	31.01	33.54	37.79
-5 dB	31.25	29.17	27.11	25.77	28.33	24.33	28.11	28.31	26.11	26.72	24.43	21.25	22.84	26.58
Average	74.05	69.67	70.18	72.53	71.61	70.90	71.41	69.91	68.04	70.07	67.43	65.69	66.56	69.98

Table 3 Performance of the PS-PA method with WF-based noise reduction, modulation enhancement, and masking, all together

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.98	99.90	99.97	99.90	99.94	99.88	99.89	99.98	99.87	99.91	99.94	99.81	99.88	99.91
20 dB	98.91	98.95	99.95	98.87	99.17	97.06	98.79	98.88	98.93	98.42	99.23	99.04	99.14	98.86
15 dB	97.04	97.99	99.12	97.64	97.95	93.92	97.78	97.10	97.99	96.70	97.67	98.14	97.91	97.44
10 dB	91.84	92.70	97.51	93.73	93.95	86.95	92.54	90.41	91.80	90.43	89.97	92.45	91.21	91.99
5 dB	71.15	73.23	87.82	78.61	77.70	70.02	76.50	75.99	84.11	76.66	68.74	73.45	71.10	75.96
0 dB	44.95	44.89	51.92	49.34	47.78	42.51	51.16	50.81	55.82	50.08	39.96	46.94	43.45	47.83
-5 dB	33.51	30.93	29.93	28.26	30.66	26.91	29.67	29.39	27.99	28.49	26.92	23.63	25.28	28.71
Average	80.78	81.55	87.26	83.64	83.31	78.09	83.35	82.64	85.73	82.45	79.11	82.00	80.56	82.42

Table 4 PS-PA method with modulation enhancement

	A	B	C	Overall
Average	72.56	71.00	67.27	70.88

Table 5 PS-PA with masking

	A	B	C	Overall
Average	76.52	74.63	71.86	74.83

Table 6 PS-PA method with WF noise reduction

	A	B	C	Overall
Average	82.06	80.97	78.94	81.00

Table 7 Relative performances (over MFCC) of the methods

	Overall
PS-ZCPA	42.37%
PS-PA	44.23%
WI008 (WF+MFCC)	59.09%
PS-ZCPA+WF	64.29%
PS-PA+WF	64.70%
PS-ZCPA+WF+modu+mask	66.16%
PS-ZCPA+WF+mask	66.87%
PS-PA+WF+mask	67.11%
PS-PA+WF+modu+mask	67.33%

and the PS-ZCPA with WF has it 80.78% [3]. Modulation enhancement and masking also improves the performance significantly to the PS-PA method (Tables 4, 5 and [2,3]).

The best result, overall accuracy of 82.42% (Table 3), is achieved using the PS-PA method with WF-based noise reduction, modulation enhancement, and masking, all together, while the best result using the previously proposed PS-ZCPA method is 82.17% [3]. It can be mentioned that the PS-ZCPA method with WF-based noise reduction, modulation enhancement, and masking shows some degraded performance (Table 7, and [3]), due to difficulties of reconstructing zero-crossing after modulation filtering. The proposed PS-PA method has no problem concerning this, and it does not show subdued performance while using all the procedures mentioned. Moreover, architecture complexity of the PS-PA is less than the PS-ZCPA method, as there is no zero-crossing detection and histogram calculation in the PS-PA method.

From Table 7, we can see that the relative performance over the conventional MFCC of the proposed method is the best comparing to the PS-ZCPA and the WI008. The proposed PS-PA method extracts features using more pitch-synchronized manner than in the PS-ZCPA method, and frame lengths are different for voiced and unvoiced/silent segments, which make the proposed method more noise-robust.

5. CONCLUSION

A pitch-synchronous peak-amplitude (PS-PA)-based feature extraction method was proposed. The proposed PS-PA method, though simple in architecture, was proved more robust than the previously proposed PS-ZCPA method, and MFCC. WF-based noise reduction, modulation enhancement, and auditory masking were also successfully integrated into the proposed method to enhance its robustness.

Acknowledgement

This work was supported in The 21st Century COE Program "Intelligent Human Sensing", from the ministry of Education, Culture, Sports, Science and Technology, Japan.

7. REFERENCES

[1] M. Ghulam, et al, " A noise-robust feature extraction method based on pitch-synchronous ZCPA for ASR," In *Proc. ICSLP04*, pp. 133-136, 2004.

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

[3] M. Ghulam, et al, "Effect of modulation enhancement and noise reduction on PS-ZCPA based feature extraction method," In *Proc. ASJ Autumn Meeting*, pp. 81-82, 2005.

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

[5] M. R. Schroeder, "Modulation Transfer Functions: Definition and Measurement," *IEEE Trans. ASSP*, vol. 26, pp. 179-182, 1978.

[6] Arai, et al, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, vol. 105(5), pp. 2783-2791, 1999.

[7] ETSI ES 202 050 V1.1.1, "Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[8] S. Nakamura, et al, "Data collection and evaluation of AURORA-2 Japanese corpus," In *Proc. IEEE ASRU03,* pp. 619-623, 2003.