PITCH-SYNCHRONOUS ZCPA (PS-ZCPA)-BASED FEATURE EXTRACTION WITH AUDITORY MASKING

Muhammad Ghulam, Takashi Fukuda, Junsei Horikawa, and Tsuneo Nitta

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

ABSTRACT

A pitch-synchronous (PS) auditory feature extraction method based on ZCPA (Zero-Crossings Peak-Amplitudes) was proposed in [1] and showed more robust over the conventional ZCPA [2]. In this paper, we examine the effect of auditory masking, both simultaneous and temporal, into the proposed PS-ZCPA method. We also observe the effect of varying the number of histogram bins on the way to find out the optimum parameters of the proposed method. The experimental results demonstrated improved performance of the PS-ZCPA method by embedding auditory masking into it, for example, with both the masking embedded the performance increased to 73.71% from 69.92% obtained without masking for PS-ZCPA; while it showed mere improvement with increased number of histogram bins.

1. INTRODUCTION

The use of auditory-based feature extraction methods for automatic speech recognition (ASR) has been increased in recent years for their robustness in presence of noise. EIH model [3], proposed by Ghitza, uses an array of level-crossing detectors attached to the outputs of band-pass filters to generate an interval histogram. The EIH model produces dominant periodic temporal structures by analyzing zero-crossing intervals in frequency band. The ZCPA method [2], which is an improvement of the EIH model, uses peaks rather than the level-crossings to measure the intensity of each zero-crossing interval. The ZCPA method was proved more robust and computationally efficient than the EIH model.

It is well known that an auditory nervous system has a pitch-synchronous mechanism [4], which can be useful for speech detection; however neither the ZCPA method nor the EIH model utilizes this mechanism. We proposed the PS-ZCPA method [1] that extracts pitch-synchronous features by using the ZCPA method. In the ZCPA method, the positive zero-crossings in each subband are detected, and their intervals are calculated. Then a histogram of the intervals for all bands is collected with the peaks within the interval contributing as a weighting factor. In the proposed PS-ZCPA method, at first, a noise-robust, non-delayed pitch detection algorithm (PDA) is applied to extract the pitches of the speech signal, and also to detect the voiced (V) and the unvoiced or silent (U/S) segments of the signal. The highest peak ($P_{highest}$) in each pitch interval for each subband is also detected. The peaks that are above a threshold determined by the $P_{highest}$, rather than all the peaks as in the ZCPA method, are to contribute in histogram bin count. For the unvoiced or silent segments, feature are extracted same as with the ZCPA method.

A perceived histogram from ZCPA is influenced by various kinds of auditory effects. One of the important auditory effects is masking. Masking functions in such a way that a masker component inhibits other components in its vicinity. From a signal-processing point of view, masking enhances peaks on a time-spectrum pattern that are expected to bring robust speech recognition.

In [1], the superiority of the proposed PS-ZCPA method over the original ZCPA method was justified. A simple noise subtraction (NS) mechanism was also applied to enhance the performance of the proposed method. In this paper, the performance is further enhanced by incorporating auditory masking into it. The effect of changing the number of histogram bins is also observed. Moreover, a comparative study on the PDA used in the proposed method is reported using a larger dataset.

The paper is organized as follows. Section 2 presents the system configuration, where both the PDA and the PS-ZCPA method are described in short; section 3 describes the implementation detail of auditory masking into the PS-ZCPA method; section 4 gives the experimental results with discussion. Finally, section 5 draws some conclusions.

2. SYSTEM OVERVIEW

Fig. 1 shows the block diagram of the proposed PS-ZCPA method. The proposed method is divided into two parts: a) pitch determination, and voiced and unvoiced/silent segments detection, and b) feature extraction.

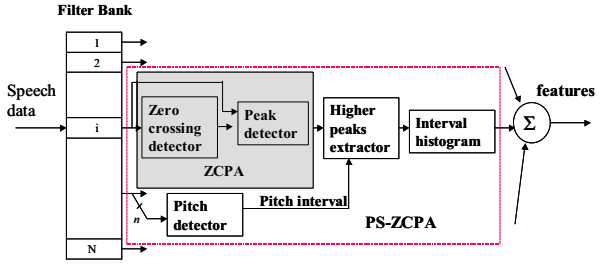

Fig. 1: Block diagram of the proposed PS-ZCPA method

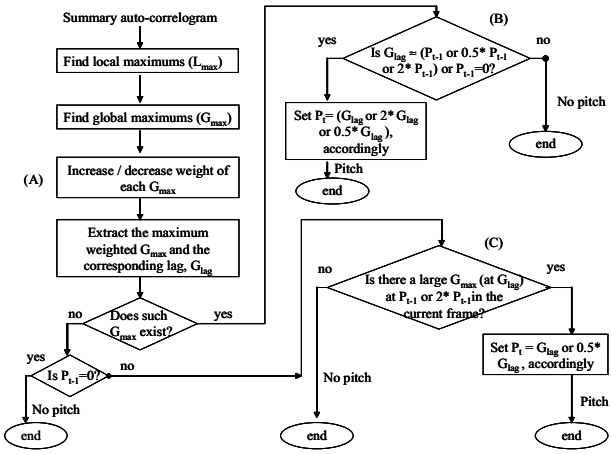

Fig. 2: Flow chart of the proposed pitch detection algorithm (PDA)

2.1. Pitch determination algorithm

A sophisticated pitch detection algorithm (PDA) was developed for the proposed method, and was described in detail in [1]. Each of the first *n* filter outputs is half-wave rectified, center-clipped, and then an auto-correlation function (ACF) is applied to give an auto-correlogram. A summary auto-correlogram is obtained by summing up all the auto-correlograms. The PDA then extracts the pitches and detects the voiced and unvoiced/silent segments of the speech signal. A flow chart of the proposed PDA is shown in Fig. 2.

In the block marked (A) in Fig. 2, the PDA increases the candidacy of a pitch candidate if there are peaks at 2nd, 3rd or 4th multiple of its lag. It also decreases the candidacy correspondingly in the absence of peaks at its multiple lags. For example, if a candidate has peak at its 2nd multiple lag, but does not have a peak at its 3rd multiple lag, then its candidacy is increased by $x/2$, and decreased by $x/3$, where x is a constant integer.

The decision block marked (B) checks for any unwanted pitch resulted from noise, or for any half-pitch or double-pitch error. If there is any half-pitch or double-pitch error, then the pitch is adjusted accordingly.

The decision block marked (C) eliminates the possibility of finding 'no pitch' towards the end of voiced segments. If a pitch, P_{t-1}, is found in previous frame, but no pitch in current frame, then this block checks whether there is a large peak in the current frame at around (or twice) the lag similar to the pitch lag in the previous frame. If such a large peak is found, then a pitch is set for the current frame.

2.2. The PS-ZCPA-based feature extraction method

The proposed PS-ZCPA method uses pitch-synchronized peaks to extract the features. The PS-ZCPA-based features are computed by the following procedure: (1) detects all the zero crossings from each filter output (subband signal), (2) calculates the inverse of the successive positive zero-crossing interval lengths that corresponds with the dominant frequencies, (3) collects histograms of the inverse zero-crossing lengths over all the subband signals, (4) increases the histogram bin count by the logarithmic value of the peak detected in between the corresponding zero crossing interval. A noise-subtraction method [1] is applied to each subband signal. The averaged noise level (P_{avg}) for each subband signal is found. For a voiced segment, the highest peak ($P_{highest}$) within a pitch period is detected. The peaks that are above some threshold are to contribute in the histogram bin count. The threshold is set to n% (n=20 in the experiment) of $P_{highest}$ for each pitch interval, or to P_{avg}. For higher SNR, where the noise level is very low, the threshold is automatically adjusted to n% of $P_{highest}$ within each pitch period, and for lower SNR, where the noise level is very high, it is automatically set to P_{avg}. For the unvoiced/silent segments, the threshold is fixed to P_{avg}. The PS-ZCPA method thereby increases the robustness by not considering the smaller peaks that are heavily affected by noise. The detail is described in [1].

3. PS-ZCPA 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 masking, and non-simultaneous (temporal) masking. Simultaneous masking is a frequency domain phenomenon where a low level signal can be made inaudible by simultaneously occurring stronger signals if the masker and the maskee are close enough to each other in frequency. The maskers affect not only the frequencies within a critical band, but also in surrounding bands (Fig. 3). A spreading function represented by a matrix $S(Z_i, Z_j)$ is to estimate the effects of masking across the critical bands. The function used in this work has been proposed in [5]:

$$S(Z_i, Z_j) = 15.81 + 7.5(Z_i - Z_j + 0.474) - 17.5\sqrt{(1 + (z_i - z_j + 0.474)^2)}$$

where Z_i is the Bark frequency of the masked signal, and Z_j is the Bark frequency of the masking signal. The PS-ZCPA generated histogram, $B(k,Z_i)$ where k is the frame

number, is then multiplied with $S(Z_i, Z_j)$ as follows:

$$C(k, Z_i) = \sum_{z_j} S(Z_i, Z_j) \times B(k, Z_i) \text{ , for all } Z_i$$

The value of $C(k, Z_i)$ denotes the spread masked histogram of Z_i-th histogram for the k-th frame of the speech.

On the other hand, the temporal masking is a function of the intensities of the masker and the probe. It is also a function of the time delay between masker and the probe. In the present work, temporal masking is implemented using the following unilateral integration model [4]:

$$y(n) = x(n) + A\sum_k \alpha^k x(n-k) - B\sum_k \beta^k x(n-k)$$

where A and B are the constants reflecting the amount of integration, and α and β are the exponential decays of the previous response and masking term, respectively.

4. EXPERIMENT

4.1. Experiments on the proposed PDA

30 isolated Japanese words spoken by 10 male speakers and 3 connected Japanese words spoken by 3 female speakers were used as test dataset. The sampling rate was 16 kHz. There were a total of 2234 frames of which 1406 were voiced and the rest were unvoiced/silent. White Gaussian noise was added to the clean speech at SNR = 10dB, 5dB and 0dB. The reference pitches were extracted manually checking the speech waveforms. The experiments were performed using the following methods:

(1) The proposed PDA

(2) Omit the enhancement blocks (B) and (C).

(3) Omit 'decrease weight' in block (A) in Fig. 2. The weight is only increased in presence of peaks in multiple lags. Also omit the enhancement blocks (B) and (C). With this it becomes a rather conventional method.

The experimental results are shown in Table 1. The results are given in the number of frames. Table 1 depicts the strength of the proposed PDA. Without enhancement blocks (B) and (C), the performance is poor. The performance was greatly affected in voiced segments, which are more important in the PS-ZCPA method. Also, the 'reduce weight' showed positive effect in the proposed PDA.

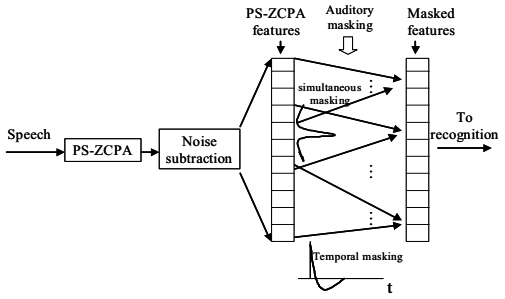

Fig. 3: PS-ZCPA feature extraction followed by auditory masking

Table 1: Erroneous no. of frames for the variation of PDA

Error	Method	SNR (dB)			
		Clean	10	5	0
Gross	(1)	4	13	31	42
	(2)	9	20	50	94
	(3)	11	24	61	118
V to U/S	(1)	4	9	26	40
	(2)	8	18	43	81
	(3)	10	21	54	93
U/S to V	(1)	5	10	14	20
	(2)	9	18	28	49
	(3)	10	21	34	58

4.2. Experiment on PS-ZCPA with masking

4.2.1. Database and experimental setup

The performance of the proposed PS-ZCPA method with auditory masking was evaluated using the Aurora-2J database [7]. The sampling rate was 8k Hz, and the utterances were connected digit strings.

20 FIR hamming band-pass filters of order 61, with center frequencies uniformly spaced on the Bark scale between 150 Hz and 3.7k Hz, were used in the experiment. Frequency range between 0 and 4k Hz was partitioned into 18 histogram bins uniformly distributed on the Bark scale. The same frequency length was again partitioned into 26 histogram bins for a comparative study. Frame length was set to $30/f_{ck}$, where f_{ck} were the center frequencies of the filters in kHz. The frame rate was 10 ms. A noise subtraction procedure was applied to the proposed method.

4.2.2. Results and discussion

The experimental results are shown in Table 2 to table 6. Table 2 to Table 5 show the results with 18 histogram bins using the PS-ZCPA method without masking, with simultaneous masking, with temporal masking, and with simultaneous and temporal masking together, respectively. A summary result using the above procedures with 26 bins, MFCC with spectral subtraction, and ZCPA with noise subtraction is given in Table 6. From Table 2 to Table 5, we can see that embedding auditory masking increases the performances of the proposed PS-ZCPA method. Simultaneous masking has lesser effect comparing to temporal masking effect. It means that the PS-ZCPA method already has some sort of spectral masking effect integrated. In a previous experiment [6], it was found that the simultaneous masking had adverse effect on the ZCPA method. But in the present experiment, it increased the performance, because a noise-subtraction procedure was applied in the PS-ZCPA method; in our experiment, the overall performance increased from 67.41% (without noise-subtraction, not shown in the tables) to 70.29% (Table 3). Table 6 indicates a little improvement using 26 bins instead of using 18 bins. Because the PS-ZCPA method has a poor frequency resolution at higher frequency regions, we cannot increase the performance by simply increasing the number of histogram bins.

Table 2: Performance of the PS-ZCPA method without masking, using 18 histogram bins

	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	99.87	99.98	99.86	99.90	99.94	99.80	99.87	99.91
20 dB	97.84	93.01	93.96	97.11	95.48	91.34	95.67	94.21	93.09	93.58	94.94	94.01	94.48	94.52
15 dB	90.19	86.09	83.12	88.76	87.04	82.20	83.24	78.11	81.93	81.37	77.85	80.88	79.37	83.24
10 dB	80.10	73.06	75.78	77.92	76.72	78.03	76.87	73.02	73.21	75.28	72.15	69.34	70.75	74.95
5 dB	58.25	55.98	54.89	56.12	56.31	60.32	56.28	60.00	53.20	57.45	53.18	50.56	51.87	55.88
0 dB	39.07	35.28	37.70	37.71	37.44	37.29	39.17	37.95	33.86	37.07	35.42	29.87	32.65	36.33
-5 dB	29.76	27.62	26.19	24.12	26.92	23.11	26.94	26.64	24.63	25.33	23.67	20.43	22.05	25.31
Average	73.09	68.68	69.09	71.52	70.60	69.84	70.25	68.66	67.06	68.95	66.71	64.93	65.82	68.98

Table 3: Performance of the PS-ZCPA method with simultaneous masking, using 18 histogram bins

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.95	99.85	99.92	99.84	99.89	99.86	99.84	99.96	99.81	99.87	99.90	99.78	99.84	99.87
20 dB	97.90	93.63	94.34	97.84	95.93	91.87	95.85	94.65	93.63	94.00	95.04	94.45	94.75	94.92
15 dB	90.99	87.32	84.93	89.32	88.14	83.45	84.67	79.82	82.76	82.68	78.99	82.08	80.54	84.43
10 dB	81.21	74.27	77.21	79.05	77.94	79.82	78.41	74.72	74.55	76.88	73.67	70.12	71.90	76.30
5 dB	59.87	57.27	56.58	58.32	58.01	62.44	57.89	61.93	55.02	59.32	55.02	52.32	53.67	57.67
0 dB	40.68	37.09	39.88	39.31	39.24	39.54	41.01	39.79	35.56	38.98	37.34	31.22	34.28	38.14
-5 dB	31.38	29.42	28.47	26.45	28.93	25.02	29.04	28.51	26.88	27.36	25.86	22.37	24.12	27.34
Average	74.13	69.92	70.59	72.77	71.85	71.42	71.57	70.18	68.30	70.37	68.01	66.04	67.03	70.29

Table 4: Performance of the PS-ZCPA method with temporal masking, using 18 histogram bins

	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	99.91	99.94	99.90	99.87	99.98	99.88	99.91	99.94	99.81	99.88	99.91
20 dB	97.93	93.92	94.92	97.96	96.18	91.95	95.95	94.81	93.82	94.13	95.51	94.82	95.17	95.16
15 dB	91.45	88.87	86.06	90.23	89.15	84.79	85.91	80.45	83.90	83.76	80.12	83.75	81.94	85.55
10 dB	82.58	77.37	79.21	80.34	79.88	81.21	80.02	76.59	76.87	78.67	77.97	72.90	75.44	78.51
5 dB	62.99	60.41	59.56	62.19	61.29	65.14	60.53	64.10	58.00	61.94	58.82	55.55	57.19	60.73
0 dB	45.75	42.22	44.67	44.21	44.21	44.32	45.82	44.36	40.46	43.74	41.44	36.48	38.96	42.97
-5 dB	36.55	35.59	33.12	31.78	34.26	30.54	35.29	34.33	31.04	32.80	32.33	28.69	30.51	32.93
Average	76.14	72.56	72.88	74.99	74.14	73.48	73.65	72.06	70.61	72.45	70.77	68.70	69.74	72.58

Table 5: Performance of the PS-ZCPA method with simultaneous and temporal masking, using 18 histogram bins

	A					B					C			Overall
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	99.97	99.89	99.94	99.89	99.92	99.88	99.86	99.98	99.86	99.90	99.93	99.80	99.87	99.90
20 dB	98.14	94.02	95.09	98.02	96.32	92.04	95.99	94.89	93.88	94.20	95.58	94.87	95.23	95.25
15 dB	92.01	89.45	87.43	90.83	89.93	85.32	86.22	81.01	84.22	84.19	80.93	83.93	82.43	86.14
10 dB	83.98	78.21	80.32	81.24	80.94	81.91	80.78	77.41	77.42	79.38	78.04	73.06	75.55	79.24
5 dB	64.21	61.78	61.33	62.94	62.57	65.85	61.13	65.54	59.41	62.98	59.34	56.12	57.73	61.77
0 dB	47.05	44.54	46.21	45.54	45.84	45.11	46.75	45.77	41.06	44.67	42.31	37.20	39.76	44.15
-5 dB	38.31	37.42	34.57	33.43	35.93	32.04	36.73	35.21	32.24	34.06	33.10	29.76	31.43	34.28
Average	77.08	73.60	74.08	75.71	75.12	74.05	74.17	72.92	71.20	73.09	71.24	69.04	70.14	73.31

Table 6: Averaged performance using 26 histogram bins

	A	B	C	Overall
MFCC (with SS)	55.72%	63.01%	39.76%	55.44%
ZCPA	67.51%	66.63%	60.81%	65.49%
PS-ZCPA	71.49%	69.91%	66.77%	69.92%
with simultaneous Masking	72.24%	70.72%	67.63%	70.71%
with temporal masking	74.48%	73.04%	70.07%	73.02%
with both the masking	75.17%	73.76%	70.71%	73.71%

5. CONCLUSION

The effect of auditory masking on the proposed PS-ZCPA method was investigated. Integrating masking effect enhanced the performance of the proposed method, however, an increase in the number of histogram bins showed little effect on it. How to overcome the limitation of the PS-ZCPA method at higher frequency region will be our future study.

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.

6. REFERENCES

[1] M Ghulam, et al, "A noise-robust feature extraction method based on pitch-synchronous ZCPA for ASR," in *Proc. ICSLP04*, Korea, 2004, to appear.

[2] DS Kim, et al, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55-69, Jan. 1999.

[3] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115-132, Jan. 1994.

[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] MR Schroeder, et al, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.* 66(16), pp. 1647-1651, Dec. 1979.

[6] http://www.bsrc.kaist.ac.kr/seminar/Auditory

[7] K. Yamamoto, et al, IPSJ SIG Technical Reports, SLP-47-19, pp. 101-106 (2003)