社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＳＬＰ－51　（３）
2004／5／27

頑健な音声認識のためのピッチ同期ZCPAに基づく特徴抽出法

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　本報告では，雑音に頑健な特徴抽出法の一つであるZCPAを改良したピッチ同期ZCPA（PS-ZCPA）を提案する．提案方式では，まず，入力音声に対して20チャネルのBPF分析を行う．次に，各BPFの出力からピッチを計算すると同時に，音声の有声・無声区間を検出する．有声区間では，1ピッチ区間における振幅の最大値からしきい値を設定し，しきい値以上のピークを持つ1周期波形に対してヒストグラムを計算する．一方，無声区間では，従来のZCPAと同様の処理を行う．雑音を重畳した孤立単語音声による評価実験から，PS-ZCPAが従来のZCPAと比較して良好な性能が得られることを示す．また，PS-ZCPAと従来の雑音除去手法を組み合わせた方式を，Aurora-2Jを用いて評価したので報告する．

キーワード　音声認識，聴神経システム，ZCPA，ピッチ同期分析

A Pitch-synchronous ZCPA-based Feature Extraction Method for Robust ASR

Muhammad GHULAM† Junsei HORIKAWA‡ and Tsuneo NITTA‡

Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tenpaku cho, Toyohashi, 441-8580, Japan

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

Abstract　In this paper, we propose a novel feature extraction method based on an auditory neuron system for robust ASR. In the proposed method, a pitch synchronous mechanism is embedded in ZCPA (Zero-Crossings Peak-Amplitudes) method, which has previously been shown to outperform the conventional features in the presence of noise. A noise-robust non-delayed pitch determination algorithm (PDA) is also developed. The proposed pitch-synchronous ZCPA (PS-ZCPA) method was proved more robust than the original ZCPA method. Moreover, a simple noise subtraction method is also integrated in the proposed method, and the performance was evaluated using the Aurora-2J database. The experimental results showed the superiority of the proposed PS-ZCPA method with noise-subtraction over the PS-ZCPA method without noise-subtraction.

Keyword　Speech recognition, Auditory neuron system, ZCPA, Pitch synchronous analysis

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 the presence of noise. Seneff's model [1] uses a generalized synchrony detector (GSD) to identify formant peaks and periodicities of the speech signal. EIH model [2], 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 each frequency band. The ZCPA method [3], 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 the mechanism. The proposed PS-ZCPA method extracts the pitch-synchronous features based on the properties of the ZCPA method. In the ZCPA method, the positive zero-crossings in each subband are detected, and their intervals are calculated. The peaks within the intervals are also detected at the same time. Then a histogram of the intervals for all bands is collected with the logarithm of the peaks contributing as a weighting factor. In the proposed PS-ZCPA method, at first, a noise-robust, non-delayed pitch determination 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, features are extracted same as with the ZCPA method.

In heavily corrupted noisy environments, the temporal structure in high frequency bands for the ZCPA method is deteriorated, and the information from those bands is often unreliable, which affect the recognition performance. In this paper, we integrate a noise subtraction method in temporal domain to the proposed PS-ZCPA method to overcome this kind of noise effect.

The paper is organized as follows. Section 2 presents the system configuration, where both the proposed PDA and the PS-ZCPA method are described; section 3 gives the experimental results; section 4 introduces noise subtraction in the proposed method. A comparative experimental result on the proposed methods using the Aurora-2J database is given in section 5. Finally, section 6 draws some conclusions.

2. System overview

Fig. 1 shows the block diagram of the proposed PS-ZCPA-based feature extraction method. Speech signal is passed through a bank of N band pass filters (BPFs) with the center frequencies (f_c) are uniformly distributed on a Bark scale. The proposed method is divided into two parts: a) pitch determination, and voiced and unvoiced/silent segments detection, and b) feature extraction.

2.1. Pitch determination and voiced (V) / unvoiced or silent (U/S) segments detection

For pitch determination, outputs from first n filters are used. The frame length is set to 35ms, and the frame rate is 10ms. At first, each filter output is half-wave

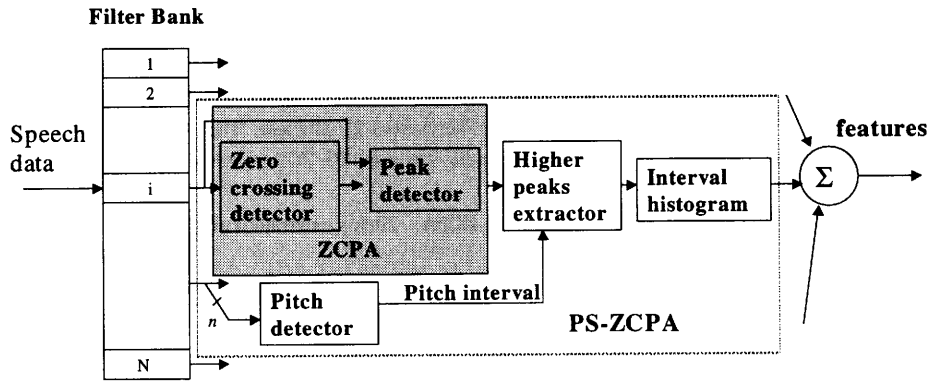


Fig. 1: Block diagram of the proposed PS-ZCPA method

rectified, and 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. A novel noise-robust, non-delayed pitch determination algorithm is then applied to the summary auto-correlogram to find out the pitches. If the pitch is equal to zero for a certain frame, then the frame is considered to be an unvoiced/silent frame, otherwise the frame is a voiced frame.

The basic idea of auto-correlation based pitch tracking is that the correlogram will have a large peak at the lag corresponding to the pitch period. But in real world, it may happen that there are many large peaks at half or double pitch periods, or there are many unwanted peaks at random lags. We need a carefully designed algorithm to ignore all these unwanted or irregular peaks, and to extract the right peak corresponding to the pitch period. In this paper, a noise-robust PDA is developed to overcome the shortcomings for pitch detection in real world.

The basic steps of the PDA are as follows:

(i) Find the local maximums (L_{max}). The L_{max} are greater than the values at ±3 lags, and higher than a threshold.

(ii) Find the global maximums (G_{max}). The G_{max} are greater than the L_{max} with some threshold at ±2 ms. If the maximum of the G_{max} lies outside half of the frame length, then there is no pitch. It means the proposed PDA can detect any pitch greater than 2ms and less than 17.5ms.

(iii) Increase the weight of the G_{max} that have peaks with height less than 110% of that G_{max} at their multiple (up to 4th multiple) lags. The weight is increased by 4/b, where b is the multiple integer (2nd multiple, 3rd multiple, etc.).

(iv) Find the G_{max} with the maximum weight. The lag of that G_{max} corresponds to the pitch period.

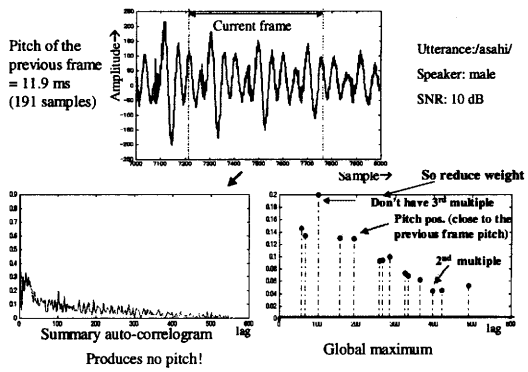
The key features of the proposed PDA are given below:

(i) The proposed PDA adopts a non-delayed approach. It means that while determining pitch at frame t, it does not check for any information of later frames, i.e., frames t+1, t+2, etc.

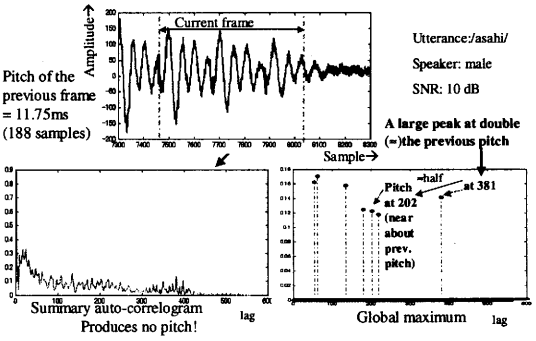
(ii) As pitches do not change abruptly in successive frames, the proposed PDA always check the pitch of the previous frame, t-1.

(iii) The proposed PDA makes a good use of the peaks located at multiple lags of a pitch candidate. An intensive observation shows that, for noisy data, a false pitch candidate may have peaks at its 2nd, 3rd, or 4th multiple lags. The proposed PDA not only increases the candidacy by corresponding weight, regarding to the presence of peaks at i-th multiple lag, but also decreases the candidacy in case where no peak is available at any multiple lag (for example, where a candidate has peak at its 2nd multiple lag, but does not have peak at its 3rd multiple lag, then the weight of that candidate is reduced). This increases the pitch accuracy and eliminates the possibility of generating half-pitch / double-pitch error. It is illustrated in Fig. 2(a).

(iv) At the end of a voiced segment of a noisy speech, it is very difficult to correctly determine the pitch. In this situation, where there is a pitch (p) at previous frame, but no pitch at current frame, the proposed PDA checks whether there is a large peak at around p, or at 2nd multiple of p in the current auto-correlogram. If there is any large peak at any of those lags, a pitch is set for the current frame (see Fig. 2(b)). This increases the accuracy of the voiced (V) segments detection.



(a) Eliminating half-pitch error, by reducing weight



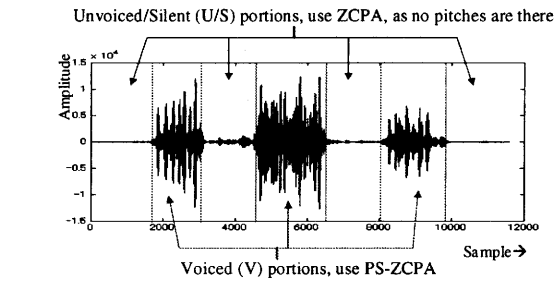
(b) Improving voiced segment detection by checking previous frame

Fig. 2: Examples for the enhancement of pitch detection (for detail, see section 2.1).

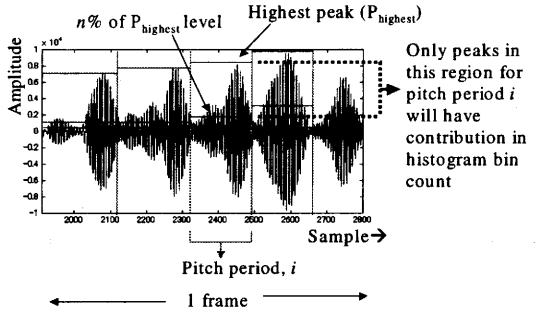
2.2. The proposed PS-ZCPA-based feature extraction method

The pitch intervals, obtained by the PDA, are used to extract the pitch-synchronized features for the proposed PS-ZCPA method. The spectrum is estimated using the ZCPA method. The original 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 subband signal, and (3) collect histograms of the inverse zero-crossing lengths over all the subband signals. At the last stage, the peak value in each zero crossing interval is detected, and then the histogram bin count for the corresponding interval is increased by the logarithmic value of the peak. However, in presence of heavy noise, the smaller peaks are corrupted by noise, which in turn have adverse effect on the performance. To overcome this shortcoming of the original ZCPA method, the proposed PS-ZCPA method does not consider the smaller peaks that are less than some threshold to contribute in the histogram bin count.

In the proposed method, for a voiced segment, the highest peak ($P_{highest}$) within a pitch period, determined by the PDA, is detected. Only those peaks that have height above n% of $P_{highest}$ within that pitch period are to contribute in the histogram bin count in the same manner as with the ZCPA method. The other peaks (smaller ones) in that pitch period are of no contribution. The n should carefully be chosen so that no important information is lost, as well as less noise-corrupted peaks are accounted. For unvoiced/silent segments, as there are no pitches, the features are extracted same as with the ZCPA method. Fig. 3 illustrates this with an example.



(a) Output of filter i



(b) A voiced segment of the output of filter i

Fig. 3: An illustration of the PS-ZCPA method. The utterance is /asahi/ by a male speaker. (a) Use the PS-ZCPA method for only voiced segments. (b) The peaks larger than some threshold have contribution in the histogram count.

Table 1: Error rate of the proposed PDA

SNR (dB)	Gross error	V to U/S	U/S to V
Clean	0 (0%)	2 (0%)	3 (0.01%)
10	7 (0.01%)	5 (0.01%)	7 (0.02%)
5	20 (0.02%)	15 (0.02%)	10 (0.02%)
0	27 (0.03%)	21 (0.03%)	14 (0.03%)

3. Experiment

3.1 Experiments on the proposed PDA

10 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 1028 frames of which 616 were voiced and the rest were unvoiced. White Gaussian noise was added to the clean speech at SNR = 10dB, 5dB, and 0dB.The reference pitches were extracted by manually checking the speech waveforms. 10 BPFs were used (highest f_c = 2kHz) to calculate the summary auto-correlogram. The experimental results are shown in Table 1. The results are given in the number of frames, and also in the percentile form in the brackets. If the PDA generated pitch of a frame is not within 20% of the actual pitch period, then it is considered as a gross error. From Table 1, we can see that the proposed PDA yields high accuracy. It can be mentioned that the U/S to V error has less significant effect on the proposed PS-ZCPA method.

3.2 Experiments on the PS-ZCPA method

3.2.1 Database

Train data set: A subset of "ASJ Continuous Speech Database" consisting of 4503 sentences spoken by 30 male speakers. The sampling rate is 16 kHz.

Test data set: 100 isolated Japanese words from Tohokudai-Matsushita database [5], spoken by 10 male speakers each. White Gaussian noise was added to the clean speech at SNR = 10 dB, 5 dB.

3.2.2. Experimental setup

20 FIR Hamming filters of order 61, with center frequencies uniformly spaced on the Bark scale between 250 Hz and 6.8 kHz were used in the experiment. Frequency range between 0 and 7 kHz was partitioned into 18 histogram bins uniformly distributed on the Bark scale. Frame length was set to $30/f_{ck}$, where f_{ck} were the center frequencies of the filters in kHz. The value of n, described in section 2.2, was varied from 20 to 70. The frame rate was 10ms. We compared the proposed PS-ZCPA method with the conventional ZCPA method and also with MFCC. The feature vectors consisted of 12 cepstrum and corresponding delta and delta-delta features. MFCC included two additional features: delta power, and delta-delta power.

3.2.3 Results and discussion

The experimental results in word accuracy (WA) [%] of the proposed PS-ZCPA, the ZCPA and MFCC are shown in Fig. 4. The value of n was set to 40, which gave a very competitive result. The optimum results for the PS-ZCPA method were obtained by adjusting n as n = 20, 40, 60 for clean, 10dB, and 5dB, respectively.

From Fig.4, we can see that the performance of the proposed PS-ZCPA gradually increased with reduced SNR. For n = 40, some peak information is lost for clean speech, resulting in slightly reduced performance. But for larger n, more noise-corrupted peaks are neglected, making the proposed method more robust. However, a large n increases the probability of losing much peak information. This result indicates that the value of n is critical for recognition performance. For optimal result, we need to adjust the value of n for different values of SNR. To overcome this problem, we integrate a time-domain noise subtraction procedure, which is discussed in the following section, in the proposed PS-ZCPA method.

4. The PS-ZCPA method with noise-subtraction

A noise-subtraction procedure is applied to the proposed PS-ZCPA method to reduce the effect of noise in the performance, and also to fix a value of n, described in Section 2.2, for the speech signals with different SNR for

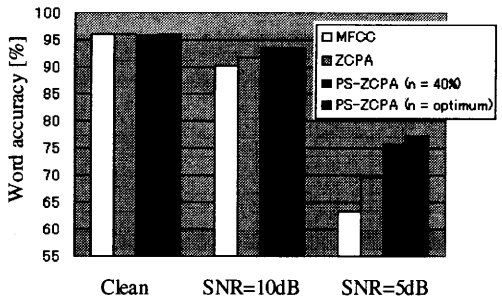


Fig. 4: Word accuracy

the optimum result. In heavily corrupted noisy environments, the temporal structure obtained by the ZCPA method in high frequency bands is deteriorated, and becomes unreliable in recognition. The noise-subtraction procedure that we developed, can easily adapt the noise threshold level with the change in SNR, and thereby reduce the need of changing the value of n for optimum performance. The noise level is checked in each channel output, and hence the temporal structure even in high frequency bands is reliable.

In the noise-subtraction procedure, the following steps were performed:

i) The average value (P_{avg}) of the peaks in the first 160ms (silent segments) for each filter output is calculated.

ii) The zero level of each filter output is virtually switched to the corresponding P_{avg} level. The histogram bin count is same as the PS-ZCPA method, where n is fixed to 20%. For example, for any pitch period, if a peak is above 20% of $P_{highest}$, described in section 2.2, and also above the P_{avg} level, then that peak will contribute in the histogram bin count. Otherwise, that peak will have no contribution. For the unvoiced and the silent segments, only P_{avg} level is activated.

5. Experiments

The performances of the proposed PS-ZCPA method both with noise-subtraction (NS) and without NS (n=40%) were evaluated using the Aurora-2J database [6]. The sampling rate was 8 kHz, and the utterances were connected digit-strings. The experimental results are shown in Table 2, where (a) and (b) show the results for the PS-ZCPA method without NS and with NS, respectively. From Table 2, it can be seen that the PS-ZCPA method with NS has higher accuracy than that

Table 2: The performance of the PS-ZCPA 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	99.97	99.82	99.94	99.80	99.88	99.82	99.80	99.96	99.76	99.84	99.90	99.72	99.81	99.85
20 dB	97.34	92.31	93.06	96.60	94.83	90.12	94.57	93.21	91.90	92.45	94.26	93.10	93.68	93.65
15 dB	88.18	84.24	81.68	87.65	85.44	80.56	81.89	76.12	80.10	79.67	76.90	79.37	78.14	81.67
10 dB	78.10	70.44	73.90	75.43	74.47	75.65	74.76	72.43	70.52	73.34	68.87	65.90	67.39	72.60
5 dB	52.11	50.67	49.11	51.87	50.94	54.25	50.11	56.22	48.34	52.23	47.98	46.38	47.18	50.70
0 dB	31.01	27.80	29.29	28.63	29.18	28.78	30.40	30.12	27.76	29.27	26.65	23.34	25.00	28.38
-5 dB	21.78	20.31	21.76	18.52	20.59	16.62	18.31	19.98	19.56	18.62	18.71	15.20	16.96	19.08
Average	69.35	65.09	65.41	68.04	66.97	65.87	66.35	65.62	63.72	65.39	62.93	61.62	62.28	65.40

(a) Without noise subtraction

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	99.87	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

(b) With noise subtraction (NS)

of without NS. The reason behind that is the automatic adjustment of the threshold level for each filter output, above which the peaks are to contribute in the histogram bin count. For higher SNR, the value of n is activated, while for lower SNR, the average value of the noise-level (P_{avg}) becomes prominent. Table 3 represents the relative performance of the PS-ZCPA method without and with NS in comparison with the MFCC features. For example, the PS-ZCPA method with NS had 42.37% relative improvement over MFCC. Fig. 5 shows the spectrogram of the utterance /roku/ ('six') for clean (upper), and subway noisy speech having SNR = 10 dB (middle) with the PS-ZCPA method with NS, and for the same noisy speech (lower) with the PS-ZCPA method without NS. The degraded temporal structure in high frequency bands (marked 'A' in lower spectrogram) is solved using NS (middle spectrogram).

6. Conclusion

A pitch-synchronous ZCPA-based feature extraction method was proposed. The proposed PS-ZCPA method was proved more robust than the original ZCPA method. A sophisticated channel-depended noise-subtraction method, which in turn was showed to be more efficient, was also integrated in the proposed method to handle the noise-threshold level for any kind of SNR. The future work will be to reduce the computational load of the proposed method.

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

Table 3: Relative performance of the PS-ZCPA with and without noise-subtraction (NS) comparing to MFCC

Relative performance				
	A	B	C	Overall
PS-ZCPA	38.26%	38.22%	24.70%	35.72%
PS-ZCPA with NS	45.03%	44.57%	31.78%	42.37%

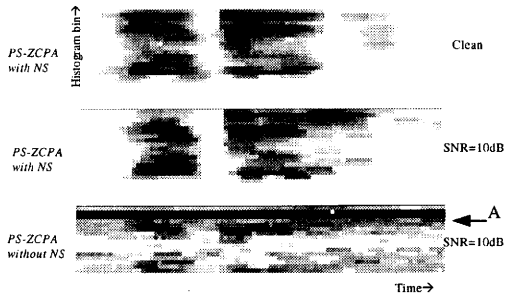


Fig. 5: Spectrogram for the speech /roku/.

References

[1] S. Seneff, in *Proc. ICASSP*, 1986, pp. 1983-1986.

[2] O. Ghitza, *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115-132, Jan. 1994.

[3] DS Kim, SY Lee, and R. M. Kil, *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55-69, Jan. 1999.

[4] T. Hashimoto, et al, *Japanese J. Physiology*, vol. 25, pp. 633-644, 1975.

[5] S. Makino et al, ASJ Trans., vol. 48, no. 12, pp.899-905, 1992.

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