A User Friendly Remote Speech Input Unit in Spontaneous Speech Translation System

Kwang-seok Lee · Heung-jun Kim · Jin-kook Song · Yeon-gyu Choo
Jinju National University
email : kslee@jinju.ac.kr

ABSTRACT

In this research, we propose a remote speech input unit, a new method of user-friendly speech input in speech recognition system. We focused the user friendliness on hands-free and microphone independence in speech recognition applications. Our module adopts two algorithms, the automatic speech detection and speech enhancement based on the microphone array-based beamforming method. In the performance evaluation of speech detection, within-200msec accuracy with respect to the manually detected positions is about 97percent under the noise environments of 25dB of the SNR. The microphone array-based speech enhancement using the delay-and-sum beamforming algorithm shows about 6dB of maximum SNR gain over a single microphone and more than 12% of error reduction rate in speech recognition.

KEYWORDS

Speech Transition Detection, Approximate-Synthesis Method, speech enhancement

Ⅰ. Introduction

Most of speech recognition systems use a single microphone as an input device. However, these single microphone -based speech recognition systems are inconvenient for the common users. They usually require the user to speak close to the microphone, within a limited range of direction. In order to improve the user friendliness, we present a remote speech input unit designed to be used as the front part of out multilingual spontaneous speech translation system.[1] Our remote speech input unit has two functions : the automatic speech detection and the microphone array-based speech enhancement, We implemented the automatic speech detection module detection the proper speech region from the incoming signal. Because the incoming speech signal has relatively low SNR for the remote seech input case, we need to improve the SNR of the signal. Our approach is based on the delay-and-sum beamforming algorithm[2,3] withe eight channel microphone array to increase the SNR, because this algorithm involving an array of microphones provided some promising solutions for the acquisition of robust and enhanced speech signal in noisy environments.[3,4,5]

This research is organized as follows. In section 2, we first represent the algorithm of our remote speech input unit. We then describe the experimental procedures and evaluations with the discussion of our findings in section 3. Finally, we conclude our works in section 4.

Ⅱ. The Algorithm of the Remote Speech Input Unit

We implemented our remote speech input unit with a TMS 320C40 DSP board. This unit generates a speech data file when the speech is detected. All of its basic algorithms are designed to run on the DSP board and sends the generated speech file to our 5,000 words spontaneous speech translation system. All of these modules are integrated to operate together such that the routines from the speech input to the speech translation are processed sequentially. Two basic functions of this remote speech input module, the automatic speech detection and microphone array-based speech enhancement, are represented as follows.

1. Automatic Speech Detection

The use of a keyboard or a mouse prior to the utterance of speech is not so convenient for users in speech recognition applications. An automatic procedure of utterance detection should be made to solve this problem. Besides, the precise classification of the silence and the speech region is also very important especially in the applications requiring real-time processing[6]. We adopt a frame synchronous speech or silence is made at avery input frame.

We then count the number of frames that are decided as the speech frames containing one current and a number of consecutive frames containing one current and a number of past input frames. When this number is greater than the threshold value, we assume that the start point of speech region is detected. In the initial stage of speech detection algorithm, we extract the feature of the silence from pause region and use it as the reference in deriving the threshold values. As feature parameters, we use the energy and the level crossing rate and compute them for each frame. In the initial pause region, we estimate the silence energy and the silence noise level. We defined energy as the integral sum of the absolute signal values. The level crossing rate is defined as the frequency that two adjacent signal pairs cross the level within one input frame . The level is chosen as two times of the average value of positive noise values to avoid the crossing due to small noises as well as to include that of unvoiced sound. From the energy and level crossing rate of the silence region, we derive the threshold values for the decision of silence and speech frame region.

After the start point of speech region is detected, we begin to detect the end point of the utterance. This end point detection algorithm is similar to that of the of the start point. At every analysis frame, we examine the number of frames that are decided as the silence frame in the pre-determined number of consecutive input frames containing the current and the past analysis frames. When this number is less than the threshold values and this state lasts for the pre-determined number of frames, we regard that the end point is detected. At the end point detection, we use smaller threshold value than that of the start point to include the utterance containing lower energy.

2. Speech Enhancement using Microphone Array-based Beamforming

At every 500 msec within the speech region, eight channel speech data are sent to the speech enhancement module, and the microphone array-based speech enhancement is performed. Our algorithm uses the delay-and-sum beamforming method and its procedure consists of following for stages; time delay estimation, time delay compensation, noise level normalization, and multi-channel speech summing. In the time delay estimation, we estimate the time delay between speech signals of different channels. Next, we synchronize the multi-channel speech signals by compensating the time delays between channels. We then normalize the noise level of eight channels. Finally, we derive the noise reduced speech signal by summing the synchronized eight channel speech signals and send to the PC side to use it as the input of the speech recognizer. In the time delay estimation, we use the time domain cross-correlation method[7]. We first calculate the cross-correlation coefficients of speech signals for different two channels. The index of maximum cross-correlation is chosen as the time delay between the two analysis channels. These are represented as following equation.

$$\tau_k = \underset{\tau}{\operatorname{argmax}} \sum_{n=0}^{N-1} x_0(n)x_k(n-\tau),\ k=1,2,\cdots,L-1 \quad (1)$$

Where, $x_0(n)$, $x_k(n)$ represent the nth speech signals of the basis channel 0 and the test channel k, respectively.

Reliable and precise estimation of time delay is critical to the performance of the delay-and-sum beamforming-based speech enhancement. Because the speech region with larger energy is more robust on the effect of noise in the cross-correlation procedure. we detect the speech region with the largest energy in each speech duration of 500 msec. Before the cross-correlation procedure, we process center-clipping[8], to this speech region. This process is effective in reducing the error on the estimation of cross-correlation coefficients. We choose the clipping level as 70 percent of the maximum signal value in the analysis frame. Next, the time delay compensation is performed to synchronize the speech signals from all eight channels. The weighting for each channel is generally followed to trade off the relationship between array beamwidth and average sidelobe level[2],[3],[9]. In our case, we choose the weight coefficients to normalize the noise level of each channel. In the final step, delay compensated signals are summed and noise reduced signal is obtained. These sequential procedures are represented as follow.

$$\tilde{x}(n) = \frac{1}{L}\sum_{n=0}^{N-1} w_k x_k(n+\tau_k),\ k=0,1,\cdots,L-1 \quad (2)$$

Where, x_k represents the signal from the kth channel, N is the length of analysis frame in the cross-correlation procedure, and L is the number of channels.

In ideal case where the speech signals and noise are incoherent each other, the time delay is accurately estimated, etc., the SNR gain of the delay-and-sum beamforming output is $\log_2 N$ dB(Our case is about 9dB) over a single microphone output.

III. Experimental Procedure and Evaluations

We made two different kinds of experiments to evaluate the performance of automatic speech detection and microphone array-based speech enhancement. In the automatic speech detection experiments, we investigate the accuracy of speech detection algorithm under different SNR environments. We then examine the improvement of the SNR and speech recognition rate for the enhanced speech signals. The details of experiments and results are discussed as follows.

1. Automatic Speech Detection

1.1 Database and Experiments

The speech database used for the performance evaluation of speech detection algorithm is consist of 105 sentences of Korean spontaneous speech dialogues that are uttered by four male speakers in the sound-proof room. All these data are then digitized with 16kHz sampling rate and 16 bit quantization level. Because these speech data are clean speech with higher SNR. We added noise data to these speech data to produce noisy speech having the desired SNRs(15, 20, and 25dB). The noise data are collected under the normal office conditions and have almost flat spectral characteristics with significant 60Hz energy components. The automatic speech detection algorithm has three parameters, that is, the threshold values of energy, level crossing rate, the number of consecutive speech frames in the decision of start and end point. The first two parameters, energy and level crossing rate are used to distinguish whether the analysis frame is speech or silence. The other parameter, the number of consecutive speech frames is adopted to reject the short time shot noise having large energy. Since we focused on the effect of two factors, the energy, and the number of consecutive frames, we chose the threshold of level crossing rate to a fixed value for all our experiments. This approach is based on the fact that the level crossing rate plays relatively less important role than the energy in speech detection.

1.2. Performance Evaluations

The results represented in Table 1 is from the experiment examining the effect of three parameters, SNR, energy, and the number of consecutive speech frames. The accuracy is defined as the probability of detecting the start point within 200 msec with respect to the manually detected point. As expected in advance, the results show better for the speech data with higher SNR. The best performance is about 97percent for start point and 92percent for end point. In the error analysis, we know that most of the start point errors are due to the utterances starting with short plosive sounds while the end point errors are resulted from various factors such as low SNR, utterance manner, long pause, etc. The undesirable accuracy in low SNR implies that we need to include some amount of signals outside the start and end point not to damage the real speech region.

2. Microphone Array-based Speech Enhancement

Table 1. The accuracy of automatic speech detection [%]

		Start point			End point		
Fea.1	Fea.2	15 [dB]	20 [dB]	25 [dB]	15 [dB]	20 [dB]	25 [dB]
5	3	42.8	80.9	97.1	48.5	80.0	92.3
5	4	37.1	71.4	95.2	47.6	76.1	89.5
7	3	30.4	69.5	92.3	18.0	79.0	91.4
4	4	26.6	53.3	91.4	45.7	75.2	89.5

Fea.1. means the number of consecutive frames and Fea.2. Represents the coefficient in energy threshold value.

2.1 Database

We located eight microphone at the edges of the PC monitor frame such two microphones are attached at each edge of the frame while keeping the two microphones about 18cm apart each other. In the performance evaluation, we did a series of experiments to investigate the gain of SNR and the improvement of speech recognition rate. For the SNR gain test, we collected speech data by the following manner. Two male speakers repeat to utter five same

Korean sentences at six different positions. These positions are the center, left, and right from the front of the PC monitor with distances to the monitor are about 40 and 80cm. For the test of the improvement in speech recognition rate, we collected speech data uttered by two male speakers under different SNRs. These speech data contain the speech signal from one of eight microphone outputs as well as that from the beamforming output. Each data consists of 115 sentences of Korean spontaneous spoken dialogues.

2.2 Experiments and Results about SNR Gain

The results from experiments about SNR gain are given in Table 2. The average gain of SNR from the microphone array-based speech enhancement is 2.3~4.0dB while the maximum gain is 5.9dB. All these results are obtained by comparing with the speech signals from one microphone output having maximum SNR. Because the sentence speech data used in the experiments contain some amounts of pause reions, the real gain of SNR should be larger than these results. Compared with the ideal case where improvement of SNR is 9dB, the performance of our result is less than half of that from the ideal case. We think this is due to some factors such as the presence of coherent noises, the location of microphones, different characteristics of microphones, and the errors in the time delay estimation, ect. After considering there factors, we expect the SNR gain could be improved further with proper study.

Table 2. The SNR gain of the microphone array-based beamforming output over a single microphone output

Speaker's distance and position[cm]		One channel output [dB]	Beam-forming output [dB]	Average gain[dB]	Max. gain[dB]
40	center	15.9	19.9	4.0	5.9
	left	18.6	21.0	2.4	4.4
	right	17.8	20.9	3.1	4.7
80	center	16.7	19.0	2.3	3.8
	left	15.1	18.3	3.2	4.9
	right	17.2	21.0	3.8	5.4

2.3 Experiments and Results about the Improvement of Speech Recognition Rate

We also investigate the improvements of speech recognition rate for the enhanced speech. We test the speech data to our spontaneous speech recognition system of which the best performance is about 72percent for 5,000 vocabulary-domain. These results are given in Table 3. Because the speech data for Test 2 have errors due to speech detection, they show lower accuracy than that of Test 1. The error reduction rate is 12.3 and 20.7percent for Test 1(#1) and Test 2(#2) respectively. Though we consider the effect of speech detection errors in Test 2, we know that the error reduction rate is at least 12percent. This implies that the adoption of microphone array-based speech enhancement is very effective in improving the performance of speech recognition system under low SNR environments.

IV. Conclusions

We proposed a remote speech input unit to efficiently input the speech without caring the position of microphones. It also need not use the mouse or keyboard to trigger the speech input. In order to achieve these functions, we adopt the automatic speech detection and the microphone array-based speech enhancement. The automatic speech detection module detects speech portion from the incoming signals. Within-200msec accuracy with respect to the manually detected positions is about 97percent under the noisy environments of 25dB of the SNR. The microphone array-based speech enhancement using the delay-and-sum beamforming algorithm shows about 6dB of maximum SNR gain over a single microphone and more than 12percent of error reduction rate in speech recognition.

As future works, we plan to adopt the adaptive beamforming algorithm in out unit to improve the noise reduction effect further. The time delay estimation delay estimation under low SNR environments is also to be studied.

Table 3. The improvement in speech recognition rate

	One channel output		Beamforming output	
	SNR [dB]	Recognition rate [%]	SNR [dB]	Recognition rate [%]
#1	14.1	55.4	20.8	60.9
#2	20.0	48.2	22.4	58.9

References

[1] J-W. Yang and Y. Lee. "Toward Translation Korean Speech into Other Language". Proc.ICSL, vol.4, pp.2368-2370, Oct. 2004.

[2] S.U. Pillai. Array Signal Processing. Springer-Verlag, New York, 2006.

[3] R.P. Ramachandran and R.J. Mammone. "Microphone Array for Hands-free Voice Communication in a Car". Modern Methods of Speech Processing, KAP, Boston, 2003.

[4] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W.Elko. "Computer-steered microphone arrays for sound transduction in large rooms". Journal Acoustical Society of America, vol. 78, pp.1508-1518, Nov. 2001.

[5] D.Giuliani, M.Matassoni, M. Omologo and P.Svaizer. "Robust Continuous Speech Recognition using a Microphone Array". Proc. Euro Speech, vol.3, pp.2021-2024, 2001.

[6] H. Lee and M. Hahn. "Development of a Real-time Endpoint Detection Algorithm" Proc. ICSPAT, vol.2, pp.1547-1553, Sept. 2005.

[7] G. Clifford Carter. "Coherence and Time Delay Estimation". Proceedings of the IEEE, vol.75, no.2, pp.236-255, Fed.2000.