A User-friendly Remote Speech Input Method in Spontaneous Speech Recognition System

*Young-Joo Suh, *Jun Park, and *Young-Jik Lee

Abstract

In this paper, we propose a remote speech input device, a new method of user-friendly speech input in spontaneous speech recognition system. We focus the user friendliness on hands-free and microphone independence in speech recognition applications. Our method adopts two algorithms, the automatic speech detection and the microphone array delay-and-sum beamforming (DSBF)-based speech enhancement. The automatic speech detection algorithm is composed of two stages; the detection of speech start point candidate and end point using energy and level crossing rate as its features and the classification of speech and nonspeech using the pitch information for the detected speech portion candidate. The DSBF algorithm adopts the time domain cross-correlation method as its time delay estimation. In the performance evaluation, the speech detection algorithm shows within-200 ms start point accuracy of 93%, 99%, and 99% under 15 dB, 20 dB, and 25 dB signal-to-noise ratio (SNR) environments, respectively and those for the end point are 72%, 89%, and 93% for the corresponding environments, respectively. The classification of speech and nonspeech for the start point detected region of input signal is performed by the pitch information-based method. The percentages of correct classification for speech and nonspeech input are 99% and 90%, respectively. The eight microphone array-based speech enhancement using the DSBF algorithm shows the maximum SNR gain of 6 dB over a single microphone and the error reduction of more than 15% in the spontaneous speech recognition domain.

I. Introduction

In order to use speech as the most natural means of communication between human and machine, massive researches are made to develop speech recognition system. However, there are still some problems that should be solved for this speech recognition system to be used successfully to the users. The main problem is the poor accuracy of the speech recognition system. The second problem would be the inconvenience and difficulty in operating the speech recognition system for the user' s point of view. This second problem is especially important in the real world applications of speech recognition. In the earlier stage of developing the speech recognition system, most of the efforts were focused on solving the first problem. As this first problem is solved within some acceptable extents, now it is necessary to move the focus of research on the second one, that is, the user friendliness.

Nevertheless, most of the current speech recognition systems still use a single microphone as their 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 method designed to be used as the front part of our spontaneous speech recognition system [1][2][3]. Our remote speech input method has two functions: the automatic speech detection and the microphone array-based speech enhancement. We implemented the automatic speech detection module that detects the proper speech region from the incoming signal. Because the incoming speech signal has relatively low SNR in the remote speech input case, we need to improve the SNR of the speech signal. Our approach is based on the DSBF algorithm [4][5], with eight channel microphone

* Spoken Language Processing Section, ETRI
Manuscript Received : July 4, 1998.

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 [5][6][7].

Ⅱ. The algorithm of the remote speech input device

We implemented our remote speech input device on a WINDOWS 95-based PC with a TMS 320C40 DSP board and depicted its block diagram in Figure 1. This device generates a speech data file when the speech is detected. All of its basic algorithms are designed to run on the DSP board and the PC sends the generated speech data to our 5,000 word spontaneous speech recognition system. All of these modules are integrated to operate together such that the routines from the speech input to the speech recognition are processed sequentially. Two principal functions of this remote speech input method, the automatic speech detection and the DSBF-based speech enhancement are presented as follows.

2.1. 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 speech recognition applications which require higher recognition performance and real-time processing [8].

To satisfy this requirement, we adopt a frame synchronous speech detection approach that consists of two stage procedures. In the first stage, the candidate of speech start point and the speech end point are detected. First, the detailed algorithm detecting the candidate of start point is described as follows.

1) Blocking into frames: N consecutive samples of input signal are used as a single frame, and these consecutive frames are also spaced N samples apart (we use N as 320 corresponding to 20 ms of input signal).
2) Computing features: The energy and level crossing rate for each input frame are calculated by the following equations.

$$ENG(i)=\sum_{j=0}^{N-1}x_i(j)^2 \qquad (1\text{-}1)$$

$$LCR(i)=\sum_{j=0}^{N-2}sgn[-(x_i(j)-LVL)\times(x_i(j+1)-LVL)] \qquad (1\text{-}2)$$

$$sgn(x)=1, \quad if \quad x)0, \qquad (1\text{-}3)$$

=0, otherwise

where $x_i(j)$ is the j^{th} sample value of the i^{th} frame. The level, LVL, is chosen as two times of the average value of positive silence samples to avoid the crossings due to small noises but to include those of unvoiced sound.

3) Deriving the threshold values: The threshold values are derived to decide whether the current analysis frame is speech or silence (we use two threshold values corresponding to the energy and the level crossing rate for each frame, respectively).

$$ENG\_THR=C_{ENG}\times ENG\_SILENCE \qquad (2\text{-}1)$$

$$LCR\_THR=C_{LCR}\times LCR\_SILENCE \qquad (2\text{-}2)$$

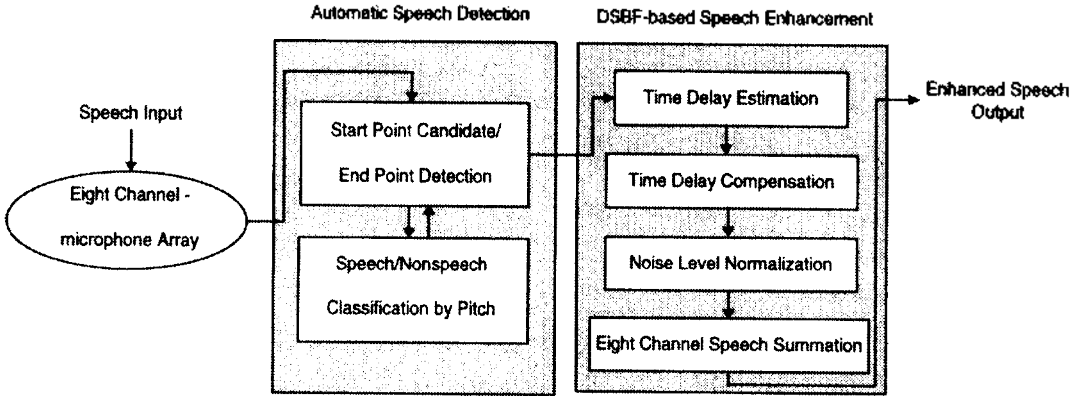
Automatic Speech Detection
DSBF-based Speech Enhancement
Speech Input
Eight Channel - microphone Array
Start Point Candidate/ End Point Detection
Speech/Nonspeech Classification by Pitch
Time Delay Estimation
Time Delay Compensation
Noise Level Normalization
Eight Channel Speech Summation
Enhanced Speech Output

Figure 1. Block diagram of the remote speech input method.

where *ENG_THR*, *LCR_THR* are the threshold values for the energy and level crossing rate, respectively. C_{ENG}, C_{LCR} are the multiplying coefficients for the energy and the level crossing rate, respectively. *ENG_SILENCE*, *LCR_SILENCE* represent average values of the energy and the level crossing rate for the initial silence region, respectively.

4) Deciding speech or silence frame: For each frame, the decision of speech or silence is made by comparing the two feature values with the previously derived threshold values as follows

$$\text{If } [(ENG(i) > ENG\_THR) \text{ or } (LCR[i] > LCR\_THR)], \tag{3}$$

Then FRM(i) = 1 (i.e., speech frame)

Else FRM(i) = 0 (i.e., silence frame)

5) Detecting the candidate of start point: Within the constant CONSEC_ST_FRM number of consecutive frames including the current frame and the past CONSEC_ST_FRM-1 frames, we count the number of speech frames. If this number is greater than the pre-determined threshold value, CONSEC_ST_FRM_THR, then we find the candidate of start point of speech region by the following equation.

Then Candidate of start point is (i-CONSEC_ST_FRM+1)th frame.

Else Candidate of start point is not detected and repeats to detect the candidate for the next analysis frame.

$$\text{If } \Big(\sum_{k=0}^{CONSEC\_ST\_FRM} FRM(i-k) \rangle CONSEC\_ST\_FRM\_THR\Big), \tag{4}$$

When the candidate of the start point of speech region is detected, we enter the procedure that classifies the detected speech region candidate is speech or nonspeech (i.e., noise). The steps in this procedure are as follows.

1) Estimating voiced region: For V_FRM number of frames after detecting the candidate of start point, we search the frame having the maximum energy value and then estimate the voiced region as this frame as well as its previous and next M frames (we choose M as 1). This approach is based on the fact that most of the voiced speech are vowel and this vowel speech shows relatively larger energy than unvoiced speech.

$$ENG\_MAX\_FRM = \text{argmax}(ENG\_FRM(i)),\ i=0,1,\cdots,V\_FRM\text{-}1. \tag{5}$$

The estimated voiced region:

$$\sum_{i=-M}^{M} FRM(ENG\_MAX\_FRM + i)$$

2) Computing correlation coefficient (r(t)): For the estimated voiced speech region, we compute the correlation coefficient, r(k), defined by the following equations.

$$r(k) = \frac{R(k) - \eta_0 \eta_k}{\sigma_0 \sigma_k},\ \text{k=P, P+1,..., Q} \tag{6-1}$$

$$R(k) = \frac{1}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} x(n)x(n+k),\ \text{k=0, P, P+1,..., Q} \tag{6-2}$$

$$\sigma_k = \sqrt{\frac{1}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} (x(n+k) - \eta_k)},\ \text{k=0, P, P+1,..., Q} \tag{6-3}$$

$$\eta_k = \frac{1}{N} \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} x(n+k),\ \text{k=0, P, P+1,..., Q} \tag{6-4}$$

where N is the size of the window that is applied to the estimated voiced region, P and Q are the lower and upper boundary of pitch estimation region, respectively.

This correlation coefficient, r(k), lies between 0 and 1 and its statistics shows superiority to the normalized autocorrelation coefficient, R(k)/R(0). So we adopt the correlation coefficient to detect the pitch period more reliably.

3) Peak-picking: For the computed correlation coefficients, we extract the pitch candidate by picking the peak point of correlation coefficients for the range between P and Q.

4) Postprocessing: Because this pitch extraction method is based on the detection of peak value, there is a possibility such that twofold of the true pitch period can be determined as the pitch candidate. To solve this problem, we adopt a postprocessing procedure in which we re-examine the values of correlation coefficient for the region around the half value of the pitch candidate. If the difference between the peak value of the correlation coefficient in the region and

the value for the firstly determined pitch candidate is within the threshold value, 0.1, then we regard this smaller point as the new pitch period.

5) Classifying into speech or nonspeech: we classify the previously estimated voiced region into the true human voiced speech region or nonspeech noise region by the following method. If the extracted pitch value lies within the human pitch range (3.4 - 12.5 ms in this case) and the value of the correlation coefficient at this pitch point, r(PITCH), is greater than the threshold value (i.e., 0.5), we decide the estimated voiced region is speech and start to find the end point of speech region. In other cases in which the region may be nonspeech or unvoiced speech, we apply this speech and nonspeech classification procedure to another region of 100 ms just after the first decision procedure. This re-examination procedure can be repeated up to twice. When the last classification is nonspeech, the region after start point candidate is assumed as nonspeech and the routine detecting a new start point candidate is resumed.

When the incoming signal is classified as speech, we start to detect the end point of speech region. This end point detection algorithm is similar to the start point one. At every analysis frame, we examine the number of frames that are decided as the silence frame in the pre-determined number, CONSEC_END_FRM, of consecutive input frames containing the current and the past analysis frames. When this number is smaller than the threshold value, CONSEC_END_FRM_THR, and this phase lasts for the pre-determined number of frames, END_LENGTH_FRM, successively, we regard that the end point is detected. Once the start point of speech signal is detected, the speech data begin to be accumulated in the buffer. When the accumulated speech region is reached 500 ms length, these multi-channel speech data are firstly sent to the microphone array DSBF-based speech enhancement module.

2.2. The DSBF-based speech enhancement

For every 500 ms length within the speech region, eight channel speech data are sent to the speech enhancement module, and the microphone array DSBF-based speech enhancement is performed. Our algorithm uses the DSDF method and its procedure consists of the following four stages: time delay estimation, time delay compensation, noise level normalization, and multi-channel speech summation. 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 recognition system.

The first step is the time delay estimation. Reliable and precise estimation of time delay is critical to the performance of the DSBF-based speech enhancement and several methods are proposed to improve the accuracy of the estimation [9][10]. Here, we use the time domain cross-correlation method [9]. The main reason for adopting this method is that it is relatively simple in algorithmic aspect and produces relatively good performance compared with other methods. In this method, we first calculate the cross-correlation coefficients of speech signal for different two channels. The point in which the maximum value of cross-correlation coefficient resides is chosen as the time delay between the two analysis channels. These are represented as following equation.

$$\tau_k = \arg\max_{\tau} \sum_{n=0}^{N-1} x_0(n) x_k(n-\tau), \text{ k=1,2, ..., L-1} \qquad (7)$$

where $x_0(n)$ and $x_k(n)$ represent the n^{th} speech signals of the reference channel *0* and the test channel *k*, respectively.

The speech region with larger energy is higher SNR if the noise level is constant with time. Then, this region is more robust against the effect of noise. In order to utilize this fact in the cross-correlation procedure, we detect the speech region with the largest energy in each speech duration of 500 ms. In this largest energy region, we estimate the time delay.

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 [4][5][11]. In our case, we choose the weight coefficients to normalize the noise level of each channel to derive the maximum SNR gain. In the final step, delay compensated signals are summed and noise reduced signal is obtained. These sequential procedures are represented as follows.

$$\bar{x}(n) = \frac{1}{L}\sum_{n=0}^{N-1} w_k x_k(n+\tau_k)\ \mathrm{k=0,1,\ ...,L\text{-}1} \qquad (8\text{-}1)$$

$$W_k = \sqrt{\frac{\sum_{n=0}^{N-1} X_{SH_m}(n)^2}{\sum_{n=0}^{N-1} X_{SH_k}(n)^2}},\ \mathrm{k=0,1,\ ...,L\text{-}1} \qquad (8\text{-}2)$$

where $x_k(n)$ represents the n^{th} signal sample from the k^{th} channel, N is the length of analysis frame in the cross-correlation procedure, and L is the number of channels. $x_{SH_k}(n)$ is the n^{th} noise sample in the silence region of the k^{th} channel.

Ⅲ. Experimental procedures and evaluations

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

3.1. Automatic speech detection

3.1.1. Database and experiments

The speech database used in the performance evaluation of speech detection is consist of 106 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 16 kHz sampling rate and 16 bit quantization level. Because these speech data are relatively clean speech with higher SNR, we added noise data to these speech data to produce noisy speech with the desired SNRs (15, 20, and 25 dB). The added noise data are collected under the normal office conditions and have almost flat spectral characteristics with significant 60 Hz energy components.

The start point candidate and end point 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 defined in equation (1-1) and (1-2) are used to distinguish whether the analysis frame is speech or silence. The other parameter, the number of consecutive speech frames described in equation (4) is adopted to reject the short impulsive noise having large energy. Since the importance of the level crossing rate is decreased as the SNR value is lowered, we choose the threshold of level crossing rate to a fixed value for all our experiments. We thus focus our experiments on the effect of two factors, the energy, and the number of consecutive frames.

The experiments about the classification of speech and nonspeech after the detection of the start point candidate of speech region is performed using the same 106 sentence database and 120 examples of various noises that are commonly generated in office environments. We test the accuracy of classification for both speech utterances with different three SNRs and several kinds of nonspeech noises.

3.1.2. Performance evaluations

Table 1. The accuracy of speech detection under 15 dB SNR environments [%].

Case 1 CONSEC_ST_FRM=7, CONSEC_ST_FRM_THR=5, CONSEC_END_FRM=10 CONSEC_END_FRM_THR=1

Case 2 CONSEC_ST_FRM=5, CONSEC_ST_FRM_THR=3 CONSEC_END_FRM=10, CONSEC_END_FRM_THR=3.

C_{ENG}	Time [ms] Cases		20	100	200	300	400	500
2	Start	Case 1	33.9	80.2	89.6	93.4	95.3	96.2
		Case 2	45.3	87.7	93.4	95.3	96.2	96.2
	End	Case 1	0.0	16.0	71.7	77.4	85.8	91.5
		Case 2	0.0	0.9	54.7	67.0	72.6	78.3
3	Start	Case 1	10.4	44.3	56.6	66.0	72.6	75.5
		Case 2	13.2	54.7	68.9	76.4	82.1	84.0
	End	Case 1	0.0	4.7	37.7	43.4	56.6	66.0
		Case 2	0.0	0.0	20.8	26.4	28.3	40.6

Table 2. The accuracy of speech detection under 20 dB SNR environments [%].

Case 1 CONSEC_ST_FRM=7, CONSEC_ST_FRM_THR=5, CONSEC_END_FRM=10 CONSEC_END_FRM_THR=1

Case 2 CONSEC_ST_FRM=5, CONSEC_ST_FRM_THR=3 CONSEC_END_FRM=10, CONSEC_END_FRM_THR=3

C_{ENG}	Time [ms] Cases		20	100	200	300	400	500
2	Start	Case 1	55.7	97.2	98.1	98.1	99.1	99.1
		Case 2	56.6	99.1	99.1	99.1	99.1	99.1
	End	Case 1	0.0	50.9	68.7	91.5	94.3	95.1
		Case 2	0.0	1.9	81.0	86.8	91.5	96.2
4	Start	Case 1	43.4	84.0	93.4	97.2	98.1	98.1
		Case 2	51.9	89.6	95.3	97.2	98.1	98.1
	End	Case 1	0.0	24.5	76.4	82.1	89.6	94.3
		Case 2	0.0	0.9	64.2	71.7	77.4	95.8

Table 3. The accuracy of speech detection under 25 dB SNR environments [%].

Case 1. CONSEC_ST_FRM=7, CONSEC_ST_FRM_THR=5, CONSEC_END_FRM=10
CONSEC_END_FRM_THR=1.

Case 2. CONSEC_ST_FRM=5, CONSEC_ST_FRM_THR=3, CONSEC_END_FRM=10
CONSEC_END_FRM_THR=1

Ceng	Time [ms] Cases		20	100	200	300	400	500
2	Start	Case 1	30.2	99.1	99.1	99.1	99.1	99.1
		Case 2	29.2	98.1	99.1	99.1	99.1	99.1
	End	Case 1	0.0	74.5	92.5	93.4	94.0	97.2
		Case 2	0.0	9.4	92.5	94.3	97.2	98.1
3	Start	Case 1	48.1	98.1	99.1	99.1	99.1	99.1
		Case 2	50.9	98.3	99.1	99.1	99.1	99.1
	End	Case 1	0.9	53.8	85.7	90.6	93.0	97.2
		Case 2	0.0	2.8	86.8	90.6	95.3	96.2

The results represented in Table 1, 2, and 3 are from the experiments examining the effect of two parameters, that is, the energy multiplying coefficient (i.e., CENG) that is used to derive the threshold value and the number of consecutive speech frames (i.e., CONSEC_ST_FRM and CONSEC_END_FRM for start and end point, respectively) under SNR environments of 15 dB, 20 dB, and 25 dB, respectively. The accuracy is defined as the probability of detecting the start point and the end point within each specified time (i.e., 20 - 500 ms) with respect to the manually detected point. As expected in advance, the results show better performance for the speech data with higher SNR value. If we choose the acceptable start point accuracy as above 93%, the results indicate that input signal at least 200 ms prior to the detected start point should be included not to exclude the speech portion. We find that lowering the two parameters, energy coefficient and the number of consecutive frames, increases the detection accuracy. However, in this case, the algorithm is getting erroneous to the incoming short duration noises. This undesirable effect requires the adoption of pitch information-based speech and nonspeech classification algorithm as the postprocessor. As can be seen in the tables, the accuracy of the end point detection is lower than that of the start point one. This is because most of the start point errors are due to relatively simple problems like the failure of detecting short plosive sound at the starting point of utterance while the end point errors are resulted from various unexpected factors such as low SNR, the presence of impulsive noise, speaker' s undesirable uttering manner, very long pause, etc. We thus need to include more amount of signals outside the end point compared to the start point case not to damage the real speech region.

Table 4. The accuracy of the speech and nonspeech classification using pitch information [%].

Input	SNR [dB]	1st trial success	2nd trial success	3rd trial success
Speech	15	95.3	96.2	99.1
	20	99.1	99.1	99.1
	25	96.2	99.1	99.1
Nonspeech	20	89.9	-	-

Table 5. The statistics for the correlation coefficient (r(PITCH)) and the normalized autocorrelation coefficient (R(PITCH)/R(0)) at the detected pitch point.

Input	SNR [dB]	r(PITCH)		R(PITCH)/R(0)	
		Average	Standard Deviation	Average	Standard Deviation
Speech	15	0.773	0.155	0.771	0.168
	20	0.871	0.083	0.869	0.109
	25	0.878	0.154	0.892	0.168
Nonspeech noise	20	0.251	0.115	0.721	4.302

Table 4 and 5 are the results obtained by applying the pitch information-based speech and nonspeech classification algorithm. In Table 4, the results indicate that the accuracy of classifying into speech is very high for various SNR environments. Even in the SNR of 15 dB, the correct rate is above 99% within three trials. But for the nonspeech input case, the accuracy is about 90% and this is relatively lower compared to the speech classification result. In the error analysis, most of the errors are results from the cases when the input noise has human pitch-like periodicity and almost all of the short duration nonperiodic noises are correctly classified into noise. These results imply the comparative analysis about the difference of periodic pattern between the human pitch and the periodic noises needs to improve the noise classification accuracy further.

The results shown in Table 5 are comparative statistics about parameters used in pitch-based speech and nonspeech classification. Both average value and standard deviation value for the proposed correlation coefficient, r(PITCH) show superiority to that of the normalized autocorrelation coefficient, R(PITCH)/R(0), which is used in many pitch detection algorithms. From the results, we also know that the adopted correlation coefficient-based method produces especially superior performance for the noise input.

3.2. Microphone array DSBF-based speech enhancement

3.2.1. Database

We locate eight microphones at the four edges of the PC monitor frame such that two microphones are attached at each edge of the frame while keeping the two microphones are about 18 cm apart from 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 collect speech data by the following manner. Two male speakers repeat to utter 10 Korean sentences five times at six positions. These positions are the center, left, and right from the front of the PC monitor with distances to the monitor are kept 40 and 80 cm, respectively. For the test of the improvement in speech recognition, we made experiments by playing the previously described spontaneous speech database of 106 utterances containing 1,084 words through a loud-speaker at the same six positions. We also collect speech data through the microphone array directly. These speech database are made up of 250 Korean spontaneous sentences consisting of 2,805 words uttered by five male speakers under different two SNRs to test the improvement of speech recognition for the real speech database. These real speech database are collected at the center position of the microphone array and the distance between the speaker' s lip and the microphone array is 60 cm.

3.2.2. Experiments and results about SNR gain

Table 6. The SNR gain of the mictophone array DSBF-based output over a single microphone output [dB].

Distance [cm]	Position	One channel output	Eight channel DSBF output	Average gain	Maximum gain
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

The results from experiments about SNR gain are given in Table 6. The average gain of SNR from the eight microphone array DSBF-based speech enhancement ranges from 2.3 dB to 4.0 dB and the maximum gain is 5.9 dB. All these results are obtained by comparing with the speech signals from one microphone output having maximum SNR. Compared with the ideal case where the improvement of SNR is 9 dB, the performance of our result is about the half of that from the ideal case. We think this is due to some factors such as the presence of coherent noises, the improper location of microphones, different characteristics of microphones, and the errors in the time delay estimation, etc. After considering these factors, we expect the SNR gain could be improved further with proper study.

3.2.3. Experiments and results about the improvement of speech recognition rate

Table 7. The SNRs and the accuracy of speech recognition for various number of microphones at each position and distance by the DSBF-based speech enhancement.

Distance	No. of Mics	Left		Center		Right	
		SNR	Accuracy	SNR	Accuracy	SNR	Accuracy
40 cm	1	19.07	61.07	19.16	54.70	18.09	57.47
	2	19.37	62.73	19.94	55.72	18.80	59.87
	4	20.99	63.01	21.85	57.10	20.14	61.07
	8	22.24	63.36	23.00	62.36	21.24	63.38
80 cm	1	14.55	49.72	14.32	51.75	13.89	43.82
	2	15.67	53.51	15.69	54.43	15.44	50.83
	4	17.15	57.38	16.84	57.38	16.69	51.11
	8	18.26	59.23	18.19	60.42	17.88	56.46

We also investigate the improvement of speech recognition for the enhanced speech. We test the speech data to our spontaneous speech recognition system of which the best performance is about 72% for 5,000 word-domain. In Table 7, we see the results about the performance of speech recognition for various number of microphones in the DSBF method at the different positions and distances. At these experiments, we use the speech database collected by playing speech utterances through a loud-speaker to keep the consistency of speech data during the experiments. As expected, the recognition accuracy and the average SNR at the 40 cm distance are better than those of 80 cm case. The improvement in the accuracy of the speech recognition and the SNR value increases according to the increase of the number of microphones. The SNR values are shown to be similar at different positions while the value at the right position is relatively low compared to those of center or left. We think this is partly resulted from the non-identical gain characteristics of microphones. Except this particularly low SNR at the right position, the results indicate the speaking position hardly affects the performance of speech recognition when the number of microphones reaches about eight.

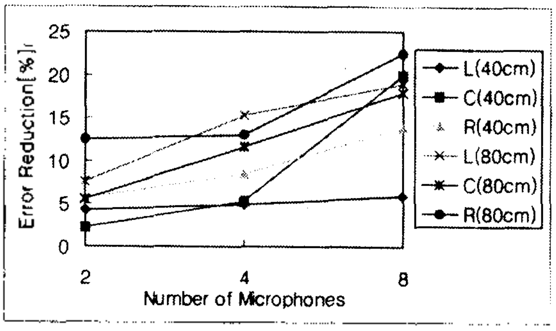

Figure 2. The error reduction rates for various number of microphones over single microphone at different positions (left, center, and right) and distances (40 cm and 80 cm) [%].

Figure 2. represents the error reduction of speech recognition for various number of microphones over single microphone. Though the trend of error reduction for the increasing number of microphones is a little irregular for the positions and distances, it is notable that most of the slopes are linearly increasing as the number of microphones increases with power of 2. This fact indicates the number of microphones plays significant role in improving the performance of speech recognition. Through these experiments, the maximum error reduction rate is 23% at the center position with 40 cm distance.

Table 8. The improvement of speech recognition under different SNR environments.

	One channel output		Eight microphone-DSBF output	
	SNR [dB]	Recognition rate [%]	SNR [dB]	Recognition rate [%]
Test 1	20.0	54.2	22.4	61.4
Test 2	14.1	41.1	20.8	50.4

The improvement of speech recognition by microphone array DSBF-based speech enhancement under different SNR environments is given in Table 8. As described previously, the speech database used in these experiments are uttered in front of the microphone array by five male speakers and collected directly through the microphone array. The error reduction rates are 15.8% and 15.7% for Test 1 and Test 2, respectively while the gain of SNR shows large difference for two tests. We think this large difference in SNR gain is due to the characteristics of environmental noises. We collect speech database used in Test 1 just in the normal office environments without generating particular noises. On the contrary, the speech database used in Test 2 contain the noises generated from two loud-speakers. So, there is some difference between the two noises. The results about improvement of speech recognition rate indicate that microphone array-based speech enhancement is generally independent of the SNR level of the input signal and produces consistent improvement in the performance of speech recognition.

Ⅳ. Conclusion

We proposed a remote speech input method 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 DSBF-based speech enhancement. The automatic speech detection module detects speech portion from the incoming signals. The percentages of detecting the start point of speech signal within 200 ms from the hand labelled points are 93%, 99%, and 99% under 15 dB, 20 dB, and 25 dB SNR environments and those for the end point are 72%, 89%, and 93% for the corresponding environments, respectively. We also adopt the speech and nonspeech classifier for the start point detected region of input signal to decide the detected speech region is real speech or noise. This speech and nonspeech classification is performed by the pitch information-based method and the accuracy for the speech is 99% and that of nonspeech input is 90%. The microphone array-based speech enhancement using the DSBF algorithm shows maximum SNR gain of 6 dB over a single microphone and error reduction of more than 15% in speech recognition.

By integrating these two modules, we introduce the remote speech input method that is very efficient and convenient for speech input. We think this remote speech input method can be widely used for the applications that require user friendliness as well as noise reduction.

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

Acknowledgements

This research was funded by the Korea Ministry of Information and Communications. The authors wish to thank all members of the Spoken Language Processing Section in ETRI for their support.

References

1. Y. Lee and K. Hwang, "Selecting good speech features for recognition," *ETRI Journal*, vol. 18, no. 1, pp. 29-40, 1996.
2. J-W. Yang and Y. Lee, "Toward translation Korean speech into other languages," *Proc. ICSLP96*, vol. 4, pp. 2368-2370,. 1996.
3. Y. J. Suh, J. Park, and Y. Lee, "A user friendly remote speech input in spontaneous speech translation system," Proc. ESCA-NATO Workshop on Robust Speech Recognition, pp171-174, 1997.
4. S. U. Pillai, Array Signal Processing, Springer-Verlag, New York, 1989, pp. 8-20.
5. R. P. Ramachandran and R. J. Mammone, "Microphone array for hands-free voice communication in a car," Modern Methods of Speech Processing, KAP, Boston, 1995, pp. 351-375.
6. 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, 1985.
7. D. Giuliani, M.Matassoni, M. Omologo, and P. Svaizer, "Robust continuous speech recognition using a microphone Array," *Proc. EUROSPEECH*, vol. 3, pp. 2021-2024, 1995.
8. H. Lee and M. Hahn, "Development of a real-time endpoint detection algorithm," *Proc. ICSPAT*, vol. 2, pp. 1547-1553, 1993.
9. G. Clifford Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, no. 2, pp.236-255, 1987.
10. M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 288-292, 1997.
11. L. J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antenn. Propagat.*, vol. AP-30, pp. 27-30, 1982.

▲Young-Joo Suh

Youngjoo Suh received the B.S. and M.S. degrees in electronics engineering from Kyungpook National University, Taegu, Korea in 1991, 1993, respectively. Since 1993, he has been with Electronics and Telecommunications Research Institute. His current research interests are speech recogntion, and neural networks.

▲Jun Park

Jun Park received the B.S. and M.S. degrees in electronics engineering from Seoul National University in 1981 and 1983, respectively, and Ph.D. degree in electrical engineering from University of Southern California, U.S.A. in 1994. Since 1983, he has been with Electronics and Telecommunications Research Institute, and participated in many R&D projects in the areas of electronic switch, centralized switch maintenance system, and speech recognition system. His current research interests are speech recogntion, speech synthesis, and neural networks.

▲Young-Jik Lee

Youngjik Lee received the B.S. degree in electronics engineering from Seoul National University, Seoul, Korea in 1979, the M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Seoul, Korea in 1981 and the Ph.D. degree in electrical engineering from the Polytechnic University, Brooklyn, New York, USA.

From 1981 to 1985 he was with Samsung Electronics Company, Suwon, Korea, where he had developed video display terminals. From 1985 to 1988 he had worked on sensor array signal processing. Since 1989, he has been with Electronics and Telecommunications Research Institute pursuing researches in neural network, pattern recognition, digital signal processing, speech recognition, speech synthesis, and speech translation.