Journal of Institute of Control, Robotics and Systems (2010) 16(12):1143-1149
DOI:10.5302/J.ICROS.2010.16.12.1143 ISSN:1976-5622

스마트 시티에서의 이머전시 사운드 감지방법

A Emergency Sound Detecting Method for Smarter City

조 영 임*
(Young Im Cho[1])
[1]The University of Suwon

Abstract: Because the noise is the main cause for decreasing the performance at speech recognition, the place or environment is very important in speech recognition. To improve the speech recognition performance in the real situations where various extraneous noises are abundant, a novel combination of FIR and Wiener filters is proposed and experimented. The combination resulted in improved accuracy and reduced processing time, enabling fast analysis and response in emergency situations. Usually, there are many dangerous situations in our city life, so for the smarter city it is necessary to detect many types of sound in various environment. Therefore this paper is about how to detect many types of sound in real city, especially on CCTV. This paper is for implementing the smarter city by detecting many types of sounds and filtering one of the emergency sound in this sound stream. And then it can be possible to handle with the emergency or dangerous situation.

Keywords: emergency, wiener filter, noise filtering, smarter city

I. INTRODUCTION

The success factor of the smart city as the next stage in the process of urbanization is how to design the role of ICT (Information Communication Techniques) infrastructure. But much research has also been carried out on the role of human capital/education, social and relational capital and environmental interest as important drivers of urban growth [1].

Recently, smarter city is a full-service communications provider across the nation. Smarter City provides technologies that make our cities smarter places to work, live, and play [2].

In this paper we are concentrated on the speech recognition system which is necessary in the emergency system as a main part of a smarter living in smarter city. Because we have many dangerous situation in real world. So, this paper is about how to detect many emergency types of sound in real city life, especially on CCTV. This paper is for implementing the smarter city by detecting many types of sounds and filtering one of the emergency sound among the sound stream. And then it can be possible to deal with the dangerous or emergency situation. For example, we can go there immediately after detecting the emergency or screaming sound(eg."ah..., please help me..., thief..."). The goal of this paper is that to detect the emergency sound and make someone who it is concerned with handled the situation immediately without hesitation. For convenience sake, we adopt a CCTV to collect many types of sound in laboratory environment. Because there are many types of noise in outside.

So, this simulation in this paper is concentrated on our laboratory environment at firstly. But next time after sucess inside speech recognition, we can apply it to the outside world.

Usually, to the extent that the speech is the basic tool in human communications, it is also being studied as an effective means of communications between various digital devices and humans in the ubiquitous environment [3,4]. One of the key factors is the noise. The noise is omnipresent. That is, it is present in indoors as well as outdoors. The noise poses greater problems in emergency situations, especially to be a smarter city, where a fast interpretation of the speech data is critical.

In speech recognition, the real situation is quite different from the controlled environment of the speech laboratory. Various factors, such as non-linearities in the microphone, surrounding noises, and the differences in the distance of the sound sources, undermine speech recognition in the real environment. The surrounding noises are a particularly difficult problem in that the sources are quite diverse: some human, some mechanical, and some natural. When superimposed on the speech, these extraneous noises can significantly reduce the accuracy of the system and could, in fact, lead to the misinterpretation of the speech data.

The difference in the controlled environment and the real environment in speech recognition comes into play in three

* 책임저자(Corresponding Author)
논문접수: 2010. 9. 10., 수정: 2010. 10. 7., 채택확정: 2010. 12. 1.
조영임: 수원대학교 컴퓨터학과(ycho@suwon.ac.kr)
※ This paper is supported by Gyeonggi-do Regional Research Center (GRRC)[(GRRC suwon2010-B3), Development of an Intelligent Speech Recognition and Information Retrieval System)].

distinct processes: signal process, feature space process, and model process. Of these three processes, the difference is most evident in the signal process [5,6].

In this work, utilizing MATLAB [7], the noise in the speech data after the signal process is filtered by a novel digital filtering system. Specifically, a FIR filter is first used to separate the speech region and the noise region, and then a Wiener filter is used to improve the overall speech recognition. In what follows, the speech recognition system and the digital filtering devised in this study is introduced in chapter II, a noise filtering scheme is proposed in chapter III, the experimental results are presented and discussed in chapter IV, and the conclusions are drawn in chapter V.

II. SPEECH RECOGNITION SYSTEM

Based on the flat lexicon and the lexical tree, the speech recognition system built in this work is designed for speedy interpretation of the voice data and is connected to CCTV's for visual in situ inspection [8].

As illustrated in Fig. 1, the speech recognition system is structured into six stages. In stage 1, voice data are inputted by converting the audio signals into the electrical digital signals. In stage 2, the voice signals are separated from the surrounding noises. In stage 3, useful traits in speech recognition are extracted by using a speech recognition model. In stage 4, a standard speech pattern database is formed by speech recognition training. In stage 5, new voice data are compared to the standard speech pattern database, and the closest match is searched. In the final stage of 6, the matched result is put to use through the user interface.

In the preprocessing (noise elimination) stage of 2, analog audio signals from a CCTV or a sensor are digitized and then fed to the digital filter. The digital filter, which is widely used and proven, selects the passband and filter out the stopband.

Depending on the presence of feedback processes, the digital filter is divided into IIR (Infinite Impulse Response) and FIR (Finite Impulse Response) filters. The latter is known to be less error-prone. For noise elimination in the subsequent processes, the Wiener and Kalman filters [9] are widely used. In emergency situations that require accurate

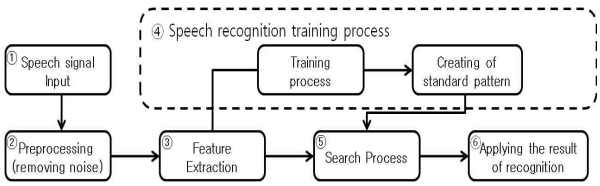

그림 1. 음성인식의 일반구조.

Fig. 1. The general structure of speech recognition system.

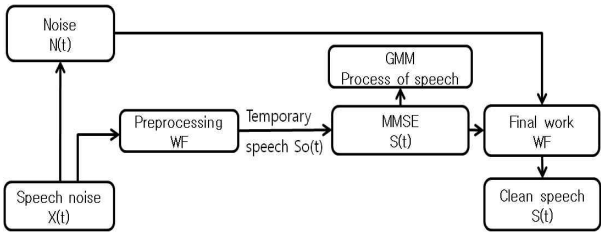

그림 2. 모델기반 위너필터 구조.

Fig. 2. The structure of model based Wiener filter.

interpretation of a rather brief voice data, the Wiener filter is usually preferred. The structure of the model-based Wiener filter is illustrated in Fig. 2 [10,11].

The general model-based Wiener filtering process can be expressed as follows:

$$\hat{s}(t) = g(t) * (s(t) + n(t)) \tag{1}$$

where $\hat{s}(t)$ is the speech to be recognized, $s(t)$ is the speech data containing noise, $n(t)$ is the noise, and $g(t)$ is the Wiener filter.

In Eq. (1), $\hat{s}(t)$ is being sought. In it, an estimate of $n(t)$ is derived from $s(t)$, and then the approximate value of $\hat{s}(t)$ is obtained by using $n(t)$. In order to achieve a better approximation of $\hat{s}(t)$, the GMM as expressed below in Eq. (2) is used. It expresses mathematically the general characteristic of speech data.

$$P(s) = \sum_{k}^{K} p(k) N(s; \mu_k; \sum k) \tag{2}$$

Based on Eq. (2), the model-based Wiener filter is designed per following steps:

① In the inputted current frame, the noise region is determined by a statistically-based VAD. In the noise region found, the noise model is renewed to the previous value.

② In the preprocess-WF block, a temporally noise-free clean speech is estimated using the decision-directed Wiener filter.

③ Using the estimated values from the previous step, the Gaussian post probabilities of the GMM are calculated. In the final WF using the MMSE method, the probabilities are used to estimate the noise-free clean speech.

④ The estimated noise-free speech and the noise model of step ① are used to design the final Wiener filter.

⑤ The current frame is processed using the Wiener filter designed, and the noise-free clean speech is obtained. Then the above five steps are repeated for the next frame.

Using the speeches obtained from the above process, the stages 3-6 (speech characteristics extraction, speech recognition training, search process, result and user interface, respectively) in Fig. 1 are then followed.

III. IMPROVED NOISE FILTER

In this chapter, a new improved noise-filtering method is proposed. The basic premise is to selectively use the audio signal being transmitted from the CCTV's. That is, from the transmitted signal, only the audio energy spectrum that is relevant to the speech is to be selected, digitized, and saved for further analysis. A high-performance FIR Wiener filer can be used to digitally filter out the unwanted portion of the audio signal, prior to actual speech recognition.

As human speech generally falls within 300-3400khz, the FIR filter [12-14] can separate the incoming audio data into passband (the speech region), stopband, and threshold-band. This will greatly reduce the time and improve the performance of a speech recognition system.

The basic mathematical concept of the FIR filter can be expressed as follows:.

$$y[x]=\sum_{k=0}^{N-1} h[k]x[n-k] \tag{4}$$

In Eq.(4), $x[n]$ is the speech information input, $y[n]$ is the output speech information after filtering, $h[n]$ is the finite impulse response characteristic, and N is the filtering step number. As the input information and coefficients are multiplied and summed, the time required for noise filtering is quite long if Eq. (4) is implemented as is. However, the multiplication steps in Eq. (4) can be eliminated if a bit-serial algorithm [15] is applied. The result is expressed in Eq. (5) below:

$$y[x]=\sum_{k=0}^{N-1}(\sum_{j=0}^{M-1} h_j[k] \cdot 2j)x[n-k] \tag{5}$$

Here, h_j, N, and M represent the coefficient h's jth bit, tab number, and coefficient bit number, respectively. The bit-serial algorithm multiplies multiplicand to the multiplier while shifting LSB to MSB and then adds the result to the previous sum. To reduce the total multiplication cycles, the odd and even part of the Eq. (5) can be separated and the result can be written as follows:

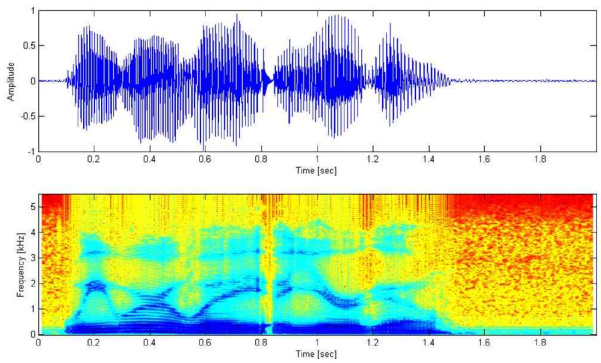

그림 3. FIR 필터링의 빈도수와 진동.

Fig. 3. The frequency and oscillation of FIR filtering.

$$y[x]=\sum_{k=0}^{N-1}\sum_{j=0}^{\frac{M}{2}-1}(h_{2j}[k] \cdot 2^{2j}+h_{2j}[k] \cdot 2^{2j+1})x[n-k] \tag{6}$$

Eq. (5) requires a total of NM multiplication cycles, while Eq. (6) requires a total of $\frac{NM}{2}$ multiplication cycles, a factor of 2 increase in the speed.

In this way, utilizing the benefits of the FIR filter, an Wiener filtering that minimizes the noise error by effectively separating the speech signal and the noise is implemented.

Afterwards, the noise signals are extracted by subtracting the output speech data from the incoming speech data. Then the extracted noise data and the incoming speech data are used in Eq. (1) to design an improved noise filter.

Fig. 3 illustrates the result in which the incoming audio wave is separated into the speech region and others by the FIR filter, so that the overall processing time can be reduced during the Wiener filtering.

In terms of mathematical expression, the general Wiener filter consists of multiplications and summations of current and past data and filtering coefficients. Thus, it can be designed using the device transfer functions and the mathematical expressions. Within the scope of this research, the physical states, such as operational stability and sensitivity and the safe transmission of data, are assumed to be steady, and the main priority is placed on minimizing the number of devices and increasing the speed of the filter operation.

Finally, the noise elimination Wiener filter is expressed as follows:

$$So(w) = H(w)S(w) \tag{6}$$

In Eq. (6), S(w) is the noise-containing speech signal, So(w) is the noise-free speech signal, and H(w) is estimation function of the Wiener filter. An effective way to determine H(w) is a major focus of this research. Accordingly, an mathematical expression for H(w) is proposed as follows:

$$H(w) = \frac{P_s(w)}{P_s(w)+P_d(w)} \tag{7}$$

Here, Ps(w) is the audio spectrum of the original speech signal, and Pd(w) is the audio spectrum of the noise signal. An error is introduced in estimating the audio spectrum of the original speech signal during the filtering process. To reduce the error, a coefficient is introduced as below:

$$H(w) = \left(\frac{P_s(w)}{P_s(w)+\alpha \cdot P_d(w)}\right)^{\beta} \tag{8}$$

Here, parameters a, β and squaring the averages of the

signals are used to reduced the error.

The Wiener filtering processes the noise-containing speech information effectively, but it takes time, so that speech recognition is delayed. To minimize the time delay, the concept expressed in Eq. (8) is applied during the statistically-based VAD process [12] of stage 1 in Fig. 2. The resulting process model is expressed in Eq. (9) below. In this model, the speech data and the noise data are considered to be asymmetric. By applying asymmetric window to these two data in designing the Wiener filter, the time required for noise filtering can be significantly reduced.

$$H(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{P_1}\right), & 0 \le n < n_0\\ \cos\left(\dfrac{2\pi(n-n_0)}{P_2}\right), & n_0 \le n < N\end{cases} \quad (9)$$

In Eq. (9), P1,P2, respectively, represents the period of the left and right portions of the asymmetric window function, n0 is the location of the maximum value, and N is the total length of the window function.

Based on the noise-free speech signal obtained thus far, a speech recognition database is complied and used in analysis of emergency situations. In this effort, the phonemic recognition of individual words is initially chosen as the key element of speech recognition, and the database is complied accordingly.

Fig. 4 illustrates the overview of the speech recognition system founded on the concepts described in Fig. 1. MATLAB is used for the proposed FIR Wiener filter, and HTK and ECHOS are used for the subsequent processes. The finished speech recognition system basically uses sound models to search key words, and the flat lexicon and lexical tree are used in this word-based speech recognition system. The lexical tree is efficient in the usage of the

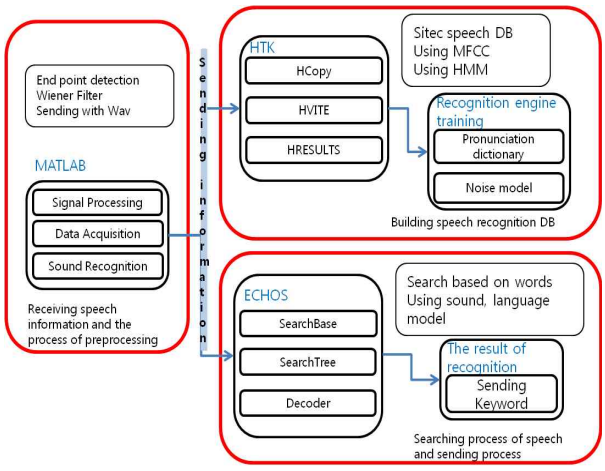

그림 4. 제안하는 음성인식 시스템의 개요.

Fig. 4. The overview of proposed speech recognition system.

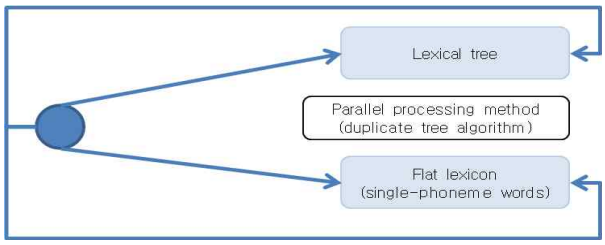

그림 5. 렉시칼 트리문제의 해결책.

Fig. 5. The solution strategy of lexical tree problem.

memory, but is somewhat slow in applying the probability values of the language models and is also somewhat complex in implementing word models. Thus, a duplicate tree algorithm is used. That is, a parallel structure is used in the lexical tree for single-phoneme words.

The speech recognition result obtained by these serious of processes is then sent to the user interface of the system.

IV. RESULTS AND DISCUSSION

To test the speech recognition system developed thus far in this work, speaking word database developed by SITEC is used. The database is recorded in 16khz/16 bit, and contained the voices of 500 individuals. For comparison to the database, voice data from a microphone or CCTV in 16khz/16 bit format are used.

As discussed in the introduction, it would be unwise to use all the collected voice information in speech recognition, as it will consume too much time and may

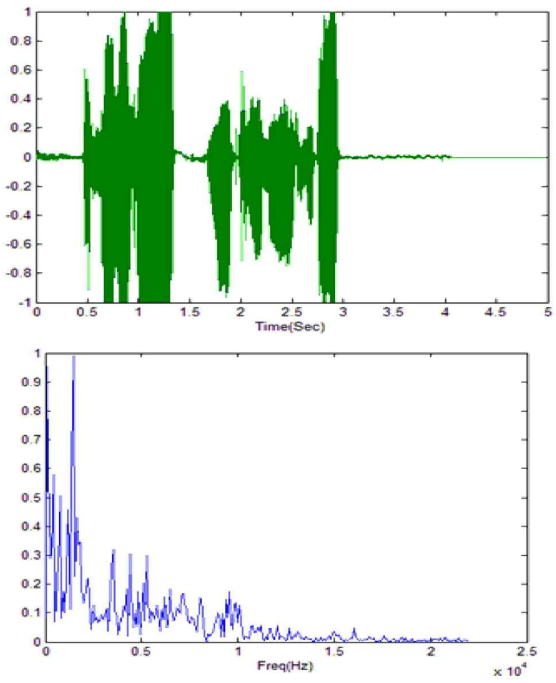

그림 6. FIR 필터링 사용 전의 음성정보.

Fig. 6. The speech information before using FIR filtering.

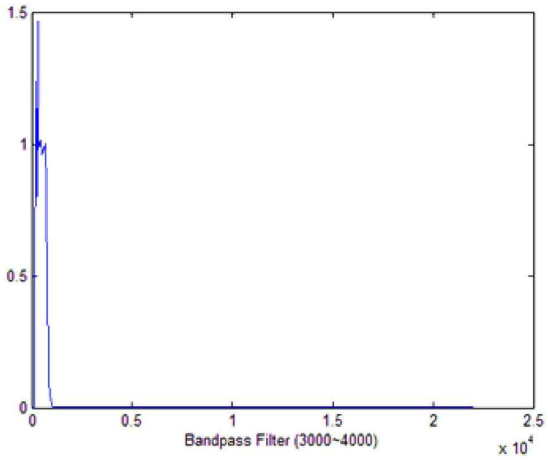

그림 7. FIR 위너필터의 파형.

Fig. 7. The waveform of FIR Wiener filter.

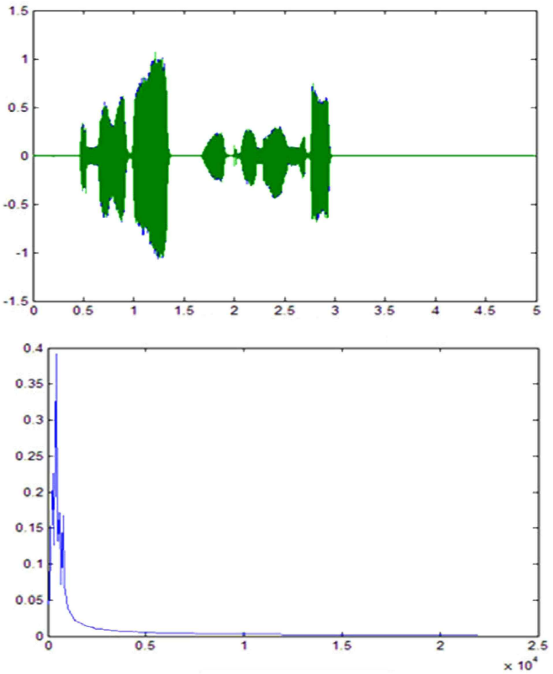

그림 8. FIR 필터링 사용 후의 음성정보.

Fig. 8. The speech information after using FIR filtering.

result in inaccurate analysis. By first extracting the audio frequency region useful for speech recognition by using the FIR filter proposed in chapter 3, the overall processing time can be greatly reduced.

The noise that escapes the initial filtering will then be eliminated by the Wiener filter. Fig. 6 illustrates the passband spectrum and frequency regions of the speech information passed through the first stage of the FIR filter. Fig. 7 illustrates the waveform of the Wiener filter.

Finally, noise-free speech information obtained by the Wiener filter is illustrated in Fig. 8.

By comparing Figs. 7 and 8, it is found that the FIR Wiener filter visibly eliminates the background noise. Also,

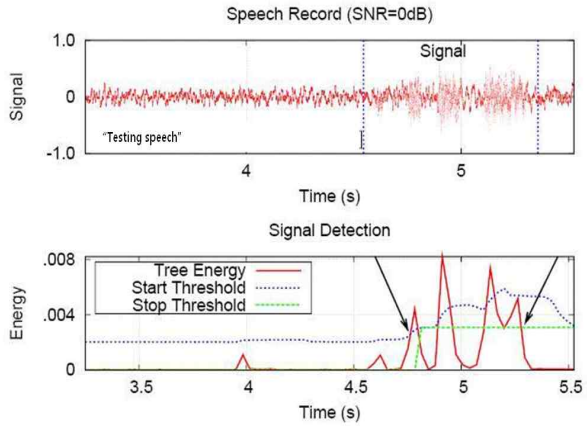

그림 9. 시그널 감지 테스트.

Fig. 9. Signal detection test.

the effect can be audibly felt by listening to the before and after sounds.

By comparing the noise-free speeches processed through the MATLAB-constructed Wiener filter, as illustrated in Fig. 1, to the existing database of key words, a very accurate speech recognition effect was realized (Fig. 9).

In this simulation in Fig. 9, SNR is 0. That is to say, our simulation environment is in our laboratory. We simulated in a environment without noise for detecting emergency sound. As we mentioned earlier, our goal is to detect some emergency sound among the stream of incoming many types of sounds. And then we can make it handle immediately the emergency situation by letting someone who it is concerned know the emergency situation.

By filtering out the unnecessary portions of the speech information, such as non-audible frequency regions, environmental noise, and transmission noise from the CCTV's, the level of speech recognition is found to be greatly increase. Also, the word models are found to be very useful in the success of speech recognition and in the reduction of the processing time. That is, the word model that considers the relationships between the words being searched had much more successes.

Using the database complied with noise-free sound data, two-pass bigram and bigram + trigram searches under ECHOS were found to indicate that the word-correlated model was much superior in terms of the recognition success rate and the search time than the simple model that does not consider the relationships between words(Eq.10). The results are summarized in Table 1.

$$P(w_n|w_1 \ldots w_{n-1}) = \frac{P(w_{1\ldots}w_n)}{P(w_1 \ldots w_{n-1})} \quad (10)$$
$$(where, n=2: bigram, n=3: trigram)$$

P is the probability to estimate the next word in n-gram. It is interesting to note that in HMM the success rate for

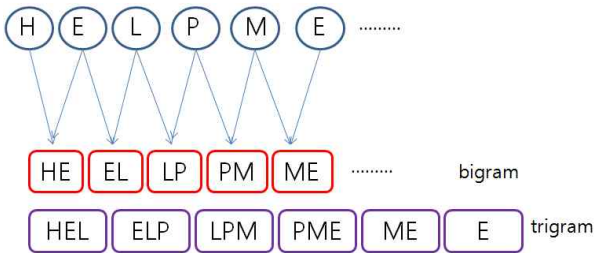

그림 10. 시뮬레이션에서 사용된 bigram과 trigram모델.
Fig. 10. Bigram and trigram model used in simulation.

표 1. 제안된 시스템의 인식결과
Table 1. The recognition result of proposed system.

Way of search	Using the model for each words	Speech recognition rate (%)	Recognition time (sec/sentence)
bigram	X	77.2	5.4
bigram + trigram	X	80.1	6.3
bigram	○	88.9	21.0
bigram + trigram	○	90.0	22.1

the standard bigram (left to right) search method in Fig. 10 is lower by 8% than the trigram search method that searches in the reverse direction and also considers the relationships between different phonemes. Nonetheless, the search time was longer for the latter method.

Lastly, a significant processing time reduction and a fast situation response were realized by selectively processing the audible voice region of the audio signals transmitted from the CCTV.

V. CONCLUSION

Unlike the controlled environment where a speech recognition system can easily filter out the extraneous noises, it is rather difficult in the real environment where a sensor, such as a CCTV, collects abundant noises from various human, mechanical, and natural sources. The success of speech recognition in the real environment thus depends critically on how well these noises are filtered. Just as important, the processing time for noise filtering needs to be reduced, as time is the most critical element in emergency situations. Thus, effective noise filtering combined with fast processing time is considered to be the essence of speech recognition. Towards these goals, an improved speech recognition system is proposed in this work. The system has the FIR and Wiener filters as the key elements and effectively filters out the extraneous noises and produces clean noise-free speech data in a reasonable time.

One of the problems cited during the work is that the extraneous noise that is present in the audible band of 300-3400khz can still pose some problems even with the proposed FIR filter. As the noise filtering in this frequency region is not yet completely understood, further research in this front is currently underway.

REFERENCES

[1] M. Deakin, “From city of bits to e-topia: taking the thesis on digitally-inclusive regeneration full circle,” *Journal of Urban Technology,* vol. 14, no. 3, pp. 131-143. 2007.

[2] K. Nicos, *Intelligent Cities: Innovation, Knowledge Systems and Digital Spaces*. London: Spon Press, 2002.

[3] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, “Toward conversational human-computer interaction,” *AI Magazine*, vol. 22, no. 4, pp. 27-37, 2001.

[4] H. Kruegle, “CCTV surveillance,” *Analog and Digital Video Practices and Technology*, Elsevier, pp. 227-239, 2007.

[5] Y. Gong, “Speech recognition in noisy environments: a survey,” *Speech Communication,* vol. 16, no. 3, pp. 261-291, Apr. 1995.

[6] C.-H. Lee, “On stochastic feature and model compensation approaches to robust speech recognition,” *Speech Communication,* vol. 25, no. 1-3, pp. 29-47, Aug. 1998.

[7] K. S. Kim, “MATLAB signal and image processing,” Ajin Publishing, Korea, pp. 213-250, 2007.

[8] Y. I. Cho and S. S. Jang, “Intelligent speech recognition system for CCTV surveillance,” *Korea Journal of Intelligent Systems,* vol. 19, no. 3, pp. 415-420, 2009.

[9] J. K. Kim, “Min/Max estimation and base esitmation for Kalman filter,” *Natural Science Research* (Korean), vol. 5, pp. 21-30, 1995.

[10] J. J. Kang, B. O. Kang, H. Y. Jung, H. Jung, and Y. K. Lee, “Long words speech recognition technology and trend,” *Electronic Communication Trend Analysis* (Korean), vol. 23, no. 1, pp. 70-76, 2008.

[11] S. Doclo, Rong Dong, T. J. Klasen, J. Wouters, S. Haykin, and M. Moonen, “Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids,” *Applications of Signal Processing to Audio and Acoustics,* vol. 16, no. 16, pp. 70-73, 2005.

[12] J. H. Jang, D. K. Kim, and N. S. Kim, “A new statistical method for speech recognition systems,” *Telecommunications Review* (Korean), vol. 15, no. 1, pp. 201-209, 2005.

[13] T. K. Ryu, K. H, Park, D. S. Hong, and C. O. Kang, “Channel estimation by sero-forcing method in the frequency region,” *Korea Journal of Telecommunications*,

vol. 31, no. 1, pp. 38-47, 2006.

[14] Y. S. Park and J. H. Jang, “Echo filtering by soft decision in the frequency region,” *Telecommunications Review* (Korean), vol. 19, no. 5, pp. 837-844, 2009.

[15] R. E. Morley, Jr. Gray E. Christensen, T. J. Sullivan, Orly Kamin, “The Design of a bit-serial coprocessor to perform mulitiplication and divison on a massively parallel architecture,” *in Proc IEEE, The 2nd Syposium on the Frontiers of Massively Parallel Computation,* Farifax, USA, pp. 419-422, 1998.

조 영 임

1988년 고려대학교 컴퓨터학과 학사. 1990년 고려대학교 컴퓨터학과 석사. 1994년 고려대학교 컴퓨터학과 박사. 1995년~1996년 삼성전자 선임연구원. 1999년~2000년 Univ. of Massachusetts, post-doc. 관심분야는 에이전트 시스템, 인공지능, 음성인식, 유비쿼터스 시스템 등.