Intelligent Speech Recognition Filtering

Young Im Cho

Dept. of Computer Science, Univ. of Suwon, San2-1, Wau-ri, Bongdam-eup, Hwaseong, Gyeonggi-do
(Tel : 82-31-229-8214; Fax : 82-31-229-8281)
(ycho@suwon.ac.kr)

***Abstract*: In Emergency situation, speech recognition speed is very important. Therefore, in this paper, we propose a fast filtering algorithm. Firstly, FIR filter selectively passes through the frequency range of speech, and secondly, the Wiener filter filters out the extraneous noises. Because of that, the processing time is reduced.**

***Keywords*: FIR filter, Speech recognition, Combination filtering algorithm**

I. INTRODUCTION

One of the key factors in the speech recognition is noise[1]. The real situation is quite different from the controlled environment of the speech laboratory. However, the surrounding noises are a particularly difficult problem in the real speech recognition. The difference in the controlled environment and the real environment in speech recognition comes into play in three distinct processes: signal process, feature space process, and model process. Of these three processes, the difference is most evident in the signal process [2].

Here, the noise in the speech data after the signal process is filtered by a novel digital filtering system. 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.

II. Combined Filtering

Generally, the speech recognition system is configured by 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.

This paper is supported by Gyeonggi-do Regional Research Center in Korea (suwon GRRC 2009-B3)

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[3] are widely used. In emergency situations that require accurate interpretation of a rather brief voice data, the Wiener filter is usually preferred. 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: 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. In the preprocess-WF block, a temporally noise-free clean speech is estimated using the decision-directed Wiener filter. 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. The estimated noise-free speech and the noise model are used to design the final Wiener filter. 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.

For emergency detection, we propose a fast recognition filtering method. The basic concept 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 [4,5] 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:.

$$y[x] = \sum_{k=0}^{N-1} h[k]x[n-k] \tag{3}$$

In Eq.(3), is the speech information input, is the output speech information after filtering, 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. (3) is implemented as is. However, the multiplication steps in Eq. (3) can be eliminated if a bit-serial algorithm [6] is applied. The result is expressed in Eq. (4) below:

$$y[x] = \sum_{k=0}^{N-1} (\sum_{j=0}^{M-1} h_j[k] \cdot 2_j) x[n-k] \tag{4}$$

Here, h_i, N, and M represent the coefficient h's j th 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. (4) can be separated and the result can be written as follows:

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

Eq. (4) requires a total of multiplication cycles, while Eq. (5) requires a total of multiplication cycles, a factor of 2 increase in the speed.

In this way, utilizing the benefits of the FIR filter, a 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.

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, $s_0(w)$ 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, a mathematical expression for $H(w)$ is proposed as follows:

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

Here, $P_s(w)$ is the audio spectrum of the original speech signal, and $P_d(w)$ 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:

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

Here, parameters α, β and squaring the averages of the signals are used to reduce 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[7] of stage. 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.

$$H(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{p_1}\right), 0\le n\le n_0\\ \cos\left(\dfrac{2\pi(n-n_0)}{p_2}\right), n_0\le n\le N\end{cases} \quad (9)$$

In Eq. (9), P_1, P_2, respectively, represents the period of the left and right portions of the asymmetric window function, n_0 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.

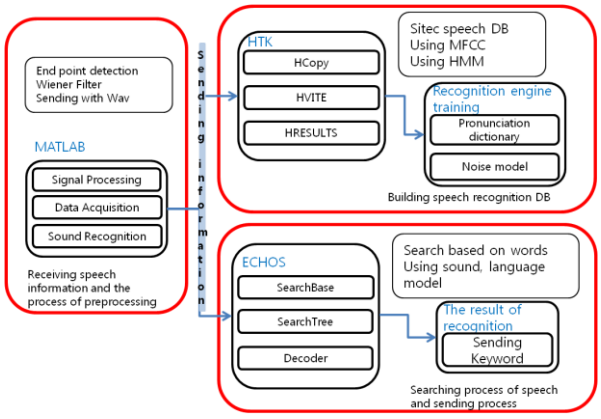

Fig.1. The overview of proposed speech recognition system

Fig. 1 illustrates the overview of the speech recognition system founded on the concepts described in Fig. 1. MATLAB[8] 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 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

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[9].

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

The noise that escapes the initial filtering will then be eliminated by the Wiener filter. By comparing two filtering, it is found that the FIR Wiener filter visibly eliminates the background noise. Also, 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, to the existing database of key words, a very accurate speech recognition effect was realized.

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 increased. 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.

As a result, in case of using model for each words, the speech recognition rate is much faster than without model. In model case, it is about 88.9% in bigram and 90.0% in bigram and trigram combination model of speech recognition rate(%) respectively. However, in without model case, it is about 77.2% in bigram and 80.1% in bigram and trigram combination model of speech recognition rate(%) respectively. Also, the recognition time(sec/sentence) is much faster than without model. It is about 21.0% in bigram and 22.1% in bigram and trigram combination model of speech recognition rate(%) respectively. However, in without model case, it is about 5.4% in bigram and 6.3% in bigram and trigram combination model of speech recognition rate(%) respectively.

It is interesting to note that in HMM the success rate for the standard bigram (left to right) search method 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.

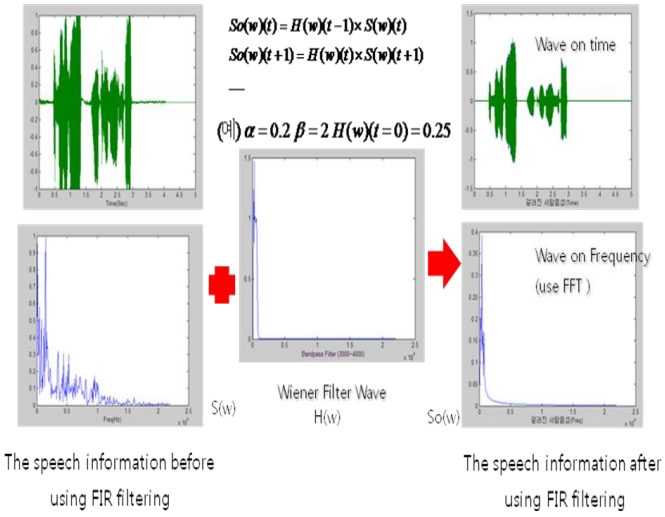

Fig.2 Simulation Result 'Hello'

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] H. Kruegle, "CCTV Surveillance", *Analog and Digital Video Practices and Technology*, Elsevier, pp.227-239, 2007.

[2] C.-H.Lee, "On Stochastic Feature and Model Compensation Approaches to Robust Speech Recognition", *Speech Communication*, pp.29-47. 1998.

[3] J.K. Kim, "Min/Max Estimation and Base Estimation for Kalman Filter", Natural Science Research (Korean), vol. 5, pp.21-30, 1995.

[4] T.K. Ryu, K.H, Park, D.S. Hong, C.O. Kang, "Channel Estimation by Sero-Forcing Method in the Frequency Region," Kor. J. Telecommunications, vol.31, no.1, pp.38-47, 2006.

[5] Y.S. Park, J.H. Jang, "Echo Filtering by Soft Decision in the Frequency Region", *Telecommunications Review* (Korean), vol.19, no.5, pp.837-844, 2009.

[6] Robert E.Morley, Jr. Gray E. Christensen, Thomas 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, U.S.A*, pp.419-422, 1998

[7] J.H. Jang, D.K. Kim, N.S. Kim, "A New Statistical Method for Speech Recognition Systems", *Telecommunications Review* (Korean), vol.15, no.1, pp. 201-209, 2005.

[8] K.S. Kim, "MATLAB Signal and Image Processing", Ajin Publishing, Korea, pp.213-250, 2007.

[9] Y.I. Cho and S.S. Jang, "Intelligent Speech Recognition System for CCTV Surveillance", Kor. J. Intelligent Systems, vol.19, no.3, pp.415-420, 2009.