A Robust Non-Speech Rejection Algorithm

*Young-Mok Ahn

Abstract

We propose a robust non-speech rejection algorithm using the three types of pitch-related parameters. The robust non-speech rejection algorithm utilizes three kinds of pitch parameters : (1) pitch range, (2) difference of the successive pitch range, and (3) the number of successive pitches satisfying constraints related with the previous two parameters. The acceptance rate of the speech commands was 95% for −2.8 dB signal-to-noise ratio (SNR) speech database that consisted of 2440 utterances. The rejection rate of the non-speech sounds was 100% while the acceptance rate of the speech commands was 97% in an office environment.

Ⅰ. Introduction

Non-speech sound rejection represents an important technology in the design of speech recognition systems. The ability to detect non-speech sound such as phone ring is essential to improve the practical performance as well as the user friendliness of a speech recognizer [1]. The performance of a non-speech sound rejection method with conventional endpoint detection algorithms is usually not good in real environments, because energy level and level-crossing rate (LCR) can not be detected reliably. It is desirable to detect the non-speech sound right after an endpoint detector which allow us to avoid the undesirable processing such as feature extraction, template search. The detection of non-speech sound eliminate some incorrect working from the recognition systems.

In a hidden Markov model(HMM)-based speech recognition system, garbage models (also referred to as filler models) or clustering techniques of noise classes have been used to model the non-speech sounds [2, 3]. To get accurate non-speech models and to achieve acceptable detection performance in the HMM-based system, it is indispensable to collect sufficient training data. In those cases, the lack of training data is a major problem.

Many algorithms have been developed for robust and accurate pitch detection. However, the pitch detector (or fundamental frequency detector) used in the non-speech rejection algorithm should have small computational requirements for realtime processing because the non-speech rejection should be processed before feature extraction for the input signal. The pitch detector helps non-speech detector to detect non-speech sounds by providing the pitch-related parameters of the input signal.

The contributions of this paper concern a new non-speech rejection algorithm. In order to detect the non-speech sounds, we use three kinds of pitch information that can distinguish between non-speech and speech sounds.

Ⅱ. Non-Speech Rejection Algorithm

The non-speech rejection algorithm was made up of the pitch characteristics in a speech command. In voiced speech, pitch (or fundamental frequency F_0) can be easily observed by pitch detection algorithms. The variability of F_0 is tightly connected with the manner of articulation. It is also associated with many other factors such as gender, language, and speech act. In our experiments, we have observed that when a user utters to speech recognition systems, the values of F_0 are concentrated within a smaller range than other situations, such as singing or shouting. From this phenomenon we can define the range of F_0 for speech input. If some part of the input signal is a speech sound, the Nth pitch value that is $F_0(n)$ should be in the predefined boundary. This is the first condition for the non-speech sound rejection algorithm. The first condition is represented by

$$F_{0,\min} \le F_0(n) \le F_{0,\max} \qquad (1)$$

where $F_{0,\min}$ is the lower bound of the pitch range and $F_{0,\max}$ is the upper bound.

In principle, F_0 of arbitrary speech may change rapidly at any time. However, when one utters to speech recog-

*Spoken Language Processing Section, ETRI

Manuscript Received : December 5, 1997.

nition systems, the differences of successive $F_0(n)$ do not make rapid changes. The differences of successive $F_0(n)$ should be smaller than the predefined difference to be a speech sound for the input signal. This is the second condition. The second condition is represented by

$$|F_0(n-1)-F_0(n)| \leq \Delta F_{0,T} \qquad (2)$$

where $\Delta F_{0,T}$ means the tolerance of differences of successive pitch values.

The last condition is the number of pitches satisfying the previous two conditions at the same time. If the input signal is a speech, the number of successive pitches should be greater than a threshold. The last condition is represented by

$$Count \geq C_{\min} \qquad (3)$$

where *Count* is the number of pitches that satisfied with the above two conditions and $C_{\min}$ is the minimum number to be a speech command. In a real environment with various kinds of noises (e.g., key click, telephone ring, door slam, lip smacks, footstep), we observed that the number of successive pitches of the non-speech sounds was smaller than that of the speech sound.

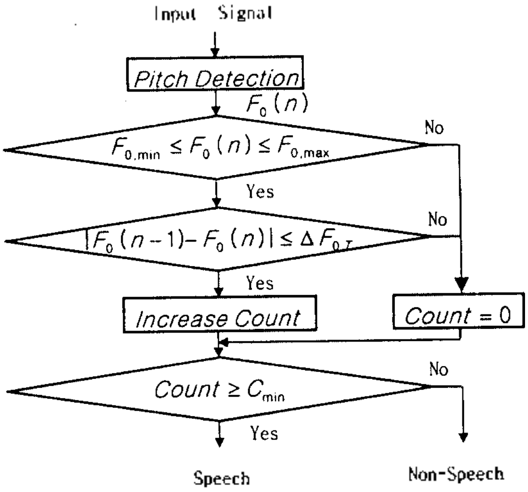


Figure 1. The block diagram of the decision logic in the non-speech sound rejection algorithm.

Fig. 1 shows the block diagram of the decision logic in the proposed algorithm. We classify the input signal into speech and non-speech sounds by checking the $F_0(n)$ range, comparing with the successive $F_0(n)$ and counting the pitches. If the input signal satisfies the above two conditions, then we start counting pitch. If the input sig-

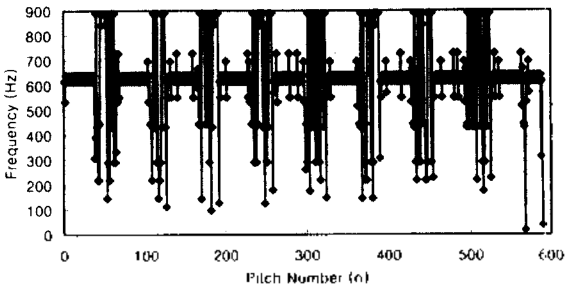


Figure 2. The values of $F_0(n)$ for the telephone ring.

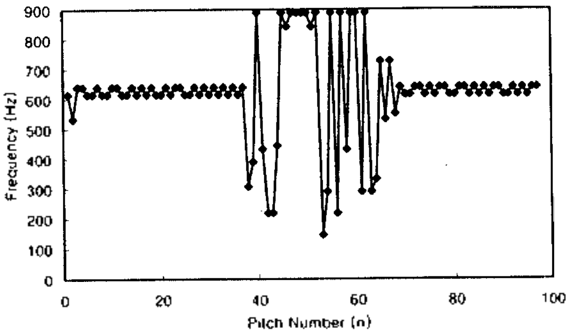


Figure 3. Expansion of Fig. 2 from $F_0(0)$ to $F_0(97)$.

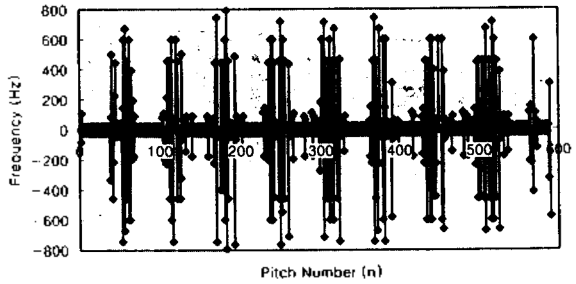


Figure 4. The differences of successive $F_0(n)$ for the telephone ring.

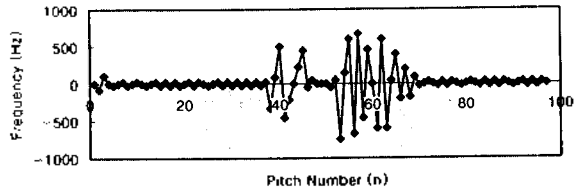


Figure 5. Expansion of Fig. 4 from $F_0(0)$ to $F_0(97)$.

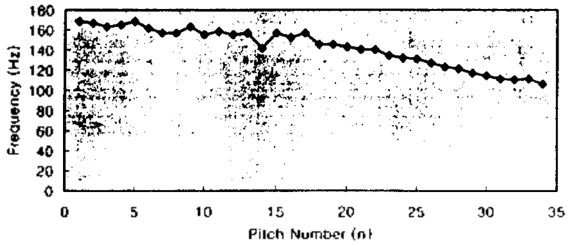


Figure 6. The values of $F_0(n)$ for speech command.

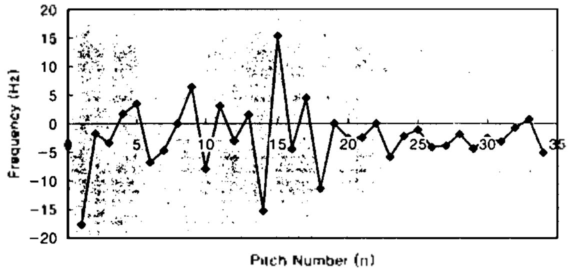


Figure 7. The differences of successive $F_0(n)$ for speech command.

nal is not satisfied with the conditions, then reset the number. If the count is greater than C_{min}, then the input signal is decided as a speech sound.

Fig. 2, 3, 4, and 5 show the pitch-related parameters for the telephone ring. In Fig. 2 and 3, most of $F_0(n)$ is greater than $F_{0,\ max}$. In Fig. 4 and 5, the difference of successive $F_0(n)$ that satisfies the first condition is greater than $\Delta F_{0,\ T}$ so that the number of count is smaller than C_{min}. Consequently, the input signal of telephone ring is rejected for speech command. The pitch-related parameters of non-speech sounds have the same characteristics like as telephone ring. Many types of noises do not satisfy the above two conditions at the same time.

Fig. 6 and 7 show the pitch-related parameters for a speech command. Under the decision logic, the maximum number of count is 15. As a result the input signal is accepted for a speech command.

In Fig. 3 and 6, we can easily distinguish between the telephone ring and the speech command for the range of F_0(n). The F_0(n) of non-speech sound have a larger value than $F_{0,\ max}$ or a smaller value than $F_{0,\ min}$. In Fig. 5 and 7, we can also distinguish between the telephone ring and the speech command for the difference of successive $F_0(n)$. In many types of noises, the difference of successive $F_0(n)$ is greater than $\Delta F_{0,\ T}$.

Ⅲ. Experimental Results

To test the usefulness of our algorithm, we have performed two types of experiments. The first experiment was performed with the speech database that consisted of 2440 utterances and the second experiment was performed in an office environment. The signal was sampled at 16 kHz with 16-bit resolution. We used $F_{0,\ min} = 80$, $F_{0,\ max} = 350$, $\Delta F_{0,\ T} = 10$, and $C_{min} = 6$.

The database for this experiment is composed of 244 isolated words, including digits related to hotel guest-side speaking and alphabet of English, spoken by 10 male speakers in the sound proof room. The white noise is added to the speech data to get the low SNR. By adding the white noise, the mean of SNR for speech database was −2.8 dB. The speech database of low SNR is used for the speech commands. The 2440 utterances of the speech database are served as the speech commands that are extracted from a speech detector. In that case, the acceptance rate was 95% so that the deletion rate of speech command was 5%.

In the experiment of an office environment, 8 male and 2 female speakers uttered 15 commands twice to a notebook PC in which the algorithm installed. They also generated 15 different non-speech sounds twice which included key click, telephone ring, door slam, lip smacks, footstep, paper rustle, cough, dragging a chair, kicking a box, knock at the desk, closing a drawer, dropping a pencil, dropping a cup, dropping a book, shaking a pencil case sounds, and so on.

The endpoint detector using frame energy and level-crossing rate finds out both speech commands and non-speech sounds. The non-speech rejector works only on the output of the endpoint detector. The rejection rate of the non-speech sounds was 100% and the acceptance rate of the speech commands was 97%.

Ⅳ. Conclusion

We proposed a new robust algorithm to distinguish non-speech from speech sounds using the three types of pitch-related parameters. Experimental results showed that the proposed algorithm was very powerful to reject many types of non-speech sounds. By rejecting non-speech sounds right after the endpoint detector, the speech recognition system is more practicable and user friendly than the conventional speech recognition system.

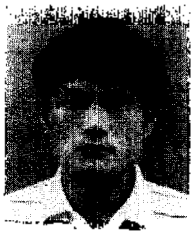
Acknowledgments

This research was funded by the Korea Ministry of Information and Communications.

References

1. Meredith, S. "Microsoft speech research program overview," in *Proc. AVIOS 97*, pp. 187-196, 1997.
2. Wilpon, J.G., Rabiner, L.R., Lee, C.H., and Goldman, E.R. "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans., ASSP*, Vol. 38, No. 11, pp. 1870-1878, 1990.
3. Schultz, T. and Rogina, I. "Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition," in *Proc. ICASSP 95*, Vol. 1, pp. 293-296, 1995.

▲Youngmok Ahn



Youngmok Ahn received the B.S. degree in electronic engineering from Hongik University in 1991. Since 1991 he has been with Electronics and Telecommunications Research Institute, where he is currently a Senior Researcher.

His research interests are speech signal processing, speech recognition and speaker recognition.