A new method to distinguish non-voice and voice in speech recognition

LI CHANGCHUN
Centre for Signal Processing
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE 639798

Abstract—we addressed the problem of remove the non-voice disturbance in speech recognition. It is always a big problem that the system will wrongly recognize our natural sound, like cough, breath, or sound of lip, nose as speech input and give "recognized " words output, when we use a speech recognition system. As we know, such non-voice speech is unavoidable for natural speaking, and if we don't supply effective control, the performance often drops to unacceptable level [1]. This paper puts forward a new method to detect fundamental frequency, and use it to distinguish real speech input and non-voice sound, like breath, lip, or noise by people walking by. Applying this method into our command recognition system, we get good results and make the system very robust and could be used in real life.

Key-words Voice distinction Auto-relation Fundamental frequency endpoint detection

1 Introduction

In speech recognition, when one only speaks what the system could recognize, no other additional noise or sound, most popular speech recognition system would work well[2]. But when we pause (not tell the system to pause too), our breath and some sounds coming from throat or nose can cause "False" speech input and give "recognized" words. Maybe, you could correct it when it is a text input system, but if a command recognition system, especially when you use your sound to control something, such error would be unbearable. Some people utilize filler models to absorb such noise, but since there are so many different non-voice sounds, it is almost impossible completely exclude them by training so many filler models.

By ways of analysis to the non-voice sound, we find that there is special character in these noises, compared with normal speech. These noises seldom have fixed (or almost) Fundamental Frequency (FF). So, we could use this property to distinguish them.

This paper paid attention to give the difference between non-voice sound and real voice, and select fundamental frequency as features. It introduced the modified FF extraction algorithm in Section 2. Its usage in voice distinction is in Section 3. In section 3, we supplied a comprehensive application of this method, combined with energy and duration feature to construct a robust system. Conclusion and remark are in the last section, Section 4.

2 Fundamental Frequency Detection

Fundamental Frequency reflects one's vocal cords. According to the mode of stimulation, sound could be divided into 3 types [3]:

1. **Vowels and semivowels**
   Vowels may be the most frequently used part in speech recognition systems in English. When speaking, the vocal cords vibrate, and produce quasi-periodic air pulse to excite fixed vocal tract shape, and then we get vowels, such as /a/ /o/, /i/, /æ/ and /u/. As for /w/, /l/, /r/, and /y/ has similar acoustic property with vowels, are called semivowels.
2. **Nasal consonants**
   Nasal consonants, like /m/, /n/. These are produced with glottal excitation, without vocal track vibration.
3. **Fricatives and stops**
   Fricatives could be divided into unvoiced /f/, /s/, and voiced like /v/, /z/. Stops also include voiced (like /b/, /d/, /g/) and unvoiced (like /p/, /t/, /k/). Stops are produced by setting up pressure behind somewhere in the oral tract and releasing it all a sudden, without vocal vibration either.

From the definition give above, we could find the major difference between the first type and the others is whether or not vocal cords vibrate.

Commonly, every word includes some vowels or semivowels (excluding few exception), so if we could determine the Fundamental Frequency, we could know if it is a voice.

To extract FF, we select auto-relation algorithm [4], and make some modification to improve it.

$$X(n) = S(n)W(n) \quad (1)$$

$$R_n(k) = \sum_{m=0}^{N-1-k} X(n+m)X(n+m+k) \quad (2)$$

Traditional algorithm

1 Center cutting: find the maximum value of first 1/3 and the last 1/3, then use the smaller (V_0) one as threshold to cut the waveform.

$$Y(n) = C[X(n)] \quad (3)$$

CL = a*V_0, commonly a = 0.6~0.8

2 Observe the figure of auto-relation function. Decide if there is Fundamental Frequency.

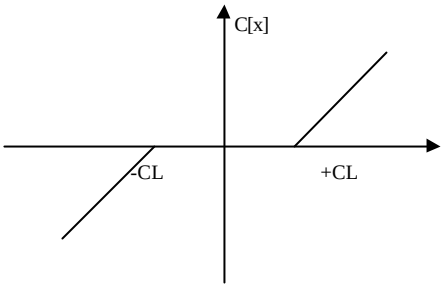

Fig. 1 Function of center-cut, attention: for upper and nether part, use same threshold

The traditional auto-relation algorithm did not consider the asymmetry of the waveform above and below axis. From the Fig. 2, we could find if we use the same threshold to cut the waveform, it will lose periodic information, which is the basement of FF detection. Therefore, we modify the algorithm to use different threshold for upper and nether waveform. From Fig. 3, you could see the essential information is reserved.

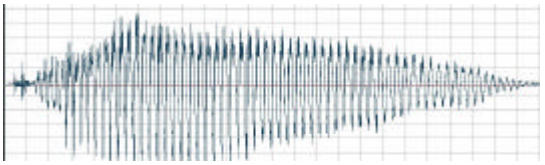

Fig. 2 waveform of /a/

From Fig. 3, we could see clearly the periodic waveform, and the upper and the nether is not symmetrical.

Cut one frame (50ms window, 10ms step, 40ms overlap for 8000Hz sample rate), and use center-cut filter on it.

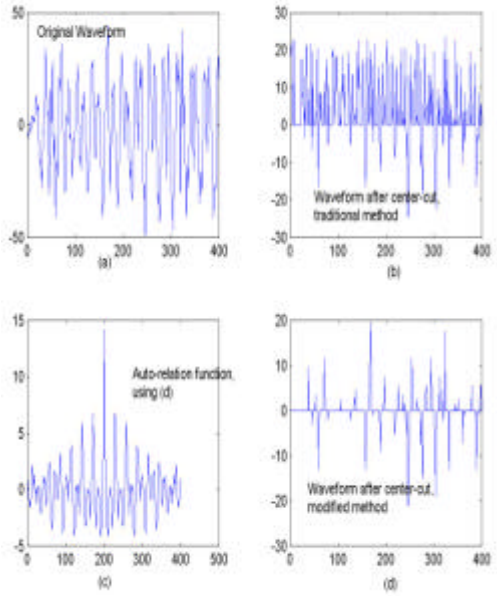

Fig. 3 (a) Original waveform (b) After center-cut with traditional method (c) Auto-relation function using (d)'s result. (d) After center-cut with improved method

Fig. 3 is the result of auto-relation. Fig. 3(d) cuts the little disturbance part and keeps the periodic information. This procedure simplifies the auto-relation function greatly, and makes it easy for one to extract fundamental frequency in next section.

3 Voice Distinction & Its Application

Now, using FF extraction algorithm, we get the robust voice distinction method. Fig. 6, Fig. 7 and Fig. 8 are the waveforms and their FF detection results. The waveform's format is 8K Hz, 8Bits mono sample.

If we observe these figures carefully, we could find that the non-voice's FF results have two main features different from real voice (means vowels or semivowels):

1) The FF values are very irregular, and almost distribute randomly.
2) Even if there are some continuous FF values, they are also below 100Hz or lower.

Thus, we set up the checking measure like these:

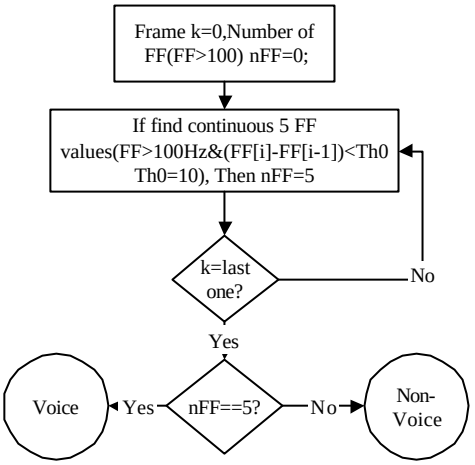

Fig. 4 Flow chart of FF extraction

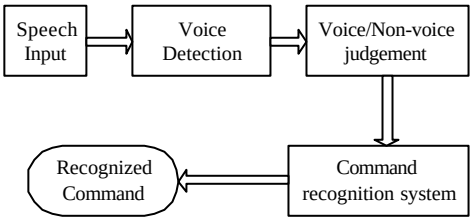

Fig. 5 Diagram for FF extraction's application in command recognition system

Table 1 is the experiment's result. From it, we could see that this algorithm could clearly distinguish voice from other

non-voice sound, like breath, cough, lip or throat sound, and other noises.

If combined with other fast algorithm to compute auto-relation, and voice detection algorithm [5][6], it could be used in speech recognition system successfully.

The Voice Detection part uses frame energy and word duration feature [7][8].

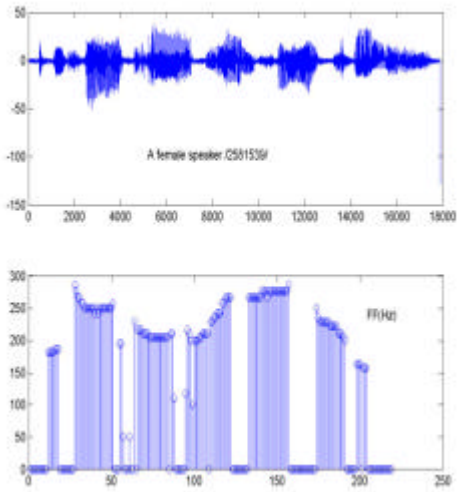

Fig. 6 Real Voice and its FF value

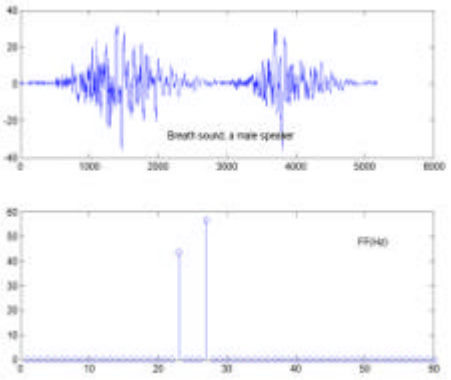

Fig. 7 Breath and its FF value

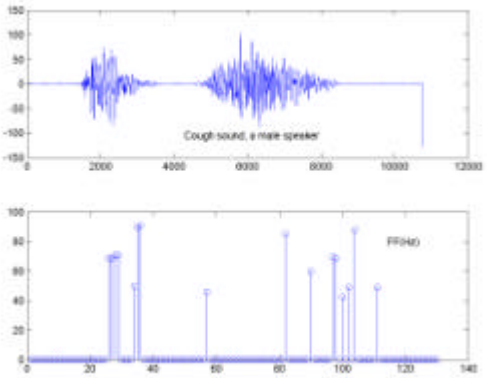

Fig. 8 Cough and its FF value

Table 1 Experiment result for voice/non-voice distinction

	Times	Correctly recognized*
Cough	20	19[#]
Breath	20	20
Lip/Throat	20	20
Other noise	20	20
Real voice	50	50

**Note*: recognized means classifying either as voice or as non-voice.

[#]*Note*: One sound is by a male speaker, who coughed on purpose, very like speaking.

4 Conclusions

This paper focused on a problem in speech recognition, and set up a new method based on fundamental frequency extraction to distinguish real voce from non-voice noise. It will find application in real speech recognition systems. Also, the paper supplied an improved algorithm to extract fundamental frequency. Because the old method did not consider the asymmetry of the waveform, which could take place when the audio input device has different response for positive and negative waveform. In fact, according to

our analysis, this is a common case. We test more than 10 microphones.

With a reliable FF extraction algorithm, we analysis the FF results applied to different sound, real voice or noise (breath, cough, lip or throat sound, nose vibration, etc). Finally, based on the difference between them, the paper put forward a distinction method. Experiments verified our analysis. When used in real system (constructed before to test speaker-independent command recognition), we get promising improvements compared with the baseline system without dong so.

References:

[1] C.H. Lee, " Some techniques for creating robust stochastic models for speech recognition" *J. Acoust, Soc, America*, suppl. 1.vol.82, Fall 1987

[2] L.R Rabiner & S.E Levinson. "Isolated and connected word recognition—theory and selected applications," *IEEE Trans. Commun.* Vol. Com 29. No 5, pp. 621-659, May 1981

[3] G.E. Peterson and H.L. Barney, " Control Methods Used in a Study of the vowels," *Journal of Acoust Soc. Ameri*. 24(2) pp. 175-194, March 1952

[4] M. M. Sondhi, " New Methods of Pitch Extraction," *IEEE Audio and Electroacoustics*, Vol. AU-16, No. 2, pp. 262-266, June 1968

[5] M. H. Savoji, " A Robust Algorithm for Accurate End-pointing of Speech," *Speech Commun*, Vol. 8, pp. 45-60,1989

[6] H. Ney, " An Optimisation algorithm for determining the endpoints of isolated utterances" , *Proc, ICASSP 81*, 1981 pp. 720-723

[7] B. Reaves, "Comments on an improved endpoint detector for isolated word recognition" , *Corresp, IEEE, Acous. Speech, signal processing*, Vol. 39, pp. 526-527, Feb 1991

[8] L. R . Rabiner and M.R. Sambur " Voiced-unvoiced-silence detection using the Itakura distance measure" , *Proc. Conf. Acoust, Speech, Signal Processing.* May 1977, pp. 323-326