Fractional Lower-order Statistics for Yangzhou Dialectal Speech Recognition

Huimin Lu^{a,b,*}, Yujie Li^b, Shiyuan Yang^b, Xuelong Hu^c, Seiichi Serikawa^b

^aState Key Laboratory of Marine Geology, Tongji University, Shanghai 200092, China

^bDept. of Electronic and Electrical Engineering, Kyushu Institute of Technology, Kitakyushu 8048550, Japan

^cSchool of Information Engineering, Yangzhou University, Yangzhou 225127, China

*Corresponding Author: luhuimin@ieee.org

Abstract

With the appearances of information time based on digital techniques, people often concert on many kinds of machines receive, transact and transfer information. As computers are wildy used, it is becoming true that the natural communication between people and machines without using keyboard or mouse will become true. Multimedia era requests speech recognition system to put into practical from laboratory. Isolated word speech recognition system will bring advantages for people in daily life. However, because of the ambient noise, such as Gaussian noise and non-Gaussian noise, the product capability of isolated word speech recognition system is hard to meet the demands. Even the isolated word recognition systems are quite mature, there are lots of problems existed and many fields need to be improved. This paper focuses on the problems of isolated word speech recognition systems as follows: 1) The problem of Pre-treatment in noisy environment. Generally, researchers consider the Gaussian Noise, but usually in our life the non-Gaussian noise are not neglected. Then we can do a good endpoint. Studies showed that a speech system utilizing an isolated word recognizer, more than 50% of error rate was credited to the endpoint detector. 2) The problem of Yangzhou dialectal. To do speech recognition of Yangzhou language by way of phonetic introduction and to establish common-used model is practical for information-exchange between dialects and speech recognition.

Keywords: non-Gaussian noise; Speech recognition; Fractional lower-order statistics (FLOS); Yangzhou dialectal.

1. Introduction

The world has a variety of natural languages, and most of natural languages contain many dialects. The difference between natural languages is very variety. Some dialect not only different in pronunciation, but writing system is also different ⁽¹⁾.

As the developed of the information communication technologies (ICTs), many speech recognition software and synthesis systems were designed in the last decade. Most of these are designed for large population speakers. Recently, the less scale population of language or dialect speech recognition systems and the speech recognition models are rarely. If we take the speech input method to identify Yangzhou dialect and to establish the speech models, can automatic exchange information between dialects. It will make the speech recognition technology for practical usage. Consequently, in this paper, we will concern about this need and propose a novel isolated words recognition system with non-Gaussian noise.

The rest of the paper is organized as follows. In Section 2, we decide the Endpoint detection, where we show that our new method can achieve well word recognition. The robust feature extraction is provided in Section 3, where we use standard Mel Frequency Campestral Coefficient (MFCC). And we use dynamic programming technique for time normalization and alignment in Section 4. Finally, we summarize our study and draw conclusions in Section 5.

2. Endpoint Detection

Fig1: The process of voice signal for pretreatment.

Endpoint detection, which aims at distinguishing speech and non-speech segments from speech signal, is considered as one of the most important steps in preprocessing components in speech recognition. The correct selection of the endpoint can improve accuracy and speed of speech recognition systems. A research showed that even small endpoint errors often result in relatively significant degradation in digit accuracy. In particular, one or more problems usually make accurate endpoint detection difficult. One particular class of problems is the background noise ⁽²⁾.

In order to solve the problem of speech endpoint detection in real world noisy environments, a new robust feature intended for speech endpoint detection is proposed. We considered the Gaussian white noise and non-Gaussian noise. And use the Low pass filter and fractional lower order lp -norm based approaches of adaptive filter for process these noise.

2.1 Short-time Average Energy and Zero-Crossing Rate

Short-time Average Energy and Zero-crossing Rate are used for distance voiceless and voice. The amplitude of voiceless is smaller than that of voice, and the energy of voiceless is obviously less than voice. So we can distance voiceless and voice by energy function.

$$e(i) = \sum_{n=1}^N |x_i(n)|, \quad (1)$$

However, in many cases, the above formula cannot distance them clearly. We can also use Zero-crossing Rate to do this. When the throat produces voiceless, more waveform crosses the zero line. So we can distance voiceless and voice by zero-crossing rate.

$$ZCR(i) = \sum_{n=1}^{N-1} |x_i(n) - x_i(n+1)|, \quad (2)$$

In practice, usually set a threshold, if the symbol of two samples is different, the voice crossing zero line. In this paper we set threshold=0.02.

Fig2: De-noising the non-Gaussian noise by LMP.

2.2 LMP

The classic method of adaptive filter is LMS method. However, it is not suitable for processing non-Gaussian noise signal. And we presented the fractional lower order lp -norm based approaches of adaptive filter for noise signal.

In this method, we set the error function to express the cost function of adaptive system. If meet $0 < p < a$. a is directly related to p . So, the cost function of adaptive system is

$$J = E[|e(n)|^p], \quad (3)$$

We get the gradient L~P-estimate is

$$\hat{\nabla}(n) = \frac{\partial J}{\partial w(n)} = p[e(n)]^{p-1} \text{sgn}(e(n))[-x(n)], \quad (4)$$

$$1 \leq p < \alpha \leq 2$$

Then, LMP can be described as follows:

Initialization:

$$w(0) = 0, R(0) = I; 0 < \mu < \frac{1}{\lambda_{\max}}, \quad (5)$$

for $n=1$ to L -length do:

① *Output of Filter*

$$y(n) = W^T(n)X_N(n), W = [w_0, \dots, w_{N-1}]^T, \quad (6)$$

$$X_N(n) = [x(n), \dots, x(n-N+1)]^T$$

② *Error Estimation*

$$e(n) = x(n) - y(n), \quad (7)$$

③ *Adaptive Formula*

$$W(n+1) = W(n) - \mu p x_n(n) |e(n)|^{p-1} \text{sign}(e(n)), \quad (8)$$

In this paper, we set $\mu=0.00001$, $p=1.1$. For the figure we can see that LMP is better for noise elimination.

3. Feature Extraction

3.1 Linear Prediction Coding

As applied to speech, the theory of linear prediction coding has been well understood for many years. The basic idea behind the LPC model ⁽³⁾ is that a given speech sample at time n , $s(n)$, can be approximated as a linear combination of the past p speech samples, such that,

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \dots + a_p s(n-p), \quad (9)$$

where the coefficients $a_1, a_2, \dots, a_p$ are assumed as constant over the speech analysis frame. We convert Eq.(9) to an equality by including an excitation term, $G^*u(n)$, giving:

$$s(n) = \sum_{i=1}^p a_i s(n-i) + Gu(n), \quad (10)$$

where $u(n)$ is a normalized excitation and G is the gain of the excitation. By expressing Eq.(10) in the z -domain we get the relation

$$S(z) = \sum_{i=1}^p a_i z^{-i} S(z) + GU(z), \quad (11)$$

Leading to the transfer function

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^p a_i z^{-i}} = \frac{1}{A(z)}, \quad (12)$$

and, its frequency response $H(e^{j\omega})$ is called LPC spectrum.

In the areas of large signal energy (at spectrum peak), LPC spectrum is similar to signal spectrum; and at spectrum bottom, the distance between them is big.

Above all of these, we considered that if the value of p is more large, the $|H(e^{j\omega})|$ is nearly to $|S(e^{j\omega})|$, can use the pole-model instead of zero-model. However, these can increase the calculation and storage. Here, we use $p=12$.

3.2 Mel Frequency Campestral Coefficient⁽⁴⁾

According to the characteristics of the human ear, we can calculate that

$$c_n = \sum_{k=1}^K (\log S_k) \cos[n(k-0.5)\frac{\pi}{K}], \quad (13)$$

where S_k is the output power of the k -th filter of the filter bank, and n is from 1 to 12. We can also calculate the logged energy of each frame as one of the coefficients,

Fig3: The figure of Mel-Frequency Cestrum, $n=24$.

$$E = \log \sum_{n=1}^{160} s_n^2, \quad (14)$$

We can calculate the Mel-Frequency cepstrum from Eq.(14).

4. Dynamic Time Wrapping

Dynamic Time Warping ⁽⁵⁾ is an extensively studied and widely used tool in operations research for solving sequential decision problems. To illustrate its applicability in speech recognition, and particularly for the problem of time alignment and normalization, the algorithm can be summarized as follows,

1. Initialization

$$\psi_1(i, n) = \zeta(i, n)$$

$$\varepsilon_1(n) = i$$

$$\text{for } n=1, 2, \dots, N, \quad (15)$$

2. Recursion

$$\psi_{m+1}(i, n) = \min_{1 \leq \ell \leq N} [\psi_m(i, \ell) + \zeta(\ell, n)]$$

$$\varepsilon_{m+1}(n) = \arg \min_{1 \leq \ell \leq N} [\psi_m(i, \ell) + \zeta(\ell, n)]$$

$$\text{for } n=1, 2, \dots, N \text{ and } m=1, 2, \dots, M-2, \quad (16)$$

3. Termination

$$\psi_M(i, j) = \min_{1 \leq \ell \leq N} [\psi_{M-1}(i, \ell) + \zeta(\ell, j)]$$

$$\varepsilon_M(j) = \arg \min_{1 \leq \ell \leq N} [\psi_{M-1}(i, \ell) + \zeta(\ell, j)], \quad (17)$$

4. Path Backtracking

$$\text{optimal path} = (i, i_1, i_2, \dots, i_{M-1}, j), \quad (18)$$

where

$$i_m = \varepsilon_{m+1}(i_{m+1}), \quad m=M-1, M-2, \dots, 1, \quad (19)$$

$$\text{with } i_M = j, \quad (20)$$

5. Summarize and Conclusions

We can conclude from TABLE 1 that LMP algorithm has higher identification accuracy in non-Gaussian noise and Gaussian noise environment ⁽⁶⁾. We found in the experiment that as the numbers of words become more, the identification accuracy becomes lower. We also found that it is difficult to distance between the number “1” and “7”. In Yangzhou dialectal, the pronunciation of these two words is similarly. And we will adopt a new method to solve this problem.

Fig4: The endpoint of Yangzhou dialectal number “1”.

Table 1. Experimental Results of Yangzhou Dialectal Recognition

<i>Experimental Results</i>	<i>10 words</i>	<i>20 words</i>	<i>50 words</i>
Before LMP	98.1%	92.5%	85.3%
After LMP	100%	98%	96.2%

(a) Contain non-Gaussian noise.

(b) After LMP filter.

Fig 5: Number “1” of Yangzhou dialectal after LMP filter.

Acknowledgment

This work is supported in part by the Japan Government Research Fellowship for Young Scientists by Japan Society for the Promotion of Science (13J10713), Grant in Aid for State Key Laboratory of Marine Geology (MGK1407), Tongji University, China ⁽⁷⁾.

References

- (1) L.Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall PTR, 1993.
- (2) C.L. Nikias, M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. New York: John Wiley and Sons, 1995.
- (3) W. Han, C.F. Chan, “An Efficient MFCC Extraction Method in Speech Recognition”, Proc. of IEEE International Conference on Circuits and Systems, pp.1-4, 2006.
- (4) D. Chen, W.G. Gong, B. Li, “High frequency weighted MFCC extraction for noise robust speaker verification”, *Chinese Journal of Scientific Instrument*, vol.29, no.3, pp.668-672, 2008.
- (5) S. Okawa, S. Katsuhiko, “Inquiry into a Common Acoustic Model to Realize Mongolian Dialectal Speech Recognition”, *Journal of Minzu University of China: Social Science edition*, vol.28, no.4, pp.114-121, 2001.
- (6) T.S. Qiu, X. Zhang, *Statistical Signal Processing—non-Gaussian Signal Processing and Application*, Publishing House of Electronics Industry, 2004.
- (7) S. Serikawa, H. Lu, “Underwater Image Dehazing Using Joint Trilateral Filter”, *Computers and Electrical Engineering*, 40(1), pp.41-50, 2014.