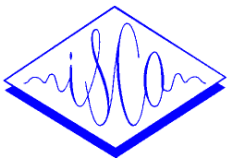

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

AN ALGORITHM FOR VOICED / UNVOICED DECISION AND PITCH ESTIMATION IN SPEECH FEATURE EXTRACTION

WANG Dong
Department of Electronical Engineering,
Tsinghua University, Beijing
wdong01@mails.tsinghua.edu.cn

CHEN Yi-Ning
Department of Electronical Engineering,
Tsinghua University, Beijing
chenyining99@mails.tsinghua.edu.cn

LIU Jia
Department of Electronical Engineering,
Tsinghua University, Beijing
liuj@tsinghua.edu.cn

ABSTRACT

An algorithm which combines voice / unvoiced decision and pitch estimation is proposed in an enhanced process of MFCC feature extraction. The residual energy of LPC analysis and normalized autocorrelation are calculated and the static and dynamic thresholds are set for the voiced, unvoiced and transitional decision. Thus speech is divided into three classes that are voiced, unvoiced and transitional. Then the pitch is estimated by a dynamic programming (DP) algorithm. In the following harmonic peak picking process, the result is refined by the additional spectral information. The algorithm is empowered by the finite state machine (FSM) embedded in U/V decision which can convert the static thresholds to dynamical variable thresholds and represent the actual speech more exactly. Experiments also show that performance gains of word recognition rate from 71.49% to 74.42% in the National 863 standard Mandarin speech Corpus.

1. INTRODUCTION

The mel-frequency cepstral coefficients (MFCC) have been the most widely used front-end feature for speech recognition in recent years. It is popular for its better representation of consonant and performance under noisy conditions than LPCC. But some drawbacks arise here. It is well known that the envelope of a short-term spectrum, not the gross spectrum, represents the linguistic information in speech [1]. Experiments also showed that while the upper envelope of power spectrum sampled at pitch harmonics remains nearly unchanged, which is due to the short range stationary property of the speech samples, the lower part envelope changed considerably [2]. So mel-frequency filtered coefficients express the transitional and the vowel part of speech with lower accuracy because of its variance from one frame to the next. The constant mel-frequency filters also can not discern the different formant width between the male and female speakers which are highly speaker dependent. This can be compensated by various method of vocal tract length normalization (VTLN), but the question remains unsolved because these methods are all post-processing methods, and the information is already lost.

Pitch is one of the most important features of speech. Here pitch refers to the fundamental frequency contour of successive frames, not the psychological one we feel. Roughly speaking, the so called prosody modeling for automatic speech recognition and understanding are mainly refer to pitch contour estimation and application. The pitch contour parameter has been widely used in the fields of speech transmission, compression and recognition, and it has been used in scene analysis and human emotion detection [3].

There are five tones in the Mandarin language. So the language is tonal and pitch variation play a significant role in the meaning convey of the language [4]. The pitch is very important both for Mandarin character recognition and other relevant recognition tasks. And that is why pitch estimation has attracted so much attention in automatic Mandarin speech recognition.

The perceptual harmonic cepstral coefficients (PHCC) add structural pitch information to MFCC. The speech is divided into voiced, transitional and unvoiced classes and weights of different factors have been given to the frames [2]. This paper summarizes our work with PHCC and gives a new accurate voiced / unvoiced decision and pitch estimation algorithm which can upgrade PHCC system performance to MFCC system level.

The paper starts with parameter estimation for the embedded FSM, pitch extraction and spectral weighting in Section 2, together with some descriptive figures. Experimental results are given in Section 3, and the paper concludes in Section 4 with an outlook on further directions towards the more accurate ASR system of Section 5.

2. PARAMETER ESTIMATION AND SPECTRAL WEIGHTING

The speech signal is generated via modulation of the vocal tract transfer function by an excitation signal. The distinction between voiced and unvoiced sounds is that the excitation is quasi-periodic for voiced sounds, and white noise for unvoiced sounds. After the modulation, the spectral distinction sometimes becomes subtle. And the transitional segments often occur between words, or even voiced and unvoiced speech parts. The normal classification of vowel and consonant roughly corresponds to voiced / unvoiced classification, but we will use the latter because of its terminology accuracy.

The algorithm below uses autocorrelation and residual energy of linear prediction as the voiced / unvoiced decision parameters and embeds a finite state machine in it, applies DP search to pitch estimation process, and refines the result in the third stage of spectral weighting.

2.1 FSM embedded in voiced / unvoiced decision

The voiced phoneme is more steady both in time and frequency domain and can be easily modeled as constant state. The unvoiced phoneme is relatively difficult to handle because of its irregularity. But the small correlation coefficient and low energy provides some usable parameters for it.

Transitional state is introduced to describe the transition between voiced and unvoiced sounds. Because of the difficulty of deciding which state it belongs to for the frames near the boundary between two unvoiced / voiced phonemes, the decisions are postponed to a later time using Viterbi decoding algorithm which is more accurate for this task. The phase shift from unvoiced to voiced usually expressed in a sharp uproar of energy, especially for the Mandarin language, and we need no elaboration here. But after the voiced state, the FSM will either fall back to unvoiced or continue to linger in another voiced state. It is difficult to decide the accurate boundary between voiced and unvoiced states even for ourselves, especially when half vowels and nasals occurs. So we find it is useful to introduce this transitional state in the feature extracting process.

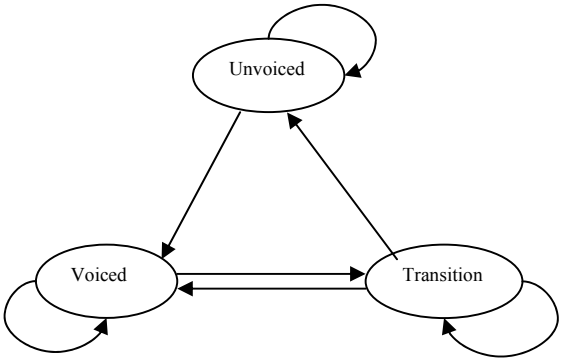

Figure 1. FSM embedded in the voiced / unvoiced decision

The three states of the embedded FSM and the state transition arcs are shown in the figure above. Because silence is well modeled in the acoustic HMM, it is omitted here.

The frame-length normalized residual energy of the frame data after passing a 6 order LP filter and the normalized autocorrelation function of the frame are calculated as the two parameters for deciding the transition action. Before the calculation, a band pass filter of 60~900 Hz filter band is applied to the windowed frame data to eliminate the illusionary effect of the first formant on robust pitch estimation.

The energy and autocorrelation parameters introduced above are calculated when each frame comes. But the static energy threshold can not meet the dynamic environment changes. So the dynamical energy threshold, $Energy\_D(k)$, is introduced for state transition and is defined as

$$Energy\_D(k) = \frac{\sum_{i=1}^{5} E(k-i+1)/i}{\sum_{i=1}^{5} 1/i},$$

where $Energy(k)$ stands for the residue energy of frame k. When every frame comes, the *Energy_D* is updated according to different states. But the autocorrelation threshold is static across time.

A duration upper bound of *Duration_Transition* for the transitional state and a lower bound of *Duration_Voice* for voiced state are also introduced to improve the robustness of this algorithm.

The state transition algorithm is listed below in the descriptive way, in which the "Energy" refers to the residue energy and the "AutoCorr" the autocorrelation. For example, in this name convention, the *Energy_UnvoiceD* means the dynamic energy threshold of Unvoice state.

State transition algorithm:

Select Current State

1. ***Unvoice:***

If *(Energy > Energy_UnvoiceD or Energy > Energy_Voice) and (AutoCorr > AutoCorr_Voice)*
State -> Voice
Else
State -> Unvoice, update Energy_UnvoiceD.

2. ***Transition:***

If *Energy > Energy_VoiceD and AutoCorr > AutoCorr_Voice*
State -> Voice
ElseIf *Energy < Energy_TransitionD or Duration > Duration_Transition*
State -> Unvoice
Else
State –> Transition , update Energy_TransitionD.

3. ***Voice:***

If *Duration < Duration_Voice or Energy < Energy_VoiceD or Energy < Energy_Unvoice*
State –> Voice, update Energy_VoiceD;

Else
State -> Transition.

The main gain of this method is the dynamic threshold which can adapt to the environment change while static ones can not.

2.2 Pitch Estimation using DP search

In resent years, various kinds of new methods of pitch contour extraction have been put forward. The novel one using nonlinear state-space [5], the more complex one of so called instantaneous frequency amplitude spectrum [6], the improved AMDF of circular AMDF [7], and the smoothed one of spline-based [8] are all put forward in ICASSP 2002. But the traditional method of DP plus autocorrelation is proved to be very useful. And pitch contour modification is not used because we only use it in separate frames.

Spectral envelope requires robust pitch estimation. The main obstacles are so called pitch multiples and the formant interference. The latter effect is controlled in stage 2.1 by a band pass filter. But the former calls for the following DP search.

The five maximal spectral peaks are selected as the fundamental frequency candidates, and each path is extended when every new frame arrives. A cost function gives high score to punish paths which join far apart pitch candidates together. Thus by minimizing the cost function of transitions among the consecutive frame of fundamental frequency candidates we select the proper pitch of the speech accurately and circumvents the usual pitfall of pitch multiples. A time delay of two frames is introduced here as a natural result of DP search.

2.3 Refining in the harmonic weighting process

The voiced sound has fine structure in its spectrum, and this is not fully represented in the parameters of stage A. On the harmonic weighting process, the harmonic peaks are found by searching the spectrum with step corresponding to fundamental frequency. The local peak-picking algorithm corrects minor pitch estimation errors or non-integer pitch effects. It looks for local maxima in a search interval that excludes neighboring harmonics and each time a harmonics peak comes out, the pitch is refined by the additional information. Then different weights are given to these pitch multiples [2].

Figure 2 shows the result of Unvoiced / Voiced decision with one sentence in the 863 database. The horizontal line is the U/V decision, and the red line is the energy plot, the black correlation. Seen from the figure, the algorithm can keep up the dynamic property of the speech utterance. Figure 3 shows the pitch estimation result of the sentence, and the black line is the correlation.

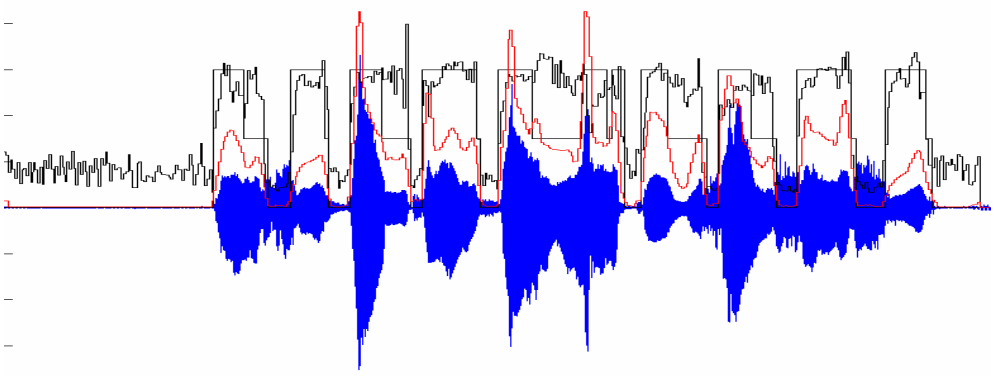

Figure 2. Unvoiced / Voiced decision

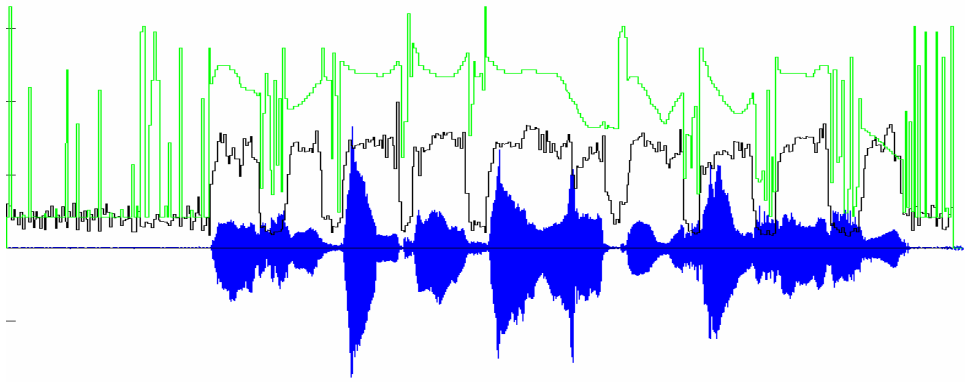

Figure 3. Pitch Estimation

3. EXPERIMENTAL RESULTS

To test the performance of the enhanced U/V decision, comparisons of MFCC and PHCC feature are made. Experiments are carried out on a database of National 863 standard Mandarin speech Corpus [9]. Training database includes 20 hours data spoken by 34 female speakers. Testing database contains 3.6 hours data spoken by 6 female speakers. All the speech is sampled by 16KHz and quantified into 16 bits.

In our experiment, 31-dimension speech features were used, including 14 cepstral (MFCC or PHCC) parameters, log energy, and their dynamics (time derivatives), and the second order dynamics of log energy. We used an analysis frame of width 30ms and step of 10ms, and a Hamming window. Gaussian mixture HMM models are used. The experiment results of PHCC and MFCC are summarized in Table 1. The overall recognition rate is the average of 6 speakers.

Acoustic Models	Character Recognition Rate
MFCC	74.95%
PHCC_initial	71.49%
PHCC_RE	71.67%
PHCC_RE+DP	73.44%
PHCC_RE+DP+FSM	74.42%

Table 1. Test-set error rate based on PHCC and MFCC and improvements on PHCC

Table 1 shows that error rate of the initial PHCC feature is very poor compared to that of MFCC feature. And the pitch refining process (RE) has little effect on the system performance. But the DP search significantly promotes the performance by nearly 2%. After the FSM is included and parameters adjusted, another 1% update is acquired and the performance has been increased to MFCC level. The improvement of PHCC_RE+DP+FSM as the final result can be explained in the U/V decision and pitch extraction figures, where the dynamic range of the residue energy is matched very well by the FSM, and the accuracy of pitch estimation is upgraded by the DP search process.

4. CONCLUSION

The embedded FSM method is put forward mainly for solving the dynamic threshold problem of U/V decision. Two different parameters of residue energy and autocorrelation are used and the simple decision rules are set. After the two improvements of DP search and pitch refining process, satisfactory experiment result on LVCSPR database is given compared to MFCC feature.

5. FUTURE DIRECTIONS

Our future work will focus on the combination of HMM and the front-end U/V decision and pitch estimation, and integrate the pitch contour information in the recognition process to further improve the performance of PHCC features.

6. REFERENCES

[1] H.Hermansky, “Auditory Modeling in Automatic Recognition of Speech”, in Proceedings of the First European Conference on Signal Analysis and Prediction, pp. 17-21, Prague, Czech Republic, 1997

[2] Liang Gu and Kenneth Rose, “Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition”, Proc. International Conference of Spoken Language Processing, pp. 309-312, Beijing, China, 2000

[3] Feng Yu, Eric Chang, Ying-Qing Xu and Heung-Yeung Shum, “Emotion Detection from Speech to Enrich Multimedia Content”, IEEE Pacific Rim Conference on Multimedia 2001, pp550-557

[4] 林焘, 王理嘉, 语音学教程, 北京大学出版社, 1992, 北京.

[5] Dmitry E. Terez, Robust pitch determination using nonlinear state-space embedding, ICASSP 2002, Orlando.

[6] Tomohiro Tanaka, Takao Kobayashi, et al, Fundamental frequency estimation based on instantaneous frequency amplitude spectrum, ICASSP 2002, Orlando.

[7] Wenyao Zhang, Gang Xu, Yuguo Wan, Pitch estimation based on circular AMDF, ICASSP 2002, Orlando.

[8] Andrei Jefremov and W. Bastiaaan Kleijn, Spline-based continuous-time pitch estimation, ICASSP 2002, Orlando.

[9] R. H. Wang, “National performance assessment of speech recognition system of Chinese”, In Proc. Oriental COCOSDA workshop ’99 Taipei 1999, pp41-44