Research paper

A Study of the Application of an Average Energy Entropy Method for the Endpoint Extraction of Frog Croak Syllables

Shan-Chih Hsieh,[1)] Wen-Ping Chen,[1,3)] Wen-Chih Lin,[2)]
Fu-Shan Chou,[2)] Jiunn-Ru Lai[1)]

【Summary】

Energy-based endpoint detection is commonly used in time domain analyses of speech segments of extracted signals to reduce the amount of computation required. However, this approach may extract incorrect speech segments due to interference by noise, which can significantly impair its recognition ability when analyzing sound files recorded in the wild. In contrast, entropy-based endpoint detection performs better in terms of noise suppression. Unfortunately, background noise that has a non-stationary frequency distribution causes drastic fluctuations in entropy values of silent segments, and weakens endpoint detection. This paper proposes using average energy entropy (AEE) endpoint detection to address these issues, and compares the AEE method with 3 other endpoint detection methods-energy-based, zero-crossing rate, and entropy-based detection methods. In experiments on frog voice-print recognition, 18 types of frog croaks recorded from the wild were analyzed, and the results revealed that the AEE method had the optimal endpoint extraction capability; and when used in concert with the linear predicative cepstral coefficients, Mel-frequency cepstrum coefficients with dynamic time warping algorithm, the AEE capability for recognition was optimized.

Key words: energy-based endpoint detection, entropy-based endpoint detection, voice-print recognition.

Hsieh SC, Chen WP, Lin WC, Chou FS, Lai JR. 2012. A study of the application of an average energy entropy method for the endpoint extraction of frog croak syllables. Taiwan J For Sci 27(2):177-89.

[1)] Department of Electrical Engineering, National Kaohsiung Univ. of Applied Sciences, 415 Jiangong Rd., Sanmin, Kaohsiung 80778, Taiwan. 國立高雄應用科技大學電機工程系，80778高雄市三民區建工路415號。

[2)] Liouguei Research Center, Taiwan Forestry Research Institute, 198 Chunghsing Village, Liouguei, Kaohsiung 84443, Taiwan. 林業試驗所六龜研究中心，84443高雄市六龜區中興村198號。

[3)] Corresponding author, e-mail:pen@mail.ee.kuas.edu.tw 通訊作者。

Received October 2011, Accepted March 2012. 2011年10月送審 2012年3月通過。

研究報告

平均能量熵值法應用於蛙鳴音節端點萃取之研究

謝勝治[1)] 陳文平[1,3)] 林文智[2)] 周富三[2)] 賴俊如[1)]

摘 要

能量(energy)端點偵測法經常被用於擷取信號的語音片段之時域(time domain)分析，以節省計算量，但此法容易受到雜訊影響而擷取不正確的語音片段，這對於分析野外所錄製之音檔而言，辨識能力將大受影響；而熵值(entropy)端點偵測法雖有較佳的抗噪能力，但背景雜訊不穩定的頻譜分佈，會導致非有聲段部份的熵值起伏劇烈而影響端點的偵測。因此本文提出平均能量熵值端點偵測法(average energy entropy (AEE) endpoint detection)來改善上述問題，並與能量、越零率、熵值等三種端點偵測法做比較，而在蛙鳴聲紋辨識實驗上，經實驗18種野外蛙類音檔分析後發現，平均能量熵值端點偵測法有最佳的端點萃取能力，而搭配線性預估倒頻譜係數與動態時軸校正演算法則有最佳的辨識能力。

關鍵詞：能量端點偵測法、熵值端點偵測法、聲紋辨識。

謝勝治、陳文平、林文智、周富三、賴俊如。2012。平均能量熵值法應用於蛙鳴音節端點萃取之研究。台灣林業科學27(2):177-89。

INTRODUCTION

Due to the relentless destruction that mankind has brought against nature, successive natural catastrophes have severely affected wildlife habitats. The effects have been particularly prominent in recent years, as extreme climate change is being witnessed all over the world, imposing serious impacts on the livelihoods of human beings, animals, and plants. In response, ecologists from across the world have set out to investigate the ecology of wildlife habitats to assess the degree of impacts that ecosystems have experienced. Traditionally, ecological survey fieldwork is manually carried out, whereby information on plants and animals is obtained through laborious trips into the wild by researchers. This not only consumes time and money, but also puts the safety, even the lives, of researchers in jeopardy.

Fortunately, thanks to technological advancements in eco-informatics, along with related applications in the field, much of the manpower and resources traditionally used for ecological investigations can be replaced by integrating sensor and network technologies. Deploying sensors in the wild and transmitting the data back through a wireless network can substantially reduce the number of risky trips into the wilderness by ecological researchers. This remarkable breakthrough in traditional ecological investigation techniques was achieved through information communication technology. In addition, the digitization of raw data recorded with sensors facilitates access and sharing. In recent years, the recognition of frog croaks has become extremely crucial in monitoring and controlling biological and ecological environments.

Advances in technology have enabled the gradual prevalence of user-friendly human-machine operations, including the development of speech recognition systems through

which humans can directly speak to computers. Apart from speech recognition systems that take humans as their main subject for study (Ishimitsu et al. 2007, Janakiraman et al. 2010), there are also studies on biological voice-print recognition, in which the main research subjects are animals (Taylor et al. 1996, Harma 2003, Lee et al. 2006, Fagerlund 2007, Zhao and O'Shaughnessy 2008, Huang et al. 2009, Jančovič 2011). Through recognizing the clamor of animals using a biological voice-print recognition system, various species of animals can be discerned.

However, background noise often seriously interferes with recordings from the wild. Taking frogs as an example, since Swinhoe's brown frogs like to live near creeks, recordings may contain serious interference from the sound of running water; while recordings of tree frogs inevitably contain the sound of the wind or of cicadas in the background, as they are also active in the woods. Therefore, conventional processing of voice recordings from the wild with energy-based endpoint detection techniques is unable to correctly divide the sound into syllables because of the background noise; hence the recognition and efficiency of the voice-print recognition systems are affected.

Noise can be classified into convolutional noise and additive noise. The former refers to noise generated when it passes through the transmission channel of the sensors; this is usually referred to as a channel effect. Generally, convolutional noise is related to the quality of the sensors. Additive noise refers to the linear summation of the target sound source and the background noise in the audio recording; this type of background noise negatively impacts the process of syllable extraction of frog croaks.

Commonly used endpoint detection methods can be classified into 2 categories: time-domain and frequency-domain endpoint detection. Detection through the time-domain endpoint generally obtains energy values of sound recordings by calculating the absolute value or square of the amplitude (Lamel et al. 1981), and the speech and silent segments are distinguished by fluctuations in the energy magnitude. The energy-based method is usually augmented using the zero crossing rate (ZCR) (Tian et al. 2002) to compensate for regions that cannot be discerned by the energy-based method alone. However, the energy-based method is prone to faulty results in noisy environments. As to endpoint detection in the frequency domain, although the amount of computations required is larger than that of the time-domain approach, frequency-domain endpoint detection nevertheless has better noise suppression.

Tu and Hung (2007) proposed 4 research methods for frequency-domain endpoint detection: low-frequency spectral magnitude, full-band spectral magnitude, cumulative quantized spectrum, and high-pass log energy. Among the 4 methods, the low-frequency spectral magnitude method takes the strength of low-frequency spectra of each sound frame as the endpoint for discernment, which has been empirically proven to be the most effective method, with a rather high recognition rate. However, because recordings of frog croaks recorded in the wild often contain low-frequency noises from the natural environment, the error rate in recognizing the voice-print rises significantly as a result.

Zhao and O'Shaughnessy (2008) proposed a hybrid segmentation method for speech segmentation, suggesting that there is a block for pauses between syllables; therefore, segments containing speech elements can be distinguished by identifying the blocks of pauses. It begins by discerning the short-time energy of each sound frame, and executing

Hamming short-time sliding windows twice to obtain a more-obvious wave trough. The hybrid method then carries out wave-crest and spectral-variation analyses. Finally, the consonant and vowel blocks are rectified through the ZCR. Shen et al. (1998) proposed an entropy-based endpoint detection method, which first obtains the probability mass function (PMF), then a weighted entropy value through calculations with a weighted value. The purpose is to further adapt to the frequency components. However, errors of discernment are easily caused by the blurry division between noise and speech segments.

To summarize, this paper proposes an average energy entropy (AEE) value method to enhance the endpoint detection for dividing sound syllables. As an application of the probability function, the proposed method possesses an important feature for a series of events: the higher probability of the occurrence of a particular event, the lower its entropy value. In other words, entropy measures the degree of uncertainty of a particular event; a high entropy value indicates a high degree of uncertainty. Likewise, a lower entropy value suggests higher stability. This characteristic can be applied to endpoint detection, where speech segments have lower stability than silent ones.

When entropy values are used for sound analyses, we can reasonably judge that portions with higher negative entropy values are segments containing speech signals. However, if the audio file contains too much noise, it becomes very difficult to distinguish the speech portions from noise ones using an entropy value analysis. Therefore, this paper improves on the entropy algorithm by enhancing the ability to suppress noise, thus significantly strengthening its ability to distinguish speech sounds from noise. It is anticipated that the AEE algorithm proposed in this paper can be used to suppress interference caused by background noise, thus improving the accuracy of endpoint detection and the recognition rate of frog croaks.

MATERIALS AND METHODS

Pattern recognition of frog croaks is much simpler than that for human speech because they only contain the fundamental frequency in the frequency distribution. Hence, endpoint detection can suitably be used to extract frog croak syllables. The AEE algorithm proposed in this paper is an improved entropy-based method, and the workflow of its application on endpoint detection of frog croak syllables is shown in Fig. 1 This workflow consists of a series of procedures, including signal preprocessing, signal conversion, addition of the average energy, computation of the PMF, calculation of the entropy value, and detection of the endpoint. Four different methods of endpoint detection are compared in terms of their recognition rate using a voice-print recognition system, which adopts linear predicative cepstral coefficients (LPCCs) (Rabiner and Juang 1993) and Mel-frequency cepstrum coefficients (MFCCs) as feature parameters. The process is elaborated below.

Signal preprocessing

To maintain consistency of the sampling rate, each sound signal file is re-sampled to 22.05 kHz in mono format, with the amplitude normalized to within the [1, -1] range. Due to attenuation of the high-frequency portion of sound signals when the frequency increases in the process of sound emission to sound reception by the recording device, the sound must be processed with a pre-emphasis, to compensate for the attenuation of highly-frequency energy.

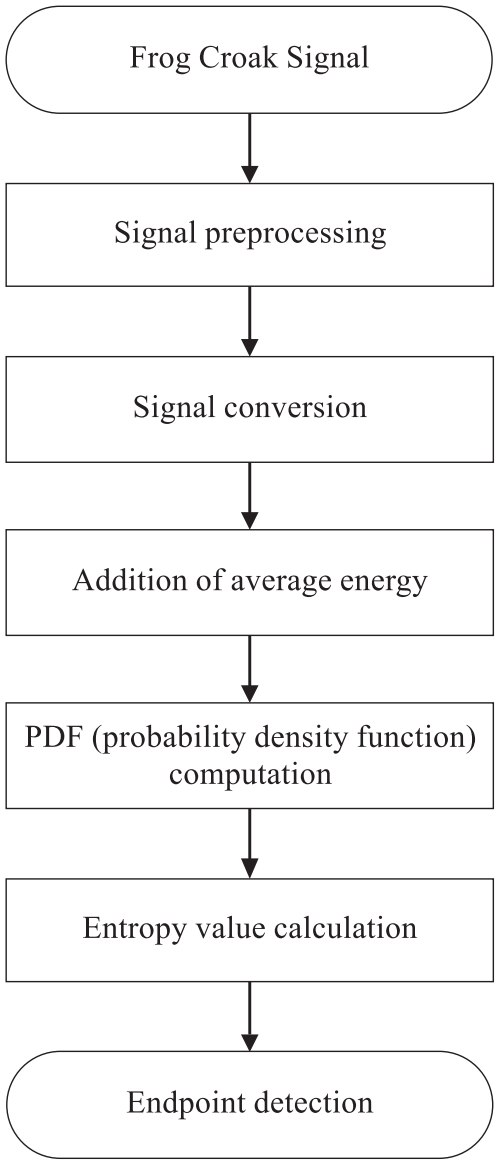


Fig. 1. Workflow of average energy entropy endpoint detection.

This can be viewed as passing the signal through a group of high-pass filters, with its value being denoted by formula (1):

$$\hat{s}(n) = s(n) - \alpha s(n-1); \quad (1)$$

where $s(n)$ represents the original signal, and α is a constant between 0.9 and 1. Because of abrupt changes between signal points, the signals need to be divided into frames for stability considerations. The change between neighboring frames should not be too sheer, such that some parts of adjacent frames overlap to maintain features of the signal over the short time.

To remove edge effects at the endpoints of a frame, a window is usually added to each of the frames. The most commonly used one is the Hamming window; and the signal, $w(n)$, after the addition of the window is shown in formula (2):

$$w(n) = \hat{s}(n) \times \left[0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\right]; \quad (2)$$

where $\hat{s}(n)$ refers to the signal of each frame, N is the length of the frame, and n is between 0 and N-1.

Signal conversion

The preprocessed sound signals are still time-domain signals, and each sound frame needs to go through a fast Fourier transform to obtain the spectral energy. The corresponding frequencies, $X(k)$, are shown in formula (3):

$$X(k) = \sum_{n=0}^{N-1} w(n) e^{-j2\pi n \frac{k}{N}}, \ 0 \leqq k \leqq K-1; \quad (3)$$

where k is the frequency component, and K is the Fourier transform number.

Basic spectral entropy value

According to Shen et al. (1998), the PMF of a spectrum must be generated before one can obtain its entropy value, so that the normalized frequency components of all frequency components within the sound frames can be obtained, as shown in formula (4):

$$p_k = \frac{X(k)}{\sum_{k=0}^{\frac{K}{2}-1} X(k)}, \ 0 \leqq k \leqq \frac{k}{2} - 1; \quad (4)$$

where K is the Fourier transform number, $X(k)$ is the spectral energy of spectrum k, and p_k is the corresponding PMF.

To improve the PMF's discernment of speech and non-speech segments, 2 constraints should be put in place:

(a) because the majority of speech frequencies fall in the range of 250~6000 Hz, the constraint is set to

$X(k) = 0$, *if* $f_k < 250Hz$ *or* $f_k > 6000Hz$, and (5)

(b) the upper and lower bounds for the PMF should be restricted as shown by the following formula:

$$p_k = 0, \text{ if } p_k < \sigma_2 \text{ or } p_k > \sigma_1; \quad (6)$$

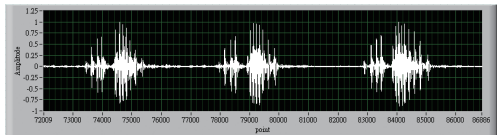
where the lower bound, δ_2, is used to remove white noise, while the upper bound is to remove noise in specific frequency bands. After the above processes, the negative spectral entropy value representing each sound frame can be calculated, with values defined by formula (5):

$$H = \sum_{k=0}^{\frac{K}{2}-1} p_k \log p_k. \quad (7)$$

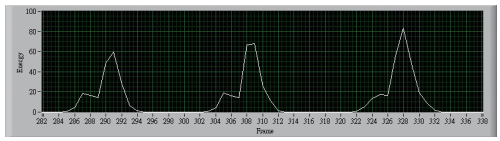
AEE method

The entropy value, representing randomness, is a type of parameter represented by a frequency domain. According to accounts of previous researchers (Fagerlund 2007), the value of background noise is non-stationary when recorded in silent segments; therefore, it becomes difficult to set bounds of endpoints for syllable detection. Take signal processing of Kuhl's creek frog croaks in Fig. 2 as an example. For the negative entropy value in Fig. 2C, due to abrupt changes in the values, the system is unable to accurately determine whether the end point of the syllable is point A or B during endpoint detection; discernment mistakes thus occur in endpoint detection.

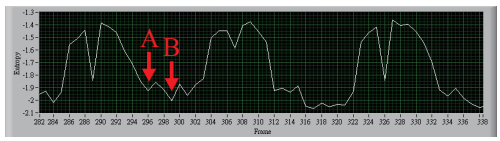
However, when the magnitude of the background noise is lower than that of the speech segments, the entropy values of the speech segments become stable. This paper



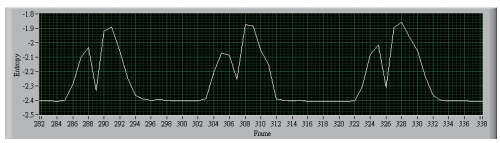
A. Time domain signals



B. Energy values



C. Negative entropy value



D. Improved negative entropy value

Fig. 2. Croaking signals of the Kuhl's creek frog.

proposes an AEE method, which can be applied to syllable endpoint extraction for audio data collected in wild environments containing interference from background noise. This method lowers the entropy value of the background noise of the silent segments by adding signals of certain energy to the original audio data, in an effort to overcome the shortcomings of conventional entropy-based algorithms. Details of the steps are elaborated as follows.

(a) The average energy of the input signal is calculated, with its value being defined by formula (8):

$$u = \frac{1}{M} \sum_{m=0}^{M-1} |A(m)|; \quad (8)$$

where u is the average energy point of the entire signal segment, $A(m)$ is the magnitude of point m, and M stands for the total points of the signals.

(b) Fourier transform is executed on every sound frame, and signals are converted into time-frequency signals.

(c) The average energy value is added several times to the frequency components of each frame to calculate its PMF:

$$p_k' = \frac{(X(k) + cu)}{\sum_{k=0}^{\frac{K}{2}-1}(X(k) + cu)}, 0 \leqq k \leqq \frac{k}{2} - 1; \quad (9)$$

where K is the number of Fourier-transform points, $X(k)$ is the spectral energy of frequency k, p_k' is the corresponding probability mass, and c is a parameter value, the value of which is decided according to the signal-to-noise rate (SNR) of signals in the background environment.

(d) The negative entropy value for individual frames is calculated, with the value denoted by formula (10):

$$H' = \sum_{k=0}^{\frac{K}{2}-1} p_k' \times \log p_k'; \quad (10)$$

where H' stands for the recalculated average energy entropy.

Endpoint detection

When AEE values for all sound frames are calculated, syllables of frog croaks can be properly extracted using the AEE endpoint detection method proposed in this paper. The steps are as follows:

(a) Set $n = 1$.

(b) Locate the maximum entropy value and its corresponding sound frame, and respectively set them to $H_n(m)$ and H_{ma}.

(c) Starting from $H_n(m)$, sequentially read the entropy values of $H_n(m + e)$ and $H_n(m - s)$, and determine whether $H_n(m)/A$ is larger than $H_n(m + e)$ and $H_n(m - s)$, where s and e are increments; A is a constant used to calculate the threshold value. If $H_n(m)/A$ is larger than $H_n(m + e)$ and $H_n(m - s)$, end this step; otherwise, continue searching in a wider range.

(d) Mark the segment from $H_n(m + e)$ to $H_n(m - s)$ as the n^{th} syllable.

(e) Replace the segment from $H_n(m + e)$ to $H_n(m - s)$ with the minimum value, ε.

(f) Specify $n + 1$, locate the maximum entropy value of the current sound frame, set it to $H_n(m)$, and determine whether H_{max}/A is larger than $H_n(m)$. If H_{max}/A is larger than $H_n(m)$, end the search; otherwise, return to step (c).

Taking Kuhl's creek frog as an example, after a temporal domain analysis, the signal, energy signals, negative entropy signals, and waveform after AEE signal processing are shown in Fig. 2A-D. A comparison of these figures reveals that for frog croak signals containing a peculiar clamor, the AEE method can effectively improve the entropy stability of the silent segments, and the peak points of the speech segments are more complete and obvious than the energy value, making it much easier to set the thresholds during endpoint detection. Therefore, when aided with appropriate endpoint deletion and combination, frog croak syllables can be appropriately segmented to facilitate successful recognition of the frog voice-print.

RESULTS

The voice-print recognition system referred to in this paper was programmed and developed using the graphic programming language, LabVIEW (National Instruments, Austin, TX, USA). The source audio files used in the experiments consisted of 12 types of frog croaks, partly provided by the Shanping Ecological Science Park of Liouguei Township of Kaohsiung City from their archives of recordings in the wild. Apart from that, files were also collected from the archives of wilderness recordings at numerous locations, including Yuanshan

Township of Ilan County, Xindian District of New Taipei City, Yuchi Township of Nantou County, Guanmiao and Guiren Districts of Tainan City, Maolin District of Kaohsiung City, Chunri, Manzhou, and Shizi Townships of Pingtung County, and Dawu and Daren Townships of Taitung County. The remaining types of frog croaks were collected from the internet. Therefore, 18 types of frog croaks (Table 1) were studied in this paper, with a total of 1079 frog croak syllables. Each sound file was re-sampled at 22.05 kHz, with the sound volume set to 16 bits and digitized in mono format.

Experiment of endpoint detection

To investigate the effectiveness of the AEE endpoint detection method proposed in this study, each segment of the audio recordings was mixed with white Gaussian noise at 6 different SNR levels (30, 25, 20, 15, 10, and 5 dB) to simulate different noisy scenarios. To each frog croak syllable, the correct starting and ending points of the syllable were first manually marked, followed by the automatic extraction of these points using the energy-based, ZCR, entropy-based, and AEE methods. If the automatically extracted starting point fell within the range of ±30% of the length of the manually marked starting point, it was considered correct; and likewise for the ending point. Taking Fig. 3 as an example, if the length a particular frog croak was about 540 ms, and the manually marked starting point was positioned at the 250-ms spot, then the correct range for the automatically marked point would be the duration of 88~412 ms; and similarly for the ending point. In total, 1079 syllables of frog croaks were manually marked during this experiment.

Table 2 shows the accuracy rates of endpoint detection of the 4 methods under an SNR environment of 5~30 dB. After plotting a chart with the data as shown in Fig. 4, we

Table 1. Types of frog croaks in the experiment

Family	Scientific Name	Common Name
Ranidae	*Rana adenopleura*	Olive frog
	Rana psaltes	Harpist frog
	Rana catesbeiana	Bull frog
	Limnonectes kuhlii	Kuhl's frog
	Rana latouchii	Latouche's frog
	Fejervarya limnocharis	Indian rice frog
	Pseudoamolops sauteri	Sauter's frog
	Rana swinhoana	Swinhoe's frog
Rhacophoridae	*Chirixalus eiffingeri*	Eiffinger's tree frog
	Polypedates megacephalus	White-lipped tree frog
	Rhacophorus moltrechti	Moltrecht's green tree frog
	Rhacophorus taipeianus	Taipei green tree frog
	Buergeria japonica	Japanese tree frog
Microhylidae	*Microhyla ornata*	Ornate narrow-mouthed toad
	Kaloula pulchra	Malaysian narrow-mouthed toad
	Microhyla heymonsi	Heymons's narrow-mouthed toad
	Microhyla steinegeri	Stejneger's paddy frog
Bufonidae	*Bufo bankorensis*	Common toad

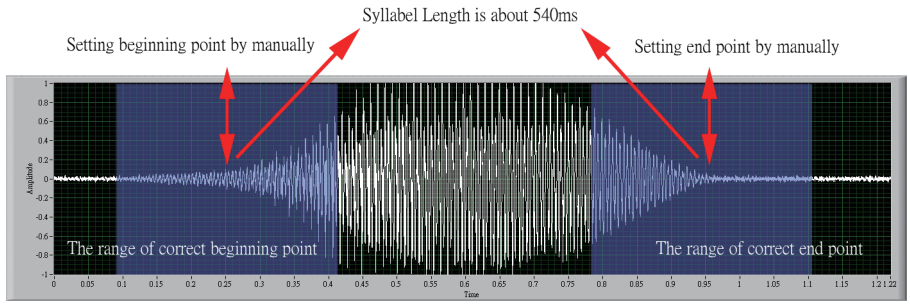


Fig. 3. An example of the correct range for automatic syllable segmentation.

Table 2. Experimental results of endpoint detection

SNR		Method			
		Energy	ZCR	Entropy	AEE
30 dB (%)	start	87.79	85.75	84.00	97.20
	end	90.08	87.02	78.12	96.18
25 dB (%)	start	86.00	83.97	83.72	96.69
	end	89.06	86.00	78.12	95.67
20 dB (%)	start	85.50	82.70	83.46	95.93
	end	88.30	85.24	77.61	95.42
15 dB (%)	start	83.72	81.68	82.70	95.93
	end	85.75	83.21	77.35	94.91
10 dB (%)	start	79.64	78.63	80.66	93.38
	end	79.90	79.13	76.08	91.86
5 dB (%)	start	73.03	73.03	78.70	89.31
	end	72.77	73.79	74.81	88.55

SNR, signal to noise ratio; ZCR, zero crossing rate; AEE, average energy entropy.

discovered that AEE outperformed the other 3 methods by 10~15% in terms of endpoint detection. In an SNR environment with a relatively smaller noise of 30 dB, the accuracy rates of AEE respectively reached as high as 97.2 and 96.18% for the start and end points; higher than the 87.79 and 90.08% of the energy-based method. Under the influence of background noise in the silent segments, the entropy-based method produced an entropy value that changed so abruptly that the detection of endpoints was seriously affected, and this method had the lowest accuracy rates among the 4 methods; namely 84 and 78.12%, respectively.

When the value of the background noise increased, the energy of the silent segments accordingly increased. This largely affected the accuracy of endpoint detection for the energy-based and ZCR methods. However, the entropy value of the entropy-based method did not correspondingly increase; therefore, the capability to suppress noise by the AEE methods was significantly superior to that of the energy-based and ZCR methods. [the ZCR method is better than the ZCR method??] For the energy-based and ZCR methods, the accuracy rates in recognizing the start and end points kept dropping with increases in the background noise.

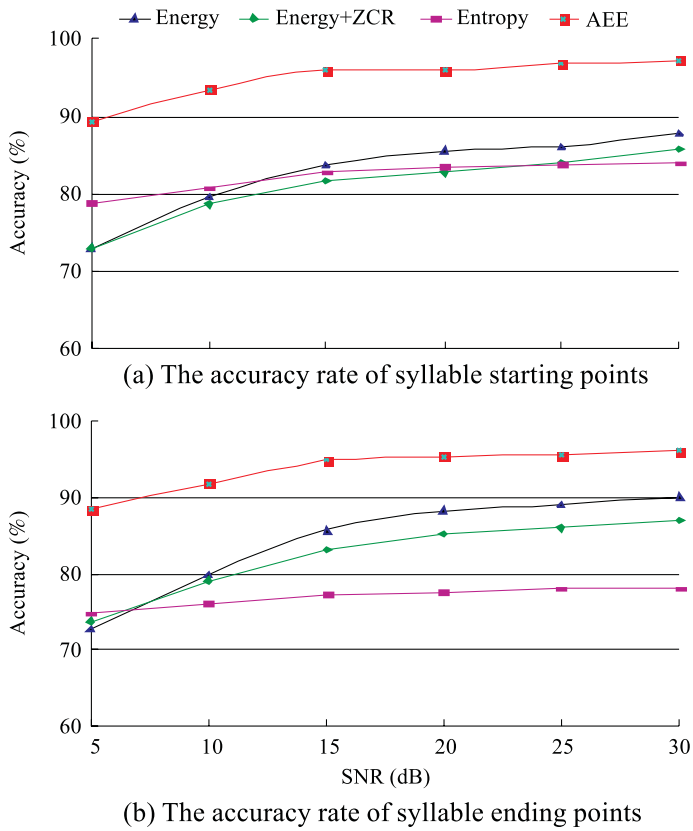


Fig. 4. Experimental results of the syllable extraction of endpoint detection.

When the SNR was at 5 dB, the accuracy rates of the energy-based and ZCR methods were even worse than those of the entropy-based method; however, the accuracy rates of the AEE method were still higher than those of the other 3. For the entropy-based method, which has a relatively good ability to suppress noise, the decrement in the accuracy rate was slower than those of the energy-based and ZCR methods. The AEE method proposed in this paper had a relatively high accuracy rate of frog croak syllable extraction, and was also less prone to interference in a high-noise environment. When detecting silent segments, the AEE method had a more-reliable performance than the conventional entropy-based approaches. From the results of this experiment, we observed that the AEE method was perfectly suitable for the endpoint detection of frog croaks.

Experiment in frog croak recognition

To compare the effects of the 4 methods of endpoint detection applied to recognizing frog croaks, 119 syllables of frog croaks were randomly selected from the 1079 total syllables collected for this experiment, and standard samples of frog croaks were made by manually extracting syllables. These syllables were saved in a database, and the remaining files were subsequently subjected to automatic syllable extraction using the 4 methods of endpoint extraction, to analyze the accuracy rates of syllable extraction. The LPCCs and MFCCs were feature parameters adopted in the voice-print recognition, and the feature

parameters included 14 dimensions. Dynamic time warping (DTW) (Myers and Rabiner 1998) was applied for recognition. Syllables were compared against all of the sample syllables, and the tested syllables with the minimal Euclidean distance to the sample syllable were the result of recognition. The accuracy was assessed with formula (11):

$$Accuracy(\%) = \frac{N_c}{N_s} \times 100\%; \quad (11)$$

where *Nc* is the number of the syllables that were discerned accurately, and *Ns* is the total number of syllables involved in the experiment.

In total, 960 syllables were involved in the experiment of this study. Each was compared using the 4 different endpoint detection methods along with DTW. The results of recognition by the LPCCs and MFCCs are listed in Tables 3 and 4, from which one can observe that the syllables extracted using AEE endpoint detection were more accurate. The AEE method was therefore superior to the other 3 endpoint detection methods. As to the recognition of the 18 types of frog croaks, the recognition rate reached as high as 90.21%, reinforcing the idea that the accuracy rate was closely related to the recognition rate.

DISCUSSION AND CONCLUSIONS

This paper proposes an AEE method to enhance the extraction of syllable endpoints. This method improves the stability of the energy distribution of spectra by adding an average energy value to each frequency component, and stabilizes the entropy values in silent segments; thereby reducing noise interference in a wild environment. Each sound clip recorded in the wild was first preprocessed, then the average energy points of the sound signals were added to the frequency components of each frame to obtain the parameter value of the AEE. Finally, syllables of frog croaks were segmented through the endpoint detection algorithm using the obtained parameters. The result was separately compared against those of the energy-based, ZCR, and entropy-based methods.

In the endpoint detection experiment,

Table 3. Experimental results of frog croak recognition with linear predicative cepstral coefficients

Method	Total syllables	Correct syllables	Recognition rate (%)
Energy + DTW	960	841	87.60
Energy + ZCR + DTW	960	827	86.14
Entropy + DTW	960	813	84.68
AEE + DTW	960	866	90.21

DTW, dynamic time warping; ZCR, zero crossing rate; AEE, average energy entropy.

Table 4. Experimental results of frog croak recognition with Mel-frequency cepstrum coefficients

Method	Total syllables	Correct syllables	Recognition rate (%)
Energy + DTW	960	861	89.69
Energy + ZCR + DTW	960	845	88.02
Entropy + DTW	960	840	87.05
AEE + DTW	960	891	92.81

DTW, dynamic time warping; ZCR, zero crossing rate; AEE, average energy entropy.

the method proposed in this study produced higher accuracy rates of extracting frog croak syllables, and also expanded its competitive advantage over the other methods in terms of accuracy. In the frog croak recognition experiment conducted using the 4 endpoint detection methods, this paper adopted the LPCCs and MFCCs with the DTW algorithm in the experiments. To express results of the recognition rate, the AEE proposed based on LPCCs and MFCCs in this paper achieved respective recognition rates of 90.21% and 92.81%, which were better than the other methods.

Sound files recorded in the wild typically contain background noise and sound artifacts derived from human interference. As a result, the spectra of frog croak signals are often seriously distorted in terms of the sound waveform, thereby weakening the representativeness of the feature parameters. Therefore, in future studies, our research team will continue to study noise suppression of feature parameters and alleviate the interference of noise with automatic frog croak recognition, in the hopes that manual work can eventually be replaced by automatic computer programs for audio file analysis.

LITERATURE CITED

Fagerlund S. 2007. Bird species recognition using support vector machines. EURASIP J Appl Sign Process 2007:1-8.

Harma A. 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. April 2003, Finland. Vol. 5. p 545-8.

Huang CJ, Yang YJ, Yang DX, Chen YJ. 2009. Frog classification using machine learning techniques. Expert Syst Appl 36(2):3737-43.

Ishimitsu S, Nakayama M, Yoshirni T. 2007. Construction of the Noise-Robust Body-Conducted Speech Recognition System. The Second International Conference on Innovative Computing, Information and Control. September 2007, Kumamoto, Japan. p 123-6.

Janakiraman R, Kumar JC, Murthy HA. 2010. Robust syllable segmentation and its application to syllable-centric continuous speech recognition. National Conference on Communications. October 2010 Chennai, India. p 1-5.

Jančovič P, Kőküer M. 2011. Automatic detection and recognition of tonal bird sounds in noisy Environments. EURASIP J Adv Sign Process 2011:1-10.

Lamel L, Labiner L, Rosenberg A, Wilpon J. 1981. An improved endpoint detector for isolated word recognition. IEEE T Acoust Speech 29(4):777-85.

Lee CH, Chou CH, Han CC, Huang RZ. 2006. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recogn Lett 27(2):93-101.

Myers C, Rabiner LR. 1998. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE T Acoust Speech 28(6):623-35.

Rabiner L, Juang BH. 1993. Fundamentals of speech recognition. Englewood Cliffs, NJ: Prentice-Hall, c1993. 507 p.

Shen JL, Hung JW, Lee LS. 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments. International Conference on Spoken Language Processing. November 30-December 4, 1998. Sydney, Australia. p 1-4.

Taylor A, Grigg G, Watson G, McCallum H. 1996. Monitoring frog communities: an application of machine learning. Proceedings of the Eighth Innovative Applications of Artificial Intelligence Conference. August 1996, Portland,

OR. p 1564-9.

Tian Y, Wang Z, Lu D. 2002. Nonspeech segment rejection based on prosodic information for robust speech recognition. IEEE Signal Process Lett 9(11):364-7.

Tu WH, Hung JW. 2007. Study of the voice activity detection techniques for robust speech feature extraction. Proceedings of the 19th Conference on Computational Linguistics and Speech Processing, September 2007, Taipei, Taiwan.

Zhao X, O'Shaughnessy D. 2008. A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation. Canadian Conference on Electrical and Computer Engineering. May 2008, Niagara Falls, Ontario. p 145-8.