# A Tool for Cutting Large Speech Corpora: HCI4CS

Lin Guo, Yang Bai, Jie Su, Wenlin Pan* and Tianjun Zhang
Yunnan Minzu University, Kunming, 650500, Yunnan Province PR China
*Corresponding author

***Abstract*—There are two methods to cut large speech corpora, include traditional manual segmentation and machine automatic segmentation. The quality of segmentation can be controlled easily using traditional manual segmentation. However, the shortcomings of manual segmentation were also obviously such as inefficiency, high cost. As we all know, the method of machine automatic segmentation has the advantage of high efficiency, but the fussy work to find cutting error can't be omitted. Thus, this paper developed a tool of human-computer interaction for cutting speech corpora (HCI4CS), which provides segment algorithm, parameter to control, modifying the error of automatic segmentation results and generates labeling files for HTK toolkit. The research object was one thousand speeches of Primi. Using HCI4CS, a person with low cognitive competence about cutting speech corpora can achieve nearly one hundred percent accuracy.**

***Keywords-primi speech; human-computer interaction; speech segmentation***

## I. INTRODUCTION

At present, the audio signal is widely used in speech recognition, the construction of corpus and voice product design, etc. The early stage of the segmentation processing for the original audio signal is based on the application requirements. The methods of speech segmentation are mainly adopted traditional manual segmentation and machine automatic segmentation.

Manual segmentation mostly use speech processing software (e.g., praat, cool edit, audition, etc.), which have many disadvantages such as time-consuming, heavy workload. In addition, it is not suit for the products needed high precision corpora speech recognition. This method needs to subjective judgment by the operator who may reduce the precision of audio segmentation and have a negative effect to the product performance.

Machine automatic segmentation system which has high precision in segmenting and improves the working efficiency. Nevertheless, it requires manual processing automatic segmentation of the audio, and also needs to build speech template in earlier stage and provide a priori knowledge of the number of segmentation.

However, it hasn't enough condition to established minority language template corpus, such as Primi speech, Jing Po speech, etc. Even if the previous speech template and priori knowledge is provided, segmentation accuracy of the machine automatic is impossible to achieve 100%. It is inevitable to use manual proofreading segmentation audio one by one in later stage. Considering the number of audio is multiply *N* times after segmentation, which will lead to sharp increase of staff workload and research costs.

This paper developed a tool for cutting large speech corpora based on human-computer interaction named HCI4CS. It realizes the audio segmentation process visualization and controllable through human monitoring speech segmentation. Once an error of machine automatic segmentation occurred, the operator can realizes the problem timely and correct the error segmentation, which improve the efficiency of audio segmentation and the cutting precision. The tool reduces levels of knowledge and technical ability requirements for an operator. It is popular, general and ascendant to the late product marketing.

## II. REQUIREMENTS OF HCI4CS

HCI4CS was designed for research and application of Primi speech corpora recognition. It selects isolated words of Primi (character, word or phrase) as primitive to cut.

Speech corpora acquisition and processing requirements: before recording, calibrating Primi speech was applied to pronounce. In recording process, one speaker pronounced a single unit many times and stored as one file. In the end, all audio files are cute as many primitives for speech recognition and establishing corpus.

Firstly, HCI4CS can automatically detect starting point and endpoint of voice primitives with different signal characteristics. Then, when it goes wrong, such as treating mute parts of a primitive as a starting point of a new unit, operator can timely discover the segmentation error through the cutting window. Finally, the operator adjusts segmentation to the correct range.

Presently, speech endpoint detection is a very critical step in the pre-processing of speech signals [1-4]. The common methods of speech endpoint detection are time domain, frequency domain and model matching [5-7]. In the process of recording, however, there are uncontrollable factors, such as breathing, cough, turning page or other noises, so noise in the speech signal must be handled. According to the difference of energy, speech signal can divided into speech and mute. The noise which in spectrum and signal energy is similar to speech segment is called interfering noise. As existent algorithms cannot eliminate all interfering noise, and the speech endpoint detection method, based on spectral entropy, just cut part of speech segment interfering noise [8]. So raising precision of segmentation should take manual adjustment [9].

The process of audio recording based on text numbers exist many uncontrollable factors, like the number of speech

primitives and the interfering noises in one audio file. Hence, this paper developed HCI4CS with high efficiency, high precision, low degree of recognition, and the design of manual monitoring. The structure is shown in Figure I.

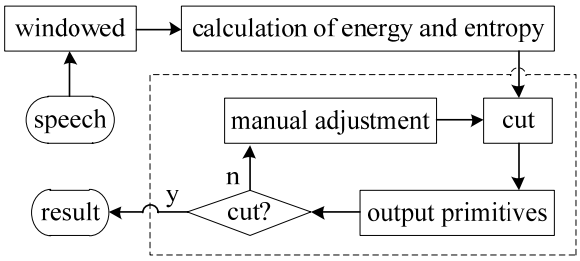


FIGURE I. STUCTURE OF HCI4CS.

HCI4CS is composed of automated processing of audio signal and manual monitoring.

*1) Signal processing:* It includes windowed and framed, calculating the ratio of energy and entropy. The effect is to calculate the ratio of energy and entropy of each frame.

*2) The segmentation:* Audio signal are cut by using threshold which is defined as the specific ratio of energy and entropy. When some interfering noises and speech segments of the audio are similar in energy and frequency, since the conventional endpoint detection algorithm cut audios simply according to threshold segmentation, that interfering noises cannot be completely distinguished. Meanwhile, audios are incapable of segmenting by prior knowledge of known numbers, because of uncertainty numbers of audio primitive and segments. Therefore, audios cutting have to join manual work. That's why the tool called the speech cutting based on human-computer interaction.

## III. MAIN SEGMENTATION ALGORITHM AND PARAMETERS

Short-time energy can detect phonetic segment in the audio signal [10], but speech endpoint cannot be precisely detected only by short-time energy. Like speech energy of endpoint is very weak when the beginning and the end of speech is weak fricative, plosive or rhinolalia, which speech endpoint may be easily confused with the interfering noise to inaccurately detect the speech endpoint and to a phenomenon of partly cutting the speech. For instance, Figure II shows the short-time energy of a Primi language "tomorrow /siãl/" which contains fricative "/s/". If fricative "/s/" that its peak is p1 cannot be completely detected, the segmentation of the speech will turn to "/diãl/". Shown in Figure III is spectral entropy of "tomorrow /siãl/", and [a, b] section is the spectral entropy of fricative "/s/". From Figure III, we find that the ratio of energy and entropy can be used to cut speech because in audio signal the spectral entropy of speech segment is lower than the spectral entropy of noise

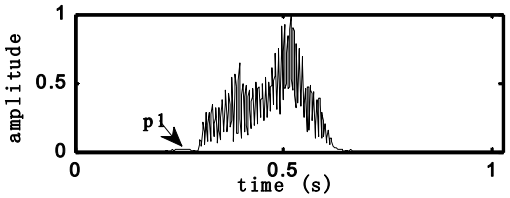


FIGURE II. SHORT-TIME ENERGY OF SPEECH.

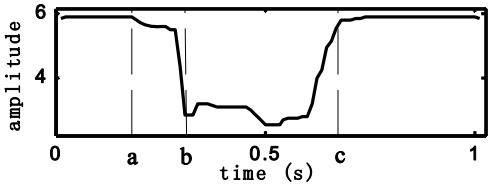


FIGURE III. SHORT-TIME SPECTRAL ENTROPY OF SPEECH.

The first cutting step is to frame the longer speech signal. In other word, the signal after sampling multiplies a finite length window function. Hamming window is generally chosen to frame the signal because of its character of bandwidth and spectrum leakage.

Let $x(n)$ be the speech signal. After adding window function $h(n)$, the $i$ frame of speech signal is $x_i(n)$. Therefore, $x_i(n)$ satisfy:

$$x_i(n) = x((i-1)T+n)h(n), n = 1,2,\cdots,L, i = 1,2,\cdots,f\ . \quad (1)$$

Let $L$ be the frame size, $T$ be the frame shift and $f$ be the total number of frames after framing. The hamming window $h(n)$ is

$$h(n) = \begin{cases} 0.54 - 0.46(2\pi n(L-1)) \\ 0 \end{cases}. \quad (2)$$

The short energy of the $i$ frame defined as

$$P_i = \log_{10}\left(1 + \frac{1}{a}\sum_{n=1}^{L} x_i^2(n)\right). \quad (3)$$

The spectral entropy of the $i$ frame defined as

$$Q_i = -\sum_{n=0}^{K} p_i(k)\log_{10}(p_i(k)), \quad (4)$$

And let $K$ be the length of FFT, $p_i(k)$ be the frequency component of the $k$ of the $i$ frame.

The rate of short energy and spectral entropy of $i$ frame is defined as

$$E_i = \sqrt{1+\left|P_i/Q_i\right|}\,. \tag{5}$$

## IV. DESIGNS OF HCI4CS

The segmentation system of HCI4CS use the rate of energy and entropy of background voice as the threshold to cut speeches, and the difficulty of the segmentation process is removing interfering noise. Commonly used treatment method is to set the auxiliary prior knowledge, like the known primitive segments of a speech file combining with similarity judgment of the audio. This method is suitable for accordant numbers between the recording of the primitive segment and the make-able text. If interference noise, adhesion of two primitives or randomness of speech primitive segments exist in recording process, adopting the method to operate will reduce cutting rate, at the same time increase manual workload to proofread segmentation audio, and reduce working efficiency. Therefore, taking manual to detective the segmentation will enhance the cutting efficiency and the segmentation accuracy. Manual adjustment of segmentation is shown in Figure IV.

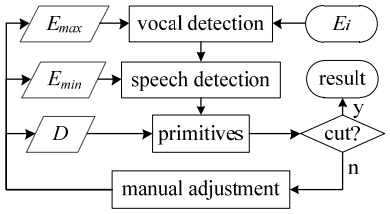


FIGURE IV. ADJUSTMENT BASED ON HCI.

Segmentation steps as follows:

*1) Setting threshold value:* The backgrounds of recording process basically remain unchanged, so $E$ which is the ratio of energy and entropy of the audio background can be selected as the basis of the threshold. Let $E_{max}=E+T_1$ and $E_{min}=E+T_2$, in which $T_1$ represent increment of high threshold and $T_2$ represent increment of low threshold.

The $i$ frame ratio of energy and entropy $E_i$ will be gotten after dividing frame of the speech. Then using $E_{max}$ detects speech segment of the radio. Finally, using $E_{min}$ detects the effective speech segment of the radio.

*2) Setting the interval value D:* After the detection of effective speech segment, for avoiding the system treat the middle pause of the speech primitive as a new start point of a speech unit, $D$ will be set to satisfy $D(V_i, V_{i+1})> D$ ($D(V_i, V_{i+1})$ is the interval time of speech primitives $V_i$ and $V_{i+1}$), and $D(V_i)< D$ ($D(V_i)$ is the adjacent interval time in the speech $V_i$).

*3) Checking the audio segmentation:* An audio file is composed of a speech which was repeatedly recorded a single word by one speaker, in order to meet the demands of Primi speech recording and make recording process easier. The automatic cutting process is unable to identify interfering noise and adhesion of two primitives, so it requires manual intervention to adjust the segmentation.

Figure V is the control panel of HCI4CS. The panel can choose the directory file of opening, saving the form of wav and labeling, and set signal to noise ratio (SNR), document number, file name, increment of high threshold, increment of low threshold and interval value according to particular case of recording speech. Click on the "RE" button, you can reset parameters. To batch large-scale speech, the tool on the basis of setting parameters automatically segment speeches and show the cutting process on a new window. The operator monitors the whole process of segmentation in real time through the window. The segmentation is shifted from automatic to manual when the operator find the cutting is wrong and click "pause/start" button in the control panel.

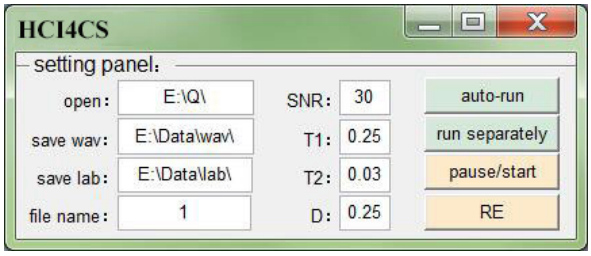


FIGURE V. THE CONTROL PANEL OF HCI4CS.

For example, Primi word "tiger teeth". Table I shows its parameter to segment and Table II is the threshold to be chose. Figure VI shows part of speech waveform and Figure VII is the rate of energy and entropy. In Figure VII, p1 represent coughing noise, p4 is the noise of turning page, p7 and p8 are other noises. The process of segmentation as follows:

TABLE I. THE SPEECH PARAMETERS.

| parameters | numerical value |
|---|---|
| encoding | ACM |
| sampling rate | 44.1khz |
| resolution | 16bit |
| frame size | 30ms |
| sound channel | single track |

TABLE II. THE ADJUSTABLE PARAMETERS.

| threshold value | numerical value |
|---|---|
| $T_2$ | 0.03 |
| $T_1$ | 0.25 |
| $D$ | 5 |

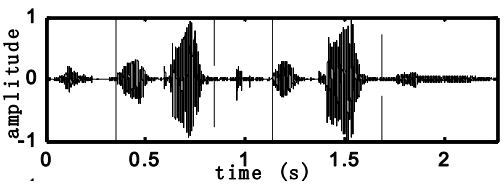


FIGURE VI. OSCILLOGRAPH OF THE SPEECH.

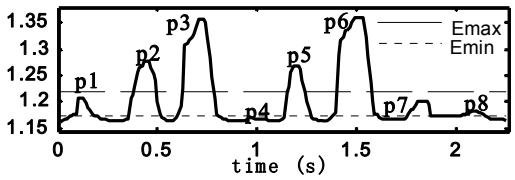


FIGURE VII. THE RATE OSCILLOGRAPH OF THE SPEECH.

The threshold increments of $T_1$ and $T_2$ are set to detect effective speech segment. Among audio segment of p1 to p8, p2, p3, p5 and p6 are detected as effective speech segment. Then p2 and p3, p5 and p6 are severally divided to one primitive because the threshold value of p2 and p3, p5 and p6 are less than $D$. Finally, the operator makes sure of right cutting method through checking the graph of speech segmentation.

Under special circumstances, part of noises will be detected as effective speech segment. The tool didn't operate a wrong segmentation to the speech, but it is improper to the cutting speech. So it needs manual intervention to denoise. Primi word "tiger teeth" which mix with interfering noise was selected at the same parameters. Figure VIII shows its oscillograph, and Figure IX is the ratio of energy and entropy. It needs manual control, on account of the noise of p1 detected as effective segment, to increase the numerical value of $T_1$. The interfering noise can removed through pulling the dividing line, when the interfering noise cannot be wiped out by adjusting parameters.

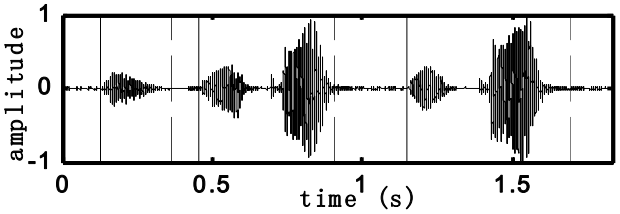


FIGURE VIII. OSCILLOGRAPH OF THE SPEECH.

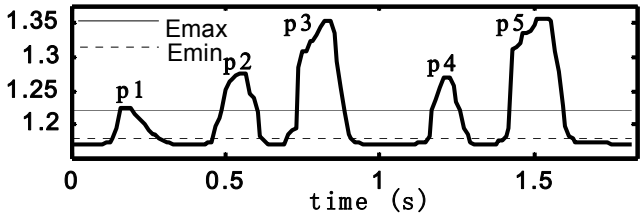


FIGURE IX. THE RATE OSCILLOGRAPH OF THE SPEECH.

## V. EXPERIMENTAL RESULT

One thousand Primi speech files that each file is a pronunciation of one word, recording in the studio, were selected to detect. The test of segmentation divided into computer and human-computer interaction, and the result is shown in Table III.

TABLE III. THE COMPARISON RESULT OF SEGMENTATION.

| | R[a] | W[b] | N[c] |
|---|---|---|---|
| **computer** | 0 | 220 | 7900 |
| **human-computer interaction** | 40 | 0 | 8103 |

a. The number of audio files adjusted by human in the process.

b. The detected number of audio files segmented wrong after speech cutting

c. The total number of audio files after speech segmentation

When R=0 that human didn't participate in the segmentation, 220 error files were detected after the segmentation, and faults are that the speech noises are cut as effective speeches, effective speeches cannot be completely segmented, and several units of the speech were cut as one primitives. When R=40 that 40 audio files were adjusted through human-computer interaction, the accuracy is 100% after checking the speech files of segmentation.

Above all, adopting human-computer interaction as a cutting method can improve machine automatic cutting efficiency to audio files. At the same time, the manual adjustments of the segmentation error enhance the cutting accuracy and avoid audio proofreading process later, etc.

## VI. CONCLUSION

Today, the method of cutting large speech corpora is mainly adopted traditional manual segmentation or machine automatic segmentation. However, manual cutting method is inefficiency and has high cost. The method of machine segmentation is unable to avoid audio cutting error. It is inevitable to check all audio files. In this paper, a new tool of HCI4CS is developed. The experiment selected speech files of Primi isolated word as a unit to cut. Using machine automatic segmentation can improve the audio files' cutting efficiency, and effectively overcome the problem of noise removal in cutting speech process. The experiment proves that it arrived at a high accuracy of speech segmentation. The follow-up work is to realize the automatic segmentation of the two adhere elements, and reduce manual operation and so on.

## ACKNOWLEDGMENT

This work was financially supported by Science Research Fund Project of Yunnan Provincial Education Department (2014Z091), Innovation Program of Yunnan Minzu University (2015YJCXY285) and Key Laboratory of IOT Application Technology of Universities in Yunnan Province.

## REFERENCES

[1] H. P. LIU, X. Li, and B. L. Xu, "Summary and survey of endpoint detection algorithm for speech signals," J. Application Research of Computer, pp. 2278-2283, August 2008.

[2] X. S. Xue, "Research on speech endpoint detection based on the improved dual-threshold algorithm," J. Electronic Design Engineering, pp. 78-81, April 2015.

[3] J. M. Zheng, and P. Zhang, "Audio segmentation based on wavelet transform," J. Computer Engineering and Applications, pp. 139-142, July 2011.

[4] M. L. Wang, “Research on the speech recognition of the isolated words in the Lahu language based on speaker independence,” J. Journal of Yunnan Minzu University of the Nationalities, pp. 337-340, April 2015.

[5] J. X. Zhang, L. X. Shi, and D. S. Wang, “Speech automatic segmentation system modeling with adaptive threshold adjustment,” J. Computer Engineering and Design, pp. 1886-1889, August 2010.

[6] G. Y. Li, H. Z. Yu, and Z. Q. WU, “An automatic phoneme segmentation method in continuous Tibetan language under the condition of resource-deficiency,” J. Computer Engineering & Science, pp. 2009-2013, October 2014.

[7] R. J. Zhang, B. C. Li, and D. Qu, “Audio Segmentation Algorithm Based on Trend of Believable Degree Change,” J. Computer Engineering, pp. 177-179, August 2010.

[8] Z. M. Li, and F. Shang, “A Speech Endpoint Detection Method Based on Spectral Entropy,” J. Electronic Technology & Software Engineering, pp. 200-202, January 2015.

[9] X. Fan, and Q. H. Zhang, “Review on Standardization for User Interface and Human–Computer Interaction,” J. Information Technology & Standardization, pp. 28-31, April 2015.

[10] G. Zhuo, and J. Jiang, “Research of Tibetan Speech Endpoint Detection Based on Short-term Average Energy and Short-term Zero Rate Algorithm,” J. Computer Knowledge and Technology, pp. 7466-7469, November 2014