Enhance Speaker Segmentation by Elaborating Utterance Detection

Min Yang, Zhaohui Wu , Yingchun Yang
College of Computer Science and Technical
Zhejiang University
Hangzhou, P.R.China, 310027
{rymth,wzh,yyc}@zju.edu.cn

Abstract

In this paper, we introduce an elaborate utterance detection algorithm to enhance speaker segmentation. Silence detector, further divider and audio type classifier are employed in this elaborate utterance detection, to make this algorithm adaptive for both silent and noisy environments. Open-set verification testing has taken on the Hub4-NE broadcasts database. The experiment results show that this enhanced segmentation method can provide better information for speaker models.

1. Introduction

Segmentation of continuous audio is widely used as an important pre-procedure of audio processing. In speaker processing such as the Automatic Speaker Recognition (ASR) and Spoken Document Retrieving (SDR), speaker segmentation is even crucial, because the performance of these processing are relied on it heavily.

Most of conventional speaker segmentation methods are distance measure of speaker features. Change points are detected by checking distances between related window pairs. The most widely adopted distances are Gaussian distances [1, 2, 3] and the Bayesian Information Criterion (BIC)[4]. While the performances of segmentation are similar among those distances, the more important factor is the selection of window pairs in distance calculation. However, in current segmentation methods, the selection of window pairs are usually violent. This makes the distance measure less efficient and less effective, because there may be noise, silence or other disturbance in chosen windows, and that makes the distance value unexpectable, hence the highest distance value does not often occur at real speaker change points. In our algorithm, elaborate utterance detection is introduced as a pre-segmentation methods, which divide audio into short voice clips. Distance values are then checked among those clips to find out the real speaker change points.

In the next section, we present our method of elaborate utterance detection. Section 3 is a brief introduction of distance based segmentation method we use. Experiment results are shown in section 4. Finally we give a summary and conclusion in section 5.

2. Elaborate Utterance Detection

The purpose of elaborate utterance detection is to detect the boundaries of each sentence. Speaker changes are more likely to happen during one sentence's ending and beginning of another sentence, compared to the possibility of that to happen during sentences. So we can only check distances of consecutive utterance clips to obtain speaker change points. This will not only save computation time but also remove the negative efforts brought by impure window pairs in traditional distance measures.

To archive this object, we use silence detection first. Pauses widely exist between sentences. In ideal environment, pauses can be detected by silence detection. But in the environment which abounded with noise or background music, using only silence detection does not work well. So we use a χ^2 distribution based method to perform further division, which is sensitive to audio content change. At last we use a trained decision tree classifier to discriminate speech and non-speech. Features used in utterance detection are zero crossing rate (ZCR) and root mean square energy (RMS), which are extracted from 20ms-long frames without overlapping.

2.1. Silence Detection

Energy is the most indicator of silence. First, to be self-adaptive, a threshold value of RMS is determined from the statistic of all RMS values in target audio. If a frame's RMS value is lower than this silence threshold, it is considered as silence frame. 10 or more silence frames in a sequence, in total length of 200ms or longer, will be taken as a potential silence segment. However, not all of these potential silence segments are finally decided to be silence.

In most languages, words start or end with surd syllables which have low energy but high zero crossing rates [5]. In Fig.1, the sixth to tenth frames are the pronouncing of /f/, but the corresponding RMS values are similar to silence. To

0-7803-9332-5/05/$20.00 ©2005 IEEE

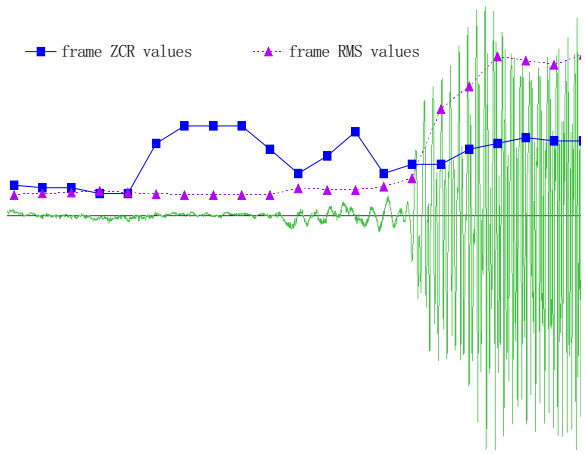

Figure 1: Wave form, ZCR and RMS values at the beginning of "four"

avoid this mistake, we use another threshold of ZCR value, which is decided from ZCR values in potential silence segments. Then, at both sides of each potential silence segment, those frames with high ZCR values are removed and merged to adjacent non-silence segments.

2.2. Further Division

In ideal environment, silence detection can find out all the utterance boundaries. But in most audio, there are not only speech but also plenty of noise or music, and this makes silence detection incapable. How to evaluate the performance of silence detection to decide whether the further division is needed or not? We treat each non-silence segments regarding by their length.

In our statistic, the most possible utterance length is between 0.5 seconds and 15 seconds, as shown in Fig. 2. According to this, short segments with length less than 0.5 seconds are abandoned as non-speech segments immediately, because these short-time sounds between silences are noise or meaningless voice. Every segment with proper length of between 0.5 and 15 seconds is considered as a sound clips and does not need further division. If a non-speech segment is longer than 15 seconds, further division is taken on this segment to obtain more potential boundaries.

C. Panagiotakis *et al.* in [6] presented a division method which assumes that acoustic features satisfy χ^2 distribution and that audio can be divided by comparing these distributions. To divide those long segments, 50-frame-long windows which shift 10 frames at one time are applied. In each window, a generalized χ^2-distribution is estimated to simulate the 50 RMS values. The probability density function of

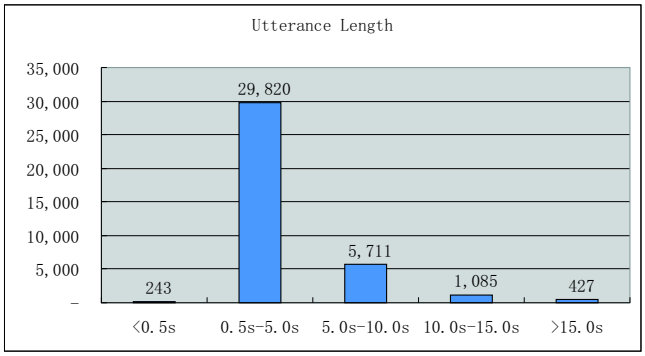

Figure 2: Statistic of utterance length in broadcasting news database

generalized χ^2-distribution is:

$$p(x) = \frac{x^a e^{-bx}}{b^{a+1}\Gamma(a+1)}, (x \geq 0). \quad (1)$$

Parameters a, b are related to the mean and the variance values:

$$a = \frac{\mu^2}{\sigma^2} - 1, \text{and } b = \frac{\sigma^2}{\mu}. \quad (2)$$

The distance value of a window is:

$$D(i) = 1 - \frac{\Gamma(\frac{a_1+a_2}{2}+1)}{\sqrt{\Gamma(a_1+1)\Gamma(a_2+1)}} \frac{2^{\frac{a_1+a_2}{2}+1} b_1^{\frac{a_2+1}{2}} b_2^{\frac{a_1+1}{2}}}{(b_1+b_2)^{\frac{a_1+a_2}{2}+1}}, \quad (3)$$

where a_1,b_1 are parameters a,b in the precede window, and a_2,b_2 are those of the succeed window. Windows with peak distance values are selected as candidates. Then, to make the division more exact, distances are computed frame by frame in these candidate windows. Finally the long non-silence segments are divided into short sound clips at those frame with highest distance.

2.3. Decision Tree Classifier

Not all sound clips are speech, so we need a classifier here to find out utterance clips. Models like GMM or HMM are not proper for this task, because clips are too short and acoustic features are insufficient. Hence we have carefully chosen some features about clips and use a decision tree to classify. Each node in this decision tree corresponds to a rule of a certain clip feature. Order of these rules and thresholds in each rule are trained to optimize its performance.

The first two clip features in decision tree is HZCRR and LRMSR, which were first introduced by Lie Lu *et al.* in [7]. According to [7], these two values are higher in speech than in music, and we found them also higher than in noise. HZCRR and LRMSR are defined as:

$$\text{HZCRR} = \frac{1}{2N}\sum_{n=0}^{N-1}[\text{sgn}(\text{ZCR}(n) - 1.5 * \overline{\text{ZCR}}) + 1], \quad (4)$$

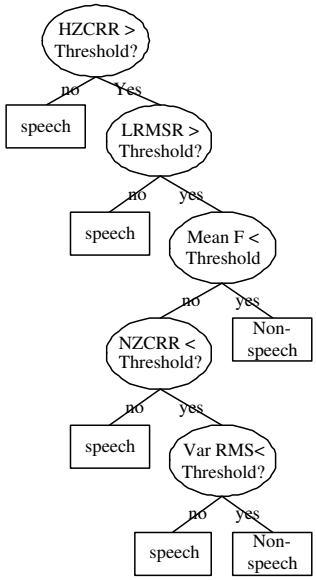

Figure 3: Topology of audio type decision tree

$$\mathrm{LRMSR} = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(0.5*\overline{\mathrm{RMS}}-\mathrm{RMS}(n))+1], \quad (5)$$

where N is the number of frames in a clip.

Maximal mean frequency is selected in this decision tree, because there is an upper limit of human voice while instrument or noise can easily surpass this value. We can use the highest ZCR value in a clip to estimate this feature.

The probability of null zero-crossings can can discriminate 40% speech against non-speech according to [6]. It is defined as the likelihood of appearances that a frame's ZCR value equals to zero. To be practical, ZCR values lower than a low value is also deemed to be zero when calculate this segment feature. Speech clips have higher value of this feature generally.

The final rule is normalized RMS variance, while it is lower in speech and higher in non-speech. Normalized RMS variance is defined as

$$\varphi = \frac{\sigma^2}{\mu^2} \quad (6)$$

in which σ^2 is the variance of RMS, and μ is the mean RMS value in this clip.

The threshold values in this decision tree are trained from recorded audio data in total length of more than 2 hours, which include one and half hours of daily life conversation and 40 minutes of music and noise. Clips are derived from these training data by using the same way described above. The average length of clips is 2.73 seconds. This simple decision tree classifier performs very well both in training and testing, the optimized accuracy of the training data is 97.65%, and with these thresholds from training, we put 94.06% of the clips in correct case in the testing data.

The final classification result is smoothed to further improve accuracy: if a clip's audio type is different from both its precedent and subsequence, none of which are silence, then the audio type of this clip is smoothed to the same as its neighbors. This smooth can increased the accuracy of training data to 97.78%, and the testing data, 95.49%.

3. Distance Based Segmentation

At the beginning of segmentation procedure, speaker features, 12-dim MFCC and energy, are extracted from 32ms-long, 10ms-shifting frames in each speech clip. Then each pair of adjacent speech clips, with only silence or without any other clips between them, are selected as window pairs to calculate distance. Among all the distance measures, we find that the T^2 statistic measure,which was first used by Rongqing Huang et al in [2], fits our algorithm best. Comparing to other distances, T^2 statistic measure has better performance when the length of speech segments are not same. The distance value is defined by:

$$D_{T^2} = \frac{ab}{a+b}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) \quad (7)$$

where features in two clips are consider to obey Gaussian distribution $N_1(\mu_1,\Sigma_1)$ and $N_2(\mu_2,\Sigma_2)$, and a, b stands for the number of frames in each clips. The covariance of all feature vectors are introduced as the equal covariance matrix.

An adaptive threshold value of distance is determined by the mean and variance of all distance values:

$$T = \mu - 1.5\sigma. \quad (8)$$

If a distance value is lower than this threshold, the corresponding pair of speech clips are deemed to be in the same speaker's speech, so they should be merged, as well as the silence segment between them, if any. This threshold value is somehow low, but it is acceptable for the penalty of false alarm is not as serious as missing.

4. Experiments and Results

To evaluate the performance of our segmentation algorithm, open-set speaker verification experiments have been taken on the 1997 Mandarin Broadcast News Speech Corpus (Hub4-NE) data, which are recorded broadcasts from CCTV, KAZN and VOA. This database has a total time of approximately 40 hours, 30 hours of which are speech. The transcript of this database is originally for speech recognition, so it also contains all the needed information. Ten most frequently appearing reporters in this database are selected, to train their 32-dim Gaussian Mixed Models. Then we apply our segmentation algorithm, BIC[4] method and KL2

distance[8] on the whole database. Segments are sent to model testing and the equal error rate(ERR) of GMM are calculated. Also, recorded segment results are compared with transcripts. Two indicators are obtained to describe the performance of segmentation methods:

a)Pure rate as the relative amount of the pure speech segments obtained in all voices:

$$pure\% = \frac{\Sigma t}{\Sigma T} \times 100\%, \tag{9}$$

where t is the length of segments which only contain voice of one speaker and T is the real speech segments' time. This indicator represent the ability of obtain useful segments.

b)Length rate as the evaluation for the average length of each pure segment:

$$length\% = \frac{\overline{t}}{\overline{T}} \times 100\%. \tag{10}$$

The meanings of t and T are same as (8).This indicator suggests the false alarms of speaker change points. The lower this rate is, the more false alarms have occurred.

In the BIC method, the penalty weight is set to 1.3 according to [4]. The threshold value of KL2 distance is set to minimize the EER.

Table 1: Evaluation of segmentation on Hub4-NE

	pure(%)	length(%)	EER(%)
BIC	72.39%	34.91%	15.91%
KL2	71.69%	36.01%	25.84%
Enhanced	80.40%	37.89%	12.94%

The results show that our algorithm can get much more pure speech segments from audio files, and also longer compared to BIC segmentation and KL2 segmentation. The ERR of GMM proves that our segment result fits speaker modelling much better. Most of impure speech segments are short sentences that are shorter than 1 second as well as some undetected noises like cough or laughter attached to speech.

5. Conclusion

We have introduced elaborate utterance detection and improved speaker segmentation in this paper. Our elaborate utterance detection can work well in both silent and noisy environments. Silence detection is designed for silent environments, while in noisy environments, the χ^2 distribution method can find out all potential point of audio content changes. A decision tree about segment features is introduced as audio type classifier to discriminate speech and non-speech.

This utterance detection method can make distance measure more effective and more efficient by providing pure segments and potential change points. Equal error rate of GMM in speaker verification experiment is 12.94%, which is much lower comparing to other segmentation methods.

Future work will be focused on fusion with prosodic features like pitch to make utterance detection more exactly.

Acknowledgments

This work is supported by National Natural Science Foundation of P.R.China (60273059), Zhejiang Provincial Natural Science Foundation for Young Scientist of P.R.China (RC01058), Zhejiang Provincial Natural Science Foundation (M603229) and National Doctoral Subject Foundation (20020335025).

References

[1] Olivier *et al.*, "Applied Clustering for Automatic Speaker-Based Segmentation of Audio Material," *Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL) Special Issue: OR and Statistics in the Universi-ties of Mons., volume 41, No. 1-2, 2001*

[2] Rongqing Huang,John H.L. Hansen, "Advances in Unsupervised Audio Segmentation for the Broadcast News and Ngsw Corpora," *ICASSP 2004*

[3] H. Gish, M. Siu, R. Rohlicek,"Segregation of Speakers for Speech Recognition and Speaker Identification," *Proc. of ICASSP-91, pp. 873-876, 1991*

[4] Alain Tritschler,Ramesh Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion," *Proc. of the EuroSpeech'1999.*

[5] L.Rabiner, M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J., Vol. 54, pp. 297-315, 1975*

[6] C. Panagiotakis, G. Tziritas,"A speech/music discriminator based on RMS and zero-crossings," *IEEE Transactions on Multimedia, 2003*

[7] Lie Lu, Hao Jiang, Hongjiang Zhang, "A Robust Audio Classification and Segmentation Method," *Proc. of the 9th ACM International Multimedia Conference and Exhibition, pp. 103-211, 2001*

[8] MA Siegler, U. Jain, B. Raj, RM Stern, "Automatic segmentation, classification and clustering of broadcast news. in DARPA Proc," *Speech Recognition Workshop, 1998*