Content-free Topic Segmentation with Acoustic Features (Report)

Jing Su
School of Computer Science and Statistics
Trinity College
sujing@tcd.ie

September 11, 2014

1 Introduction

In my previous work, content-free topic segmentation is approached by classification methods, and the unit is Vocalization [6]. Speaker ID, vocalization start time, vocalization duration, pause, overlaps and their corresponding *Horizon* features are emphasized. This followed an approach to segmentation and classification introduced by Luz [2, 3] for analysing recordings of multidisciplinary medical meetings.

In this study, I follow previous experiment settings, but focus on acoustic features, exploring the effect of acoustic features on topic segmentation/ vocalization classification. Zero-crossing rate (ZCR) and root mean square (RMS) are well studied features in audio analysis. In the following sections, I explain the method to extract ZCR and RMS from WAV files, and integrate them to ARFF files.

2 Methodology

2.1 Zero-crossing Rate

ZCR is defined as the number of sign changes in a frame divided by the length N of the frame:

$$ZCR = \frac{1}{N}\Sigma_{n=2}^{N}|sgn[x(n)] - sgn[x(n-1)]| \quad (1)$$

ZCR is highlighted since it has advantage in detecting two main audio characteristics. First, ZCR indicates the spectural centroids of a signal from which the dominant frequency can be estimated. Second, speech, music and noise signals have different ZCR distributions.

Speech is generally composed of frequently changed voice and unvoiced sounds. Voice presents low ZCR values and unvoiced sounds has high ZCR values. Consequently, speech shows relatively large variance.

A sample of ZCR values for music, silence, speech and noise is shown in Figure 1.

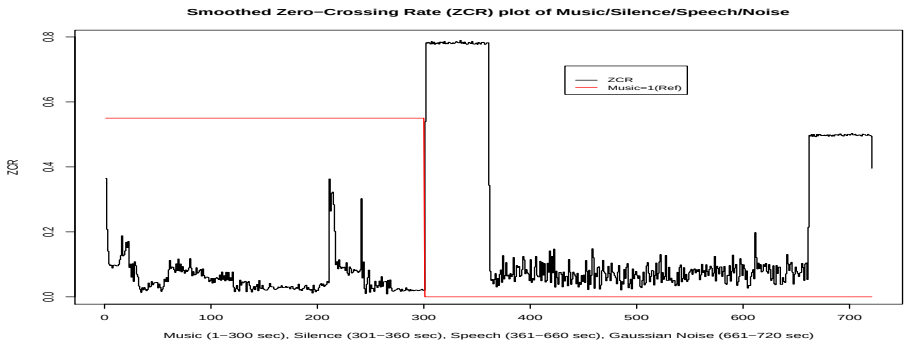

Figure 1: ZCR score of Music/Silence/Speech/Noise in 1sec unit

2.2 Root Mean Square

ZCR demonstrates the characteristics of dominant frequency in audio sources. In addition to frequency properties, signal energy is another important feature for audio analysis. Root mean square (RMS) [5] [4] and short time energy (STE) [1] are popular metrics in literature. Both of them assess signal energy via the sum of square over signal amplitude, except that RMS is the square root of STE. I take RMS as the energy measure in this section.

$$RMS = \sqrt{\Sigma_{n=1}^{N} x^2(n)} \tag{2}$$

$$STE = \Sigma_{n=1}^{N} x^2(n) \tag{3}$$

In HTK book [7], Section 5.8, signal energy is defined in log form (Equation 4, where s_n is speech samples). Signal energy **E** is an alternative feature for RMS in this study, and it will be tested in later experiments.

$$E = \log \Sigma_{n=1}^{N} s_n^2 \tag{4}$$

A sample of RMS values for music, silence, speech and noise is shown in Figure 2.

Since the vocalization boundaries are extracted from AMI reference, I expect the ZCR/RMS scores represent characters of a speaker in certain stage of a meeting. The effect of acoustic features to topic segmentation may be discovered after classification experiments.

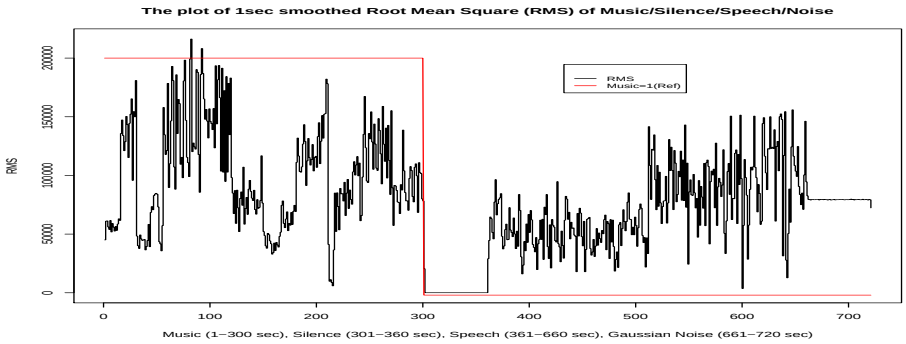

Figure 2: RMS score of Music/Silence/Speech/Noise in 1sec unit

2.3 Normalized ZCR variance and normalized RMS variance

Normalized RMS variance is proposed to indicate energy variation which may be caused by pauses or rhythm change in source [4]. Normalized RMS variance is defined as the ratio of the RMS variance to the square of RMS mean:

$$nRV = \frac{\sigma^2}{\mu^2} \tag{5}$$

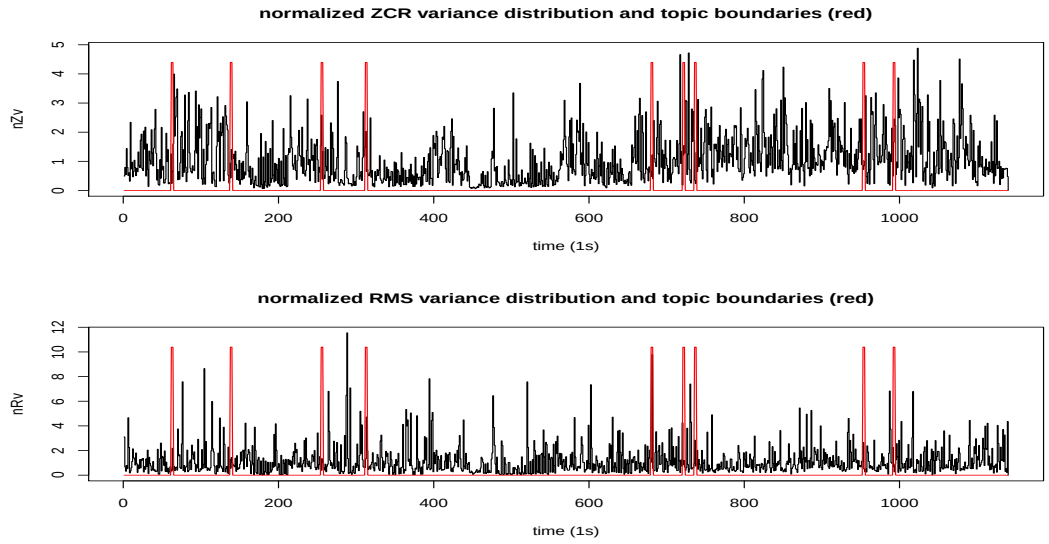

Figure 3: nZV and nRV scores of meeting ES2003a in 1sec unit, together with topic boundaries

In this study, each RMS score is generated on a 10ms frame. In order to reserve RMS variation, I take 100 frames (1 second) as a segment for nRV calculation.

Normalized ZCR variance (nZV) is defined as the ratio of the ZCR variance to the square of ZCR mean. nZV is used to indicate the dominant frequency change along speech. A plot of nZV and nRV distribution of meetng ES2003a is in Figure 3. The red spikes correspond to topic boundary (3 seconds each in current setting). Along the figures, there are no observable significant patterns around topic boundaries, or along each topic. However, I wish to judge the relevance of nRV, nZV with topic boundary in a classification model.

2.4 Null Zero-Crossing and Null Root Mean Square

A typical feature of speech recordings is that, there are always long or short pauses. The duration and frequency of pauses may be related to the content of speech. Therefore, we need a metric to reliably indicate the presence of pauses out of voice. As shown in Figure 1, a long monotone pause is clearly different from speech (60 seconds), indicated by the mean ZCR scores. However the short pauses during speech are not easily identified. For classification purpose, the mean ZCR of an instance (1 second) is the mean of 100 ZCR samples. If a short pause occupies part of two adjacent instances, the mean ZCR score of each instance may not distinguish from the ZCR of neighbouring instances of voice. Similar problems are found in mean RMS either (Figure 2). The mean RMS of short pauses hardly approaches zero, while long silence does. In order to amplify the difference between voice signal and short pauses, I propose a **frequency** measure of ZCR and RMS: Null Zero-Crossing (ZCR0, Equation 6) and Null Root Mean Square (RMS0, Equation 7). The threshold of ZCR0 is empirically set as 10, and the threshold of RMS0 is 1000.

$$ZCR0 = \frac{1}{N}\Sigma_{i=1}^{N}\Gamma(ZCR_i), \quad where\ \ \Gamma(ZCR_i) = \begin{cases} 1 & \text{if} \quad ZCR_i > Th \\ 0 & \text{if} \quad ZCR_i < Th \end{cases} \tag{6}$$

$$RMS0 = \frac{1}{N}\Sigma_{i=1}^{N}\Gamma(RMS_i), \quad where\ \ \Gamma(RMS_i) = \begin{cases} 1 & \text{if} \quad RMS_i < Th \\ 0 & \text{if} \quad RMS_i > Th \end{cases} \tag{7}$$

Panagiotakis [4] indicates that in a silent interval the number of zero-crossings is null, but in AMI speech samples I find that short pauses do not match null ZCs. They generally present higher ZCR scores than voice. In Figure 4, most ZCR scores in a 10ms frame exceed 20 and very few samples approach 0. Consequently, a ZCR score higher than a threshold indicates

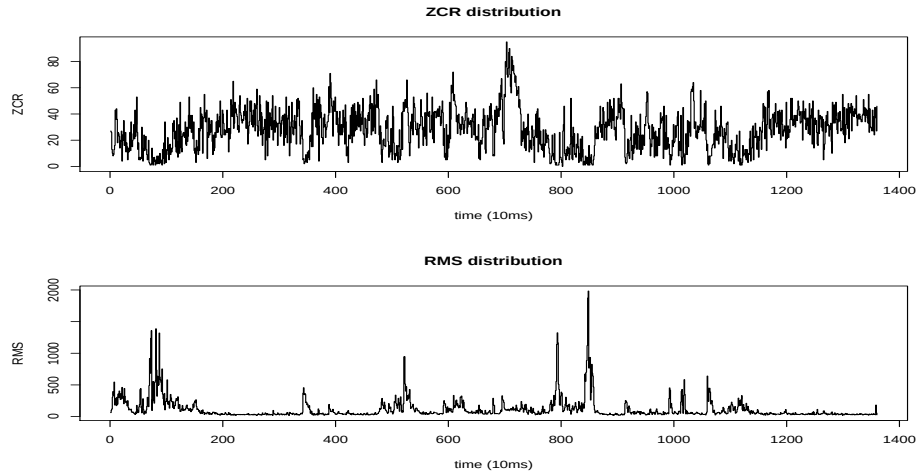

Figure 4: ZCR and RMS scores of a clip of "silence" in meeting ES2002a (from 12:20 to 12:33)

"silence", and ZCR0 counts the frequency of silence. On the contrary, RMS0 counts the silence frequency where RMS is below a threshold.

Now we can compare the differece between silence (Figure 5), male voice (Figure 6) and female voice (Figure 7) under ZCR0 and RMS0 metrics. For ZCR0, silence clip is always higher than 0.6, male voice varies from 0.3 to 0.6, and female voice is lower than 0.3. For RMS0, silence is always 1, male voice varies from 0.1 to 0.7, and female voice varies from 0.05 to 0.4. ZCR0 and RMS0 exhibit discriminating power between speech and non-speech, as well as speaker segmentation.

Feature set	horizon	Models	Pk	WD	Omega
ZCR0, RMS0	-	PT-NB	0.417	0.472	47.8/56.2
ZCR0, RMS0	1	PT-NB	**0.406**	**0.445**	**32.4/56.2**
ZCR0, RMS0	2	PT-NB	0.412	0.461	37.6/56.2
ZCR0, RMS0	3	PT-NB	0.422	0.465	32.4/56.2

Table 1: Feature Selection through ZCR0, RMS0 and horizons

3 Topic Segmentation Experiment with Unified Instances

In report v2, I explored the ways of incorporating mean ZCR and mean RMS into vocalization instances (Section 3.1), the segmentation accuracy is worse than that from original vocalization features. Afterwards, nRV and

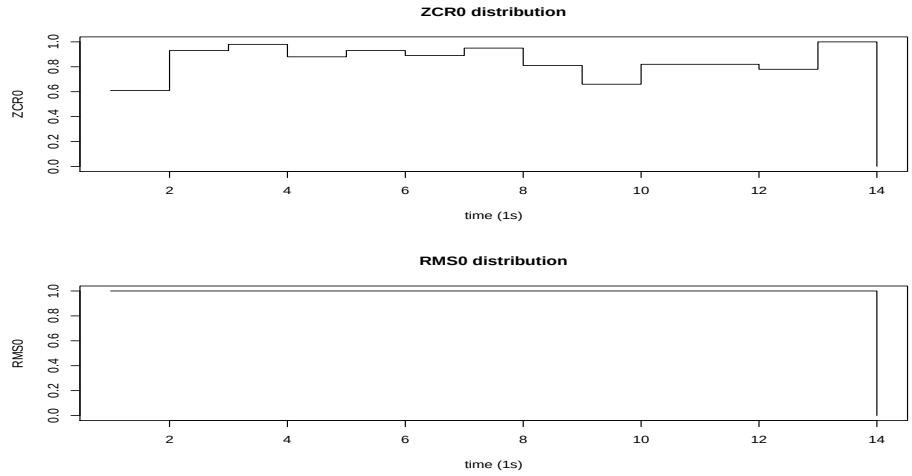

Figure 5: ZCR0 and RMS0 scores of a clip of "silence" in meeting ES2002a (from 12:20 to 12:33)

Feature set	Models	Pk	WD	Omega
ZCR	PT-NB	0.437	0.496	50.4/56.2
RMS	PT-NB	0.467	0.514	41/56.2
ZCR0	PT-NB	0.421	0.478	48.8/56.2
RMS0	PT-NB	0.529	0.788	364.4/56.2
ZCR, RMS	MAP-NB	0.411	0.416	1.4/56.2
ZCR, RMS	PT-NB	0.43	0.487	50/56.2
nZV, nRV	MAP-NB	0.459	0.517	56.8/56.2
nZV, nRV	PT-NB	0.457	0.535	55.4/56.2
ZCR0, RMS0	MAP-NB	0.406	0.411	0.0/56.2
ZCR0, RMS0	PT-NB	**0.417**	**0.472**	**47.8/56.2**
ZCR, RMS, ZCR0, RMS0	MAP-NB	0.419	0.432	10/56.2
ZCR, RMS, ZCR0, RMS0	PT-NB	0.423	0.478	47.2/56.2
ZCR, RMS, nZV, nRV	MAP-NB	0.463	0.531	64.4/56.2
ZCR, RMS, nZV, nRV	PT-NB	0.454	0.532	55.2/56.2
ZCR, RMS, nZV, nRV, ZCR0, RMS0	MAP-NB	0.445	0.506	55.2/56.2
ZCR, RMS, nZV, nRV, ZCR0, RMS0	PT-NB	0.459	0.527	54/56.2

Table 2: Feature Selection through ZCR, RMS, nZV, nRV, ZCR0 and RMS0

nZV are evaluated within unified (1 second) instances (Section 3.2). In this report, more comprehensive experiments are illustrated, all of which use unified instances. A concatenated dataset of 21 AMI meeting are in use. The total duration is 10 hour and 47 minutes. The first second of a topic is labeled as a positive instance, and all others are negative as topic boundary.

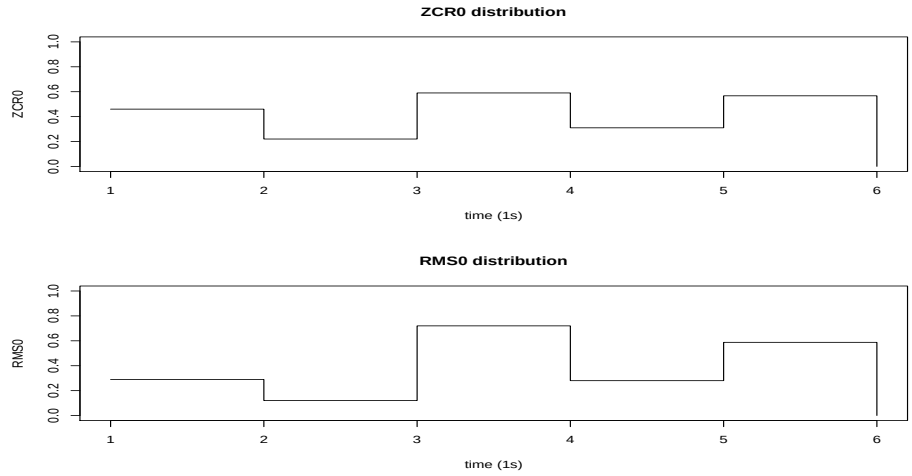

Figure 6: ZCR0 and RMS0 scores of a clip of male voice in meeting ES2002a (from 12:59 to 13:05)

Totally there are 281 topic boundaries (0.72%).

Naive Bayes classifier and proportional threshold naive Bayes classifier are selected to predict on this highly unbalanced data set. We want to know which acoustic features are the most influential factors to topic boundary classification. However, a classification model with better classification accuracy does not comply with a better segmentation accuracy. Therefore automatic feature selection is not used, I manually select feature groups and see the combination of ZCR0 and RMS0 performs good (Table 2). Inspired by the concept of vocalization horizon, I tested the horizon of acoustic features (Table 1), which improves segmentation accuracy. A comparison of topic boundary prediction sequences between ZCR0, RMS0 and $horizon = 1$ is in Figure 8 and 9.

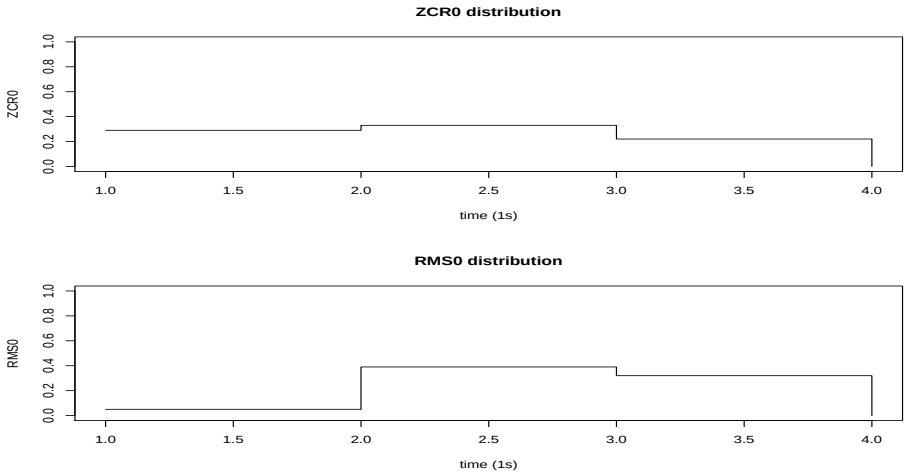

Figure 7: ZCR0 and RMS0 scores of a clip of female voice in meeting ES2002a (from 14:54 to 14:58)

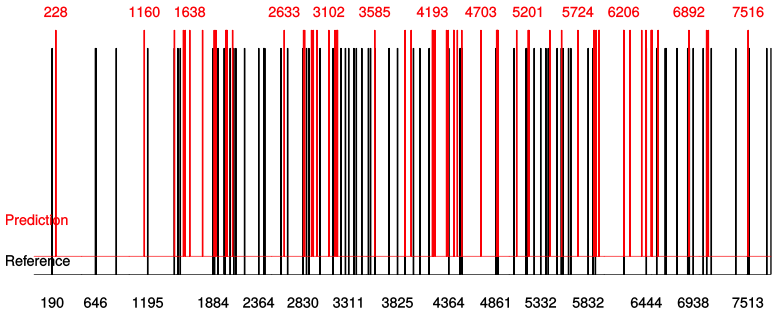

Figure 8: The fold 1 out of 5 of ZCR0 and RMS0 model prediction

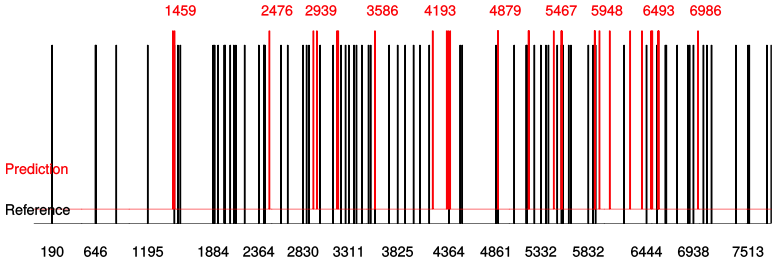

Figure 9: The fold 1 out of 5 of ZCR0, RMS0 and horizon=1 model prediction

References

[1] K. W. Jørgensen and L. L. Mølgaard. Tools for automatic audio indexing. Master's thesis, Informatics and Mathematical Modelling, Techni-

cal University of Denmark, DTU, 2006. Supervised by Prof. Lars Kai Hansen, IMM.

[2] Saturnino Luz. Locating case discussion segments in recorded medical team meetings. In *Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech (SSCS'09)*, pages 21–30, Beijing, China, October 2009. ACM Press.

[3] Saturnino Luz. The non-verbal structure of patient case discussions in multidisciplinary medical team meetings. *ACM Transactions on Information Systems*, 30(3):17:1–17:24, 2012.

[4] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on Multimedia*, 7:155–166, 2004.

[5] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE, ICASSP*, pages 1331–1334, 1997.

[6] Jing Su and Saturnino Luz. Can time dependencies and ensemble classification improve content-free dialogue segmentation? In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pages 183–188. IEEE, 2013.

[7] S. Young. *The HTK Book Version 3.4*. Cambridge University Engineering Department, 2007.