

SPEECH SEGMENTATION WITHOUT SPEECH RECOGNITION¹

Dong Wang[†], Lie Lu[‡], Hong-Jiang Zhang[‡]

[†]Department of Electronic Engineering, Tsinghua University, Beijing
wdong01@mails.tsinghua.edu.cn,

[‡]Microsoft Research Asia
{llu, hjzhang}@microsoft.com

ABSTRACT

In this paper, we presented a semantic speech segmentation approach, in particular sentence segmentation, without speech recognition. In order to get phoneme level information without word recognition information, a novel vowel/consonant/pause (V/C/P) classification is proposed. An adaptive pause detection method is also presented to adapt to various background and environment. Three feature sets, which include pause, rate of speech and prosody, are used to discriminate the sentence boundary. Experiments on broadcasting news indicate that the performance of proposed algorithm is satisfying.

1. INTRODUCTION

Sentence segmentation is very helpful in many applications, such as speech summarization [1], video summarization [9], speech document indexing and retrieval [2]. Previous researches on sentence segmentation were mostly aim at exploring the written text structure. With the fast growing needs for semantic audio content analysis, the sentence segmentation for acoustic data has been studied by many researchers recently in different applications. Shriberg [3] presented a prosody-based sentence segmentation algorithm combined with an N-gram language model. Based on this work, Shriberg [4] extended the algorithm to the scenario of multi-party meetings. Zechner [5] also applied sentence segmentation techniques to generate a summary of spoken dialogue. However, these methods rely heavily on the speech recognition results, which give the cues on both phoneme level and language level context information.

In applications such as video editing and video summarization [9], we care more about the sentence boundary than the speech content. Furthermore, speech recognition often takes too much time and the result is not reliable, especially in noisy environment, including video scenario. So it is necessary to segment the audio stream at sentence level without speech recognition. Unlike the sentence segmentation task with written text or recognized words, there is no punctuation, or language information indicating sentence boundary in acoustic spoken

data. Thus, some low level features should be extracted to discriminate the sentence boundaries.

S. Pfeiffer [6] proposed audio segmentation system without word information, where only pause feature is used to help segment the audio content at different semantic level, including sentence segmentation. However, the paper did not present an efficient method for automatic pause detection in various contexts. Furthermore, only pause is not enough for practical segmentation usage.

In this paper, a novel method is proposed for semantic speech segmentation independent of word recognition. Besides pause duration, some new features, which based on phoneme duration, Rate of Speech (ROS) and prosody, are also used to identify the speech boundaries. In order to get phoneme information without speech recognition, a novel adaptive vowel/consonant/pause (V/C/P) classification method is also proposed, which is robust and efficient in various environments.

The rest of paper is organized as follows. The overview of our system is described in Section 2. Section 3 discusses our approach to feature extraction. Section 4 gives the experimental results. And the conclusion is drawn in Section 5.

2. SYSTEM OVERVIEW

The proposed sentence segmentation system is composed of three stages, as shown in Figure 1.

In the first stage, basic features are extracted. The input audio is segmented into 20ms-long non-overlapping frames, where frame features, including frame energy, zero-crossing rate (ZCR) and pitch value, are calculated. In order to group the frames into V/C/P in phoneme level, an adaptive background noise level detection algorithm is proposed. The sentence boundary candidates are detected if the estimated pauses are long enough.

In the second stage, three feature sets, including pause features, ROS and prosodic features are extracted and combined to represent the context of a sentence boundary candidate.

In the third phase, a statistical method, AdaBoost [10], is used to detect the true sentence boundary from the candidates based on its context feature.

¹ The work was performed at Microsoft Research Asia

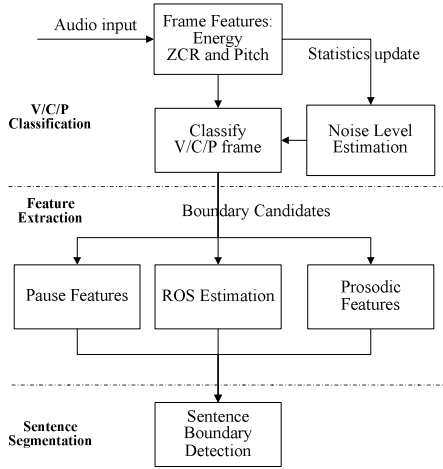


Fig. 1. An illustration of sentence segmentation system

3. FEATURE SELECTION

As mentioned above, three feature sets, including pause features, ROS and prosodic features are used to discriminate sentence boundary. Obviously, pause is one of the most important indicators of sentence boundary detection. However, pause duration between sentences is affected by the speaking rate. For example, fast-spoken sentences always have a relatively short pause. Thus, ROS is also considered in sentence detection to compensate for various speaking rate. Meanwhile, prosody shows its significant variance across the sentence boundary, as indicated in some studies [7]. Therefore, some prosodic features are also derived in our approach.

Figure 2 listed the overall feature sets used to describe the context characteristics of the sentence boundary candidate. In Figure 2, the pause features of the sentence boundary candidate are used and combined with the ROS features; and the prosodic features of vowels neighboring the candidate are also used to reflect the prosodic change across the candidate.

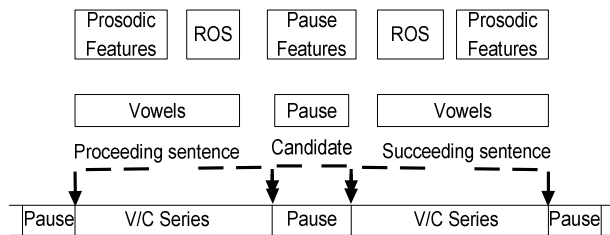


Fig. 2. The feature set used for sentence boundary candidate

In order to detect the sentence boundary candidate and extract its corresponding features, V/C/P classification is needed as a basic step. To discriminate pause, an adaptive background noise level estimation is proposed to adapt to various background. A certain energy threshold is not suitable for pause detection in such case since the environment is not constant but changeable, especially in video scenario.

3.1. Adaptive Background Noise Level Estimation and V/C/P Classification

As mentioned above, adaptive background noise level estimation is necessary for V/C/P classification. In this section, adaptive background noise level estimation and V/C/P classification is described. Fig. 3 illustrates the detailed process, and Fig. 4 shows the detailed algorithm.

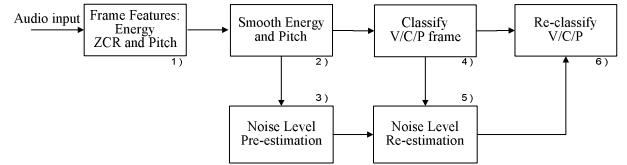


Fig. 3. Noise level estimation and V/C/P classification

1. Audio data is segmented into 20ms-long non-overlapping frames, where features, including ZCR, Energy and Pitch, are extracted.
2. Energy and pitch curve is smoothed.
3. The **Mean_En** and **Std_En** of energy curve are calculated to coarsely estimate the background noise energy level, as:

$$\text{NoiseLevel} = \text{Mean_En} - 0.75 \text{ Std_En}.$$
 Similarly the threshold of ZCR (**ZCR_dyna**) is defined as:

$$\text{ZCR_dyna} = \text{Mean_ZCR} + 0.5 \text{ Std_ZCR}.$$
4. Frames are classified as V/C/P coarsely by using the following rules, where **FrameType** is used to denote the type of each frame.
 If $\text{ZCR} > \text{ZCR_dyna}$ then **FrameType** = Consonant
 Else if $\text{Energy} < \text{NoiseLevel}$, then **FrameType** = Pause
 Else **FrameType** = Vowel
5. Update the **NoiseLevel** as the weighted average energy of the frames at each vowel boundary and the background segments.
6. Re-classify the frames using algorithm in step 4 with the updated **NoiseLevel**. Pauses are merged by removing isolated short consonants. Vowel will be split at its energy valley if its duration is too long

Fig. 4 the detail algorithm on V/C/P classification

In real implementation, the noise level estimation and V/C/P classification is performed every 5 seconds, in order to catch the environmental changes.

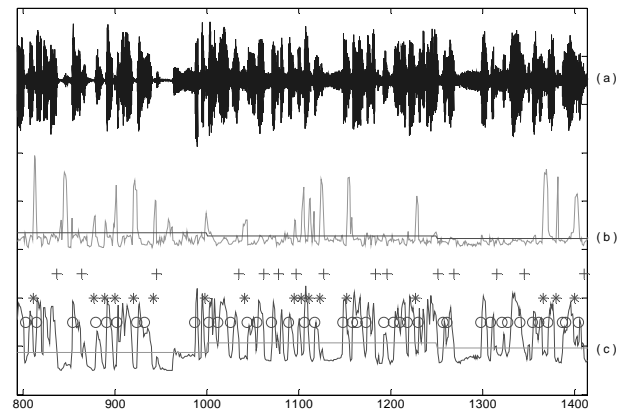


Fig. 5. V/C/P classification of an example speech clip with transition from quiet background to music background

Based this algorithm, we can get satisfactory V/C/P classification result in various background. Fig. 5 illustrates an example of V/C/P classification result on an audio clip with environment transition from quiet background to music background. In Fig. 5, (a) shows the original speech data with background transition; (b) is the ZCR curve with its dynamic threshold; and (c) is the energy curve with estimated noise level. Obviously, the estimated noise level is adaptive to background transition and suitable to be a threshold for pause detection. In the figure, '+', 'o' and '*' represent the start of pause, consonant and vowel respectively. It can be seen that V/C/P is satisfactorily segmented. If the estimated pause is long enough (in the real implementation, it is set as 5 frames), it is considered as a sentence boundary candidate. Pause/ROS features and prosodic features are then extracted to represent the characteristics of the candidate and its context.

3.2. ROS

ROS often has two main definitions based on Word/Minute (WPM) and Syllable/Second (SPS) respectively [8]. The latter one is chosen in our approach since we have no word information and it is more suitable for short-time measurement. In real implementation, we take vowel as syllable. Based on the above V/C/P classification, ROS is calculated as

$$ROS = \frac{n}{\sum d_i} \quad (1)$$

where n is the vowel count, and d_i is the i -th vowel duration. Pause durations and consonant durations are excluded in the ROS calculation.

ROS features are combined with the pause features for better performance, which is shown in the next sub-section.

3.3 Pause and ROS Features

Pause is the most basic feature for sentence segmentation. Considering this, the following pause features are extracted to represent characteristics of the sentence boundary candidate, as the Table I shows.

Table I. Pause and ROS Feature list

| Feature | Description |
|----------------------------|--|
| <i>En_Std, En_Mean</i> | The variance and mean of the energy contour |
| <i>ROS1</i> | ROS of the proceeding sentence |
| <i>ROS2</i> | ROS of the succeeding sentence |
| <i>duration</i> | pause duration of the candidate |
| <i>duration_n1</i> | duration normalized by ROS1 |
| <i>duration_n2</i> | duration normalized by ROS2 |
| <i>preceed_pause1, 2</i> | Two pause durations proceeding candidate |
| <i>succeed_pause1, 2</i> | Two pause durations succeeding candidate |
| <i>preceed_nonpause1,2</i> | Two non-pause durations proceeding the candidate |
| <i>preceed_nonpause1,2</i> | Two non-pause durations succeeding the candidate |

The feature set concerns and combines pause and ROS. Pause duration is normalized by the ROS of previous sentence and next sentence respectively. The proceeding pause/non-pause durations and the succeeding pause/non-pause durations are also

considered to depict completely the characteristics of the pause context of the sentence boundary candidate.

3.4. Prosodic Features

In general, prosody has a big variance across the sentence boundary. For example, for most of declarative sentences, the prosody curve always decline to low at the end of sentence, and rise to high at the beginning of the sentence. Based on these facts, three vowels proceeding and three vowels succeeding the sentence boundary candidate is selected to represent the prosody context of the candidate. The features extracted from each vowel are as Table II shows:

Table II. Prosodic Feature list for each vowel

| Feature | Description |
|------------------------------|---|
| <i>En_Std, En_Mean</i> | The variance and mean of the energy contour |
| <i>Pitch_std, Pitch_mean</i> | The variance and mean of the pitch contour |
| <i>Onset Pitch</i> | pitch at the begin of the vowel |
| <i>Offset Pitch</i> | pitch at the end of the vowel |
| <i>Min Pitch</i> | minimum pitch |
| <i>Max Pitch</i> | maximum pitch |
| <i>Pitch_Range</i> | range of pitch |
| <i>Sign[0]</i> | sign[0] = onset - min; |
| <i>Sign[1]</i> | sign[1] = onset - max; |
| <i>Sign[2]</i> | sign[2] = offset - min; |
| <i>Sign[3]</i> | sign[3] = offset - max; |
| <i>Sum-descent</i> | sum of all downward pitch changes |
| <i>Sum-ascent</i> | sum of all upward pitch changes |

In prosodic feature, the variance and mean, the minimum and maximum, the onset and offset of pitch curve are used to depict the pitch statistics of the vowels. Meanwhile, four *Signs* are used to describe the pitch shape of vowels. Different pitch shapes have different sign array with positive and negative values. Fig. 6 shows five basic pitch shapes of a vowel.

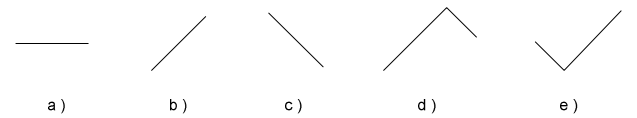


Fig. 6. Illustration of five basic vowel pitch shapes

4. EXPERIMENTAL RESULTS

In this section, we describe the experimental material and then the experiment results are presented for evaluation.

4.1. Sentence Boundary Detection Scheme

For each sentence candidate, features mentioned above are extracted. Then, a statistical method, AdaBoost [10], is used to discriminate the true boundaries from false ones. AdaBoost is an adaptive algorithm to boost a sequence of weak classifiers, where the weights are updated dynamically according to the errors in previous learning. It had good classification performance, which is proved in many fields.

4.2. Corpus for Experiments

The evaluation of the proposed sentence detection method is performed on Hub-4 1997 English Broadcast News Database, which mainly consists of conversational talk, and is from different radios, such as CNN, ABC, CRI and C-SPAN. Such corpus is used since it comprises speech of many different speakers and diversified environment. It is a good robustness test for our U/V/P classification and sentence segmentation approach.

In real implementation, only two kinds of radio, CNN Early Prime and CNN Prime News, are selected. About 20 hours of data is used, 80% for training and 20% for testing. The hand labeled sentence boundary are used as ground truth. The full stop and question mark are used as indicators of sentence boundary.

4.3. Experimental Results

Since pause duration is the basic discriminator for the sentence boundary, we first implemented a baseline system which uses only pause features. The performance is shown in the Table III, where Boundary means the candidates which are true sentence boundary; and non-boundary is on the contrary.

Table III Sentence detection results based on pause features (unit: 100%)

| | Total Number | Discriminate Results | |
|--------------|--------------|----------------------|--------------|
| | | Boundary | Non-boundary |
| Boundary | 100 | 79.8 | 20.2 |
| Non-boundary | 100 | 24.2 | 75.8 |

From the Table III, it can be seen that pause features can perform well. About 79.8 boundaries are discriminated correctly. It also proves our V/C/P classification has a good performance.

When prosodic features are introduced, the performance increases a lot, as the Table IV shows. In the Table IV, the number in parenthesis means the corresponding performance increment. It can be seen that the boundary accuracy increase 2.2%, while non-boundary 6.8%.

Table IV Sentence detection results based on pause and prosody features (unit: 100%)

| | Total Number | Discriminate Results | |
|--------------|--------------|----------------------|--------------|
| | | Boundary | Non-boundary |
| Boundary | 100 | 82.0 (+2.2) | 18.0 |
| Non-boundary | 100 | 17.4 | 82.6 (+6.8) |

In the previous experiment, all prosodic features are extracted from independent vowels. However, the variances between these vowels are not considered, which could also give us a good cue to identify sentence boundary. Therefore, we add delta prosodic feature between corresponding features of every two vowels, the final performance is listed in Table V.

Table V Sentence detection results based on pause, prosody and delta prosodic features (unit: 100%)

| | Total Number | Discriminate Results | |
|--------------|--------------|----------------------|--------------|
| | | Boundary | Non-boundary |
| Boundary | 100 | 82.3 (+0.3) | 17.7 |
| Non-boundary | 100 | 13.3 | 86.7 (+4.1) |

The final accuracy achieves 82.3% for boundary and 86.7% for non-boundary. These experiment results show that our proposed approach can achieve satisfying accuracy. It is comparable to the state-of-the-art performance on sentence segmentation which used word recognition and linguistic model [3], whose error rate is 10.8% with word alignment as a known prior on Broadcast News.

5. CONCLUSION

Unlike the traditional sentence segmentation which uses speech recognition result, in this paper, we present a novel algorithm based on vowel/consonant/pause classification result and some low-level feature sets. In our approach, an automatic pause threshold detection approach is proposed for adaptive V/C/P classification; features on pause, ROS and prosody are also considered to identify sentence boundary. Experiments evaluate the proposed approach and each feature set. The result is satisfying. It achieves comparable accuracy as the method using speech recognition but with lower computational cost.

6. ACKNOWLEDGEMENT

Thanks Lei Zhang from Microsoft Research Asia for providing the code on AdaBoost training and testing.

7. REFERENCES

- [1] K. Zechner, "Summarization of Spoken Language – Challenges, Methods, and Prospects" *Technology Expert eZine*, Issue 6, January 2002.
- [2] K. Koumpis and S. Renals, "The role of Prosody in a Voicemail Summarization System" *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*
- [3] E. Shriberg, A. Stolcke, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Communication* 32 (1-2), September 2000
- [4] E. Shriberg, A. Stolcke, D. Baron, "Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Disfluencies, and Overlapping Speech" *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pp. 139-146, Red Bank, NJ.
- [5] K. Zechner, "Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains", *SIGIR'01*, September, 2001, New Orleans, Louisiana, USA.
- [6] S. Pfeiffer, "Pause Concepts for audio Segmentation at Different Semantic Levels", *ACM Multimedia 2001*, pp. 187-193
- [7] R. Kompe, *Prosody in Speech Understanding Systems*, Springer-Verlag, 1996.
- [8] N. Morgan and E. Fosler, "Combining Multiple Estimators of Speaking Rate," *Proc. ICASSP '98*, Seattle. pp. 729-732, May 1998
- [9] Y.-F. Ma, L. Lu, H.-J. Zhang and M. J. Li. "An Attention Model for Video Summarization", *10th ACM International Conference on Multimedia 2002*.
- [10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting". *J. Comp. & Sys. Sci.*, 55(1), pp.119-139, 1997.