Soundbite Identification Using Reference and Automatic Transcripts of Broadcast News Speech

Feifan Liu and Yang Liu

The University of Texas at Dallas, USA

ABSTRACT

Soundbite identification in broadcast news is important for locating information useful for question answering, mining opinions of a particular person, and enriching speech recognition output with quotation marks. This paper presents a systematic study of this problem under a classification framework, including problem formulation for classification, feature extraction, and the effect of using automatic speech recognition (ASR) output and automatic sentence boundary detection. Our experiments on a Mandarin broadcast news speech corpus show that the three-way classification framework outperforms the binary classification. The entropy-based feature weighting method generally performs better than others. Using ASR output degrades system performance, with more degradation observed from using automatic sentence segmentation than speech recognition errors for this task, especially on the recall rate.

Index Terms—soundbite identification, term weighting, text classification, sentence boundary detection

1. INTRODUCTION

With the increasing amount of broadcast news (BN) speech, it is important to develop automatic processing techniques to effectively access these data for applications such as automatic summarization, visual browsing, speech retrieval, and question answering. Most newscasts contain interview quotations or speech clips from speakers other than anchors and reporters. These are called soundbites [1]. Obviously, identifying soundbites, together with their corresponding speaker names, would be very helpful to mine opinions from particular speakers. This is also needed for rich transcription of speech, where speech recognition output is enriched with punctuation marks (soundbite speech corresponds to quotation marks) and speaker names.

In this paper, we employ a classification framework for soundbite identification, where for a given speaker turn segment in the transcripts, the task is to determine whether it is a soundbite. We aim to address the following questions for this task. (1) What is a better problem formulation for this task, using binary classification (soundbite versus not) or multi-way classification for different speaker roles (e.g., anchor, report, soundbite)? (2) What are useful features and effective weighting methods (such as TF-IDF, entropy-based) for this classification task? (3) What is the effect of speech recognition errors and automatic sentence boundary detection on the system performance?

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the details of the components in the classification framework for soundbite identification. We describe our experiments and results in Section 4. Conclusions and future work appear in Section 5.

2. RELATED WORK

The most related work to our task so far is [1], which used a model based on conditional random fields for soundbite detection on English broadcast news and reported an accuracy of 67.4%, higher than the baseline of 46.5%. In this approach, lexical, acoustic/prosodic, and structural features at the turn level were used. However, the results were only shown for the ASR transcripts, therefore the effect of using ASR compared to the reference transcripts is unclear, which we will investigate in this paper.

Another line of work related to our task is speaker role identification, for example, [2, 3], both of which assign one of the three types of speaker roles (anchor, reporter/journalist, other/guest speaker) to each speech turn. Barzilay et al. [2] applied BoosTexter, a boosting algorithm, and a maximum entropy model for this task in an English BN corpus, obtaining the accuracy of 80.5% using reference transcripts and 77% on ASR transcripts, compared to the chance of 35.4%. Liu [3] combined a generative HMM approach with the conditional maximum entropy method in a Mandarin BN corpus, reporting a classification accuracy of 81.97% using reference transcripts against the baseline of around 50%. The category "other/guest speaker" in those studies corresponds to the soundbites used in our experiment. Our task is slightly different from [2, 3] in that we are only focusing on soundbites and do not need to distinguish the role of anchors and reporters.

Soundbite identification is also related to speaker diarization [4], which aims to find speaker changes, group the same speakers together, and recognize speaker names. It is an important component for rich transcription (e.g., in the

DARPA EARS program). Most of the early work in this area only focused on speaker segmentation and clustering, not involving speaker names or their roles. Recent studies (such as [5]) also made an attempt to add speaker names for BN speech.

3. CLASSIFICATION FRAMEWORK FOR SOUNDBITE IDENTIFICATION

We tackle the soundbite identification problem using a general classification framework as shown in Figure 1. The components in it are described in detail as follows.

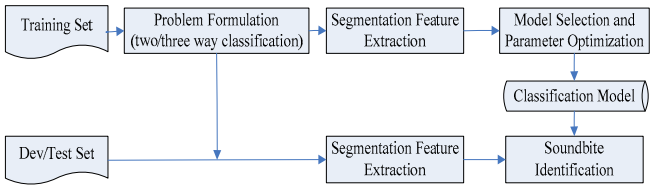

Figure 1. General Classification Framework for Soundbite Identification

3.1. Problem Formulation

Generally there are three types of speakers in broadcast news shows [2, 3], therefore for soundbite identification, we can formulate the problem using either a binary classification task — soundbite versus not, or a three-way classification task — identify whether a speech segment is from an anchor, reporter, or a sounbite. The former may reduce some inter-class noise and the latter may benefit from the more discriminative features among different classes. We will compare which of the two schemes is more effective for soundbite identification. Note that we are assuming we have the additional annotation of anchor and reporter roles in the training set. If this is not available, the three-way classification setup does not apply.

3.2. Feature Extraction

This module is designed to represent each speech turn of BN as a feature vector for classification. The features we exploit here are similar to those used in [3], including traditional lexical features as well as contextual relations between speech turns. In addition, we incorporate some length-based features.

Lexical Features (LF):

- LF-1: Unigram and bigram features in the first and the last sentence of the current speech turn. We hypothesize that the lexical cues in those sentences are more indicative to the speaker roles, as noted in [3]. Similar features are used in the previous work [1, 6, 7], but human inspection was needed to determine some special cue phrases.
- LF-2: Unigram and bigram features from the last sentence of the previous turn and from the first sentence of the following turn. We expect these to reflect some functional transition among different speakers and thus be able to model the inter-dependency relations among neighboring speech turns.

Structural Features (SF):

- Number of words in the current speech turn. Different speakers have different speaking styles as observed in [2]. Typically, the soundbite turn consists of fewer words than the turn of anchors and reporters.
- Number of sentences in the current speech turn. We expect that the different levels of length information may be complementary.
- Average number of words in each sentence in the current speech turn. Our hypothesis is that professional speakers, such as anchors and reporters, often read teleprompts or tend to use longer and more complex sentences, whereas speech in soundbites may be more spontaneous and more likely to contain shorter sentences.

3.3. Feature Weighting

Using the features introduced above, each speech turn is represented as a feature vector for classification. In text categorization and information retrieval, term weighting has been extensively studied [8, 9]. In this paper, we evaluate four different feature weighting methods for soundbite identification. The weighting is performed for different features instead of terms, that is, we distinguish the N-gram lexical features for different categories (LF-1 and LF-2 in Section 3.2), even though the terms might be the same.

The notations we use for the description of feature weighting are as follows. N is the number of speech turns in the training collection, M is the total number of features, f_{ik} is the frequency of feature φ_i in the k^{th} speech turn, n_i denotes the number of the speech turns containing feature φ_i, $F(\varphi_i)$ means the frequency of feature φ_i in the collection, and w_{ik} is the weight assigned to the feature φ_i in the k^{th} turn using different approaches.

3.3.1 Frequency Weighting

This is simply the frequency of the feature:

$$w_{ik} = f_{ik} \quad (1)$$

*3.3.2 Tf*idf (inverse document frequency) Weighting*

This weighting method was originally proposed and applied to document retrieval task [8, 10]. A feature's IDF

value $\log(N/n_i)$ represents its specificity — whether it is an indicative feature for a particular segment or it occurs in many segments.

$$w_{ik} = f_{ik} * \log(N/n_i) \quad (2)$$

*3.3.3 Tf*iwf (inverse word frequency) Weighting*

Similar to the idea of idf, Basili et al. [11] used inverse word frequency (iwf). Both idf and iwf can penalize high-frequency terms. In this paper, we used the following formula:

$$w_{ik} = f_{ik} * \log(\sum_{j=1}^{M} F(\varphi_j)/F(\varphi_i)) \quad (3)$$

3.3.4 Entropy Weighting

Entropy-based weighting has been shown to be the most effective weighting approach in comparison with others [12]. This method assigns different weights to features via the following equation:

$$w_{ik} = \log(f_{ik} + 1.0) * [1 - entropy(\varphi_i)] \quad (4)$$

where

$$entropy(\varphi_i) = -\frac{1}{\log N}\sum_{j=1}^{N}\left[\left(\frac{f_{ij}}{F(\varphi_i)}\right)\log\left(\frac{f_{ij}}{F(\varphi_i)}\right)\right]$$

is the average entropy of feature φ_i. An even distribution of feature φ_i across all the speech turns results in a high entropy value, which means the feature's discriminating ability is low, thus the feature will be given a small weight based on Equation (4).

3.4 Classification Models

The two models we considered in this paper are the maximum entropy (ME) and support vector machine (SVM) classifiers, which have been successfully applied to many natural language processing and speech processing tasks. In a preliminary study, we obtained comparable performance using the SVM and ME classifiers, therefore in the experiments in Section 4, we choose to only use SVMs, mainly because of its ability to better handle numeric features.

4. EXPERIMENTAL RESULTS

4.1. Experiment Setup

We use the TDT4 Mandarin broadcast news data in our experiment, which consists of 335 news shows from different sources. Speech turn boundaries and speaker role information (anchor, reporter, and other) were annotated manually in the transcripts. Speech turns labeled as "other" are considered as our reference soundbite segmentation. The punctuation marks in the LDC transcripts were used to obtain the reference sentence information for feature extraction. We randomly split the data set, and use around 1/10 of the data as the development set, another 1/10 as the test set[1], and the rest as our training set. The ASR output for the test set is from a state-of-the-art Mandarin speech recognizer [13]. We aligned the ASR output with the reference transcripts to obtain the speaker turn and role information for the ASR words. The statistics of the data used in our experiments is described in Table 1. Note that for training and testing, we ignored the small amount of speech turns in the corpus that were originally labeled with an "unknown" role. Those tags were used when the annotators could not determine the role for the speech segments.

Table 1. Statistics of our experiment data

Data Set	# of shows	# of speech turn	# of soundbites
Training	280	13301	1715
Dev	31	1382	255
Test	24	1211	114
Test_ASR	24	1189	109

We used the libSVM toolkit [14] and the RBF kernel function in our experiments. All the parameters for SVMs were optimized using 5-fold cross validation on the training set. The weighting terms used in different approaches were computed from the training set and applied to the dev and test sets.

For the soundbite identification performance measure, we use precision/recall/f-measure, as well as classification accuracy, as shown below.

$$acc = \frac{\#\ of\ correctly\ labelled\ speech\ turns}{\#\ of\ all\ speech\ turns} \quad (5)$$

$$\left.\begin{aligned} p &= \frac{\#\ of\ correctly\ identified\ soundbites}{\#\ of\ all\ identified\ soundbites} \\ r &= \frac{\#\ of\ correctly\ identified\ soundbites}{\#\ of\ all\ soundbites} \\ f &= 2 * p * r/(p+r) \end{aligned}\right\} \quad (6)$$

When using the three-way classification setup, the system hypotheses from the classifier are mapped into binary tags for evaluation by combining the other two classes (anchors and reporters) into non-soundbite. The baseline performance in our experiments is obtained by predicting all the speech turns as the majority class, i.e. non-soundbite.

4.2 Experimental Results

[1] Later we removed some shows from the test set because the ASR output was not available.

We first investigate the effect of different features and weighting approaches using the dev set, and then show the impact of using ASR output on the test set. The comparison between the binary and three-way classification setup is made for both cases.

4.2.1 Comparison of Different Weighting Methods

Table 2 shows the results on the dev set using different weighting methods described in Section 3.3. We used all the features listed in Section 3.2 for this experiment. Both the F-measure and accuracy ("Acc") results are presented in the table.

Table 2. Results using different feature weighting methods on the dev set. Precision/recall reslts are shown in the parenthesis under F-measure. Baseline accuracy is 81.5%.

Weighting	Binary		Three-way	
	F-measure	Acc	F-measure	Acc
Freq	84.3 (84.3/84.3)	94.2	87.5 (83.3/92.2)	95.2
tf*idf	88.2 (85.6/91)	95.5	87.2 (79.6/96.5)	94.8
tf*iwf	87.7 (83.6/92.2)	95.2	88.6 (84.4/93.3)	95.6
Entropy	87.2 (90.7/83.9)	95.4	89.8 (89.5/90.2)	96.2

The results show that three weighting methods using global information generally perform much better than simply using local information of the term frequency, with one exception of using tf*idf in three-way classification. Interestingly, different problem formulations (binary versus three-way) seem to prefer different weighting methods. "tf*idf" works best for binary classification while "entropy" is better for three-way classification. We observe that entropy-based weighting always leads to higher precision and relatively lower recall in comparison with tf*idf and tf*iwf. Consistent with the findings on the information retrieval task in [12], entropy-based weighting is more theoretic and seems to be a promising weighting choice for the soundbite identification task, compared to those two empirical solutions ("idf" and "iwf").

4.2.2 Contribution of Different Types of Features

We used five different feature sets in order to investigate the contribution of different types of features. "freq" based weighting is used for all the experiments. The results are presented in Table 3 for precision, recall, F-measure, and accuracy. "LF", "LF-1", and "SF" stand for those features described in Section 3.2; "Cutoff_1" indicates that only features in "LF+SF" occurring more than once in the training set are used; "Only_Uni" means that only the unigram features from "LF+SF" are used.

The results indicate that adding contextual features improves the performance for both binary and three-way classification (comparing LF-1 and LF in Table 3), suggesting that those features might capture some dependency information between neighboring speech turns. On the other hand, the length-based structural features (SF) are not useful in the binary classification configuration but do help three-way for all the performance measures (precision, recall, and accuracy). This indicates that length-based features may correlate more with rich speaker role information than just the two types (soundbite versus not).

Table 3. Results using different features on the dev set. Baseline accuracy is 81.5%.

	Features	Prec.	Rec.	F-value	Acc.
Binary	LF-1	84.96	75.29	79.84	92.98
	LF	84.37	84.71	84.54	94.28
	LF+SF	84.31	84.31	84.31	94.21
	Cutoff_1	87.45	81.96	84.62	94.50
	Only_Uni	88.11	78.43	82.99	94.07
Three-way	LF-1	80.94	88.24	84.43	93.99
	LF	82.69	91.76	86.99	94.94
	LF+SF	83.33	92.16	87.52	95.15
	Cutoff_1	87.85	85.1	86.45	95.08
	Only_Uni	90.5	85.88	88.13	95.73

We observe different patterns regarding the contribution of low-frequency and bigram features. Removing low-frequency features (i.e., Cutoff-1) helps in binary classification, but not for three-way classification; removing bigram features (Only_Uni) improves performance in three-way classification, but not for the binary setup. From the results in Table 3, it seems that using these feature selection always increases precision and decreases recall rate, resulting in mixed results in F-measure or accuracy.

We also tried to add all the unigrams and bigrams in the current speech segment to "LF+SF", not just using the first and last sentence. We found significant degradation from the 5-fold cross-validation on the training set. This suggests that indicative lexical cues for soundbite detection often occur in the first sentence and the last sentence, and that including more lexical features might introduce more noise.

4.2.3 Impact of Using ASR Output

Finally we test our classification framework on the test set, with a main focus on evaluating the effect of using ASR output on soundbite identification. Based on the above results, we choose to use the "LF+SF" feature set and entropy-based weighting approach in this experiment. Since the features we use rely on sentence information, we will also examine the impact of automatic sentence segmentation.

Table 4 shows the results on the test set. "REF" means the human transcripts and human annotated sentences.

"ASR_ASB" means using ASR output and automatic sentence segmentation results based on [15]. "ASR_RSB" is obtained by aligning reference sentence boundaries in the human transcripts to the ASR output.

Table 4. Results on the test set. Baseline accuracy is 90.83% for ASR, and 90.59% for REF.

	Test set	Prec.	Rec.	F-value	Acc.
Binary	ASR_ASB	77.36	37.61	50.62	92.19
	ASR_RSB	79.07	62.39	69.74	93.93
	REF	77.66	64.04	70.19	93.42
Three-way	ASR_ASB	68	46.79	55.44	93.10
	ASR_RSB	73.96	65.14	69.27	94.7
	REF	74.31	71.05	72.65	94.96

We observe that speech recognition errors hurt the system performance (comparing REF and ASR conditions). Using automatic sentence boundary detection degrades performance even more (comparing RSB and ASB). In particular, there is a significant decrease of the recall rate when using automatic sentence boundary hypotheses. It might be because that the error rate of the ASR output we used is quite low for this BN data, and the wrong sentence segmentation leads to misses of important cue words for soundbite speech. The results also show that because of the imbalance of the corpus (soundbite is the minority class), the precision/recall measure does not always correlate well with the classification accuracy.

Note that there are more VOA shows in the test set than in the dev set. VOA shows seem to be very different from other Mandarin BN sources in terms of structure and style, posing more problems to speech recognition and sentence boundary detection. That is partly why we obtain relatively worse performance on the test set, even though it has a higher baseline compared to the dev set.

4.2.4 Discussion

From all of the results above, we observe that using the three-way classification strategy generally outperforms the binary setup. This suggests that a multiple-way task formulation for soundbite identification can make use of more discriminating features among different speaker roles. Of course this requires additional annotation of speaker roles, which may not be available. We conducted a further analysis of the results on the dev set from the binary and three-way setup, both using the entropy based weighting method as in Table 2. The confusion matrix representing how their results differ is shown in Table 5. We find that for most of the instances, the hypotheses using the binary and three-way classification are the same (94.6% and 2.96% in Table 5). Compared to the results using binary classification, the three-way setup was able to correct the errors for 1.59% of all the instances, but introduced new errors to 0.8% of the instances. In addition, among those improved instances (1.59% of the instances), 27.3% of them are corrected from the hypothesis of soundbite to non-soundbite, and 72.7% of them changed from non-soundbite to soundbite. The errors introduced by the three-way classification (0.8% of the instances) are all due to false alarms (i.e., from non-soundbite hypothesis to soundbite). This indicates that using three-way classification tends to detect more soundbites and thus improves the recall rate, which may hurt the precision because of the false alarms. This observation is also consistent with the experimental results in Table 2.

Table 5. Confusion matrix using binary and three-way classification on the dev set using the entropy-based feature weighting method. The numbers shown are the percent of the instances in the dev set.

	Three-way correct	Three-way incorrect
Binary correct	94.6%	0.8%
Binary incorrect	1.59%	2.96%

For the soundbite identification task, the huge number of lexical features will inevitably introduce some noise. One simple solution is to compile some cue phrases manually [1, 6, 7], but it is time-consuming and labor-intensive. Barzilay et al. [2] tried to use machine learning methods to remove some unimportant features automatically, however, once a feature is removed, it will never be considered by the classification model. In this paper we evaluated a different approach — keeping all the features but assigning them different weights. Our experiments show that instead of a binary solution to either preserve or remove a feature, using weighting to smooth features achieves promising results.

The structural features we used in this paper are different from those in [1]. However, some of features used in that study, such as turn position, are indirectly incorporated in our contextual lexical features (LF-2), and other information, such as speakers, is usually not readily available. We did not use any prosodic/acoustic features in this paper, which we plan to investigate in our future work.

In the current study, we assume that the speaker turn segment is given, and thus the task is to determine whether a segment is soundbite or not. In a real scenario, the speaker segment information will come from speaker diarization, which will contain errors and may impact soundbite identification performance.

5 CONCLUSIONS AND FUTURE WORK

This paper presents a systematic study on the soundbite identification task under a classification framework. We have examined different problem formulation strategies (binary versus three-way classification), multiple weighting schemes for more discriminative feature representation, and the effects of using automatic speech recognition and automatic sentence boundary detection for this task. We found that in general using three-way classification

outperforms binary classification. Entropy-based weighting methods yield much better performance than the baseline frequency-based weighting method. Using automatic sentence boundary detection degrades the system performance even more than speech recognition errors for this task, especially causing the large degradation of recall rate.

In our future work, we will investigate sequence modeling approaches to model the dependency between the classes, and incorporate acoustic features for soundbite detection. In addition, we will integrate this work with soundbite speaker name identification to develop a unified framework for broadcast news processing.

6 ACKNOWLEDGMENTS

The authors thank Julia Hirschberg and Sameer Maskey at Columbia University for their help with data annotation. This work is supported by DARPA under Contract No. HR0011-06-C-0023. Distribution is unlimited.

7 REFERENCES

[1] S. Maskey and J. Hirschberg, "Soundbite Detection in Broadcast News Domain," *Proc. of Interspeech*, 2006.

[2] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts," *Proc. of AAAI*, 2000.

[3] Y. Liu, "Initial Study on Automatic Identification of Speaker Role in Broadcast News Speech," *Proc. of Human Language Technology Conference of the NAACL*, New York, 2006.

[4] S. E. Tranter and D. A. Reynolds, "An Overview of Automatic Speaker Diarisation Systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1557-1565, 2006.

[5] C. Ma, P. Nguyen, and M. Mahajan, "Finding Speaker Identities with a Conditional Maximum Entropy Model," *Proc. of ICASSP*, 2007.

[6] I. Mani, M. House, M. Maybury, and M. Green, "Towards Content-based Browsing of Broadcast News Video," in *Intelligent Multimedia Information Retrieval*, M. Maybury, Ed.: AAAI/MIT Press, 1997, pp. 241-258.

[7] J. Reynar, "Statistical Models for Topic Segmentation.," *Proc. of 37th Annual Meeting of the ACL*, 1999.

[8] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, pp. 513-523, 1988.

[9] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proc. of 14th International Conference on Machine Learning*, 1997.

[10] K. S. Jones, "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation*, vol. 28, pp. 132-142, 1972.

[11] R. Basili, A. Moschitti, and M. Pazienza, "A Text Classifier Based on Linguistic Processing," *Proc. of IJCAI-99, Machine Learning for Information Filtering*, 1999.

[12] S. T. Dumais, "Improving the Retrieval Information from External Sources," *Behaviour Research Methods, Instruments and Computers*, vol. 23, pp. 229-236, 1991.

[13] M. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin Broadcast News Speech Recognition," *Proc. of Interspeech*, 2006.

[14] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm," 2001.

[15] M. Zimmermann, D. H. Tur, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The ICSI+Multi-Lingual Sentence Segmentation System," *Proc. of Interspeech*, 2006.