CAPTION-AIDED SPEECH DETECTION IN VIDEOS *

Cong Li [1], *Zhijian Ou*[1], *Wei hu*[2], *Tao Wang*[2], *Yimin Zhang*[2]
[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Intel China Research Center, Beijing, China
ozj@tsinghua.edu.cn

ABSTRACT

This paper presents a novel audio-visual fusion method for speech detection, which is an important front-end for content-based video processing. This approach aims to extract homogeneous speech segments from the accompanying audio stream in real-world movie/TV videos with the help of video captions. Note that captions are mainly created to help viewers to follow the dialog, rather than to accurately locate the speech regions. We propose a caption-aided speech detection approach, which makes use of both caption information and audio information. The inaccurate positions of the captions are refined through using audio features (pitch and MFCCs) and BIC-based acoustic change detection. Comparison experiments against several other traditional speech detection approaches are conducted, showing that the proposed approach improves the speech detection performance greatly.

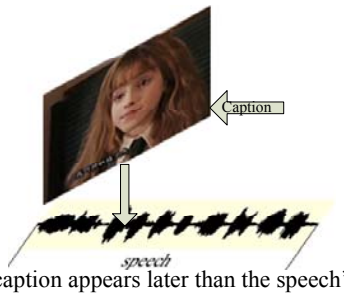
Index Terms— speech detection, caption detection, pitch, Bayesian information criterion (BIC)

1. INTRODUCTION

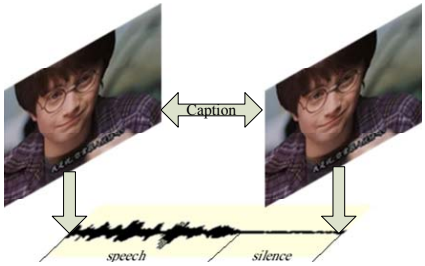
Content-based analysis, indexing and retrieval of videos are becoming more and more pervasive and attractive. Considering that video data contain both audio and visual streams, there is increasing interest in audio-visual fusion methods for video processing [1-3]. In particular, speech recognition and speaker identification techniques can be used to help indexing of video data. For these speech-related processing tasks, speech detection is an important front-end. The audio stream consists of speech, music, and various environmental sounds. Speech detection is to detect and extract homogeneous speech segments from the continuous audio stream.

Traditional frame-energy based voice activity detection [4] is sensitive to noise. More appropriately, an audio segmentation step [7] followed by some advanced speech/non-speech classification [5, 6] is often used. While speech detection using only audio information is of interest in some applications, for content-based video processing, the visual information provide some valuable supplementary cues. In this paper, we propose to exploit caption information for speech detection, which is often readily available from some simultaneous visual processing tasks.

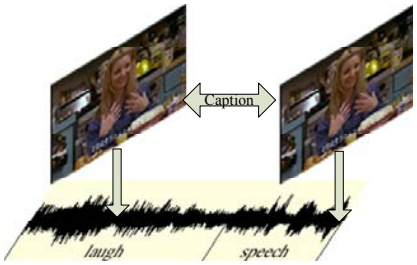
Caption (subtitle) is the text of the dialogs in movie/TV videos, usually displayed at the bottom of the screen. There are two kinds of captions. The first, called soft caption, is stored in a separate track within the video, containing the time-stamps and text content. It is displayed during video playback. The second, called hard caption, is embedded in video frames and broadly used in TV systems without close-caption channel (e.g. in China). For hard caption, text detection is needed to extract the caption information. Much effort has been taken for text detection in videos and images [8-10]. Note that for caption detection, the uniform property of captions in font, size, alignment, and position can be utilized to greatly improve the caption detection performance. Though video OCR can be used for further text recognition, what we needed for speech detection is the extracted time-stamps, indicating the begin/end points for each caption.



(a) The caption appears later than the speech's beginning



(b) The caption still stays on the screen after the speech is over



(c) The caption appears earlier than the speech's beginning, and covers non-speech (music, laugh, etc.)

Fig.1. Illustration of the inaccurate positions of the captions

* This work is supported by Intel China Research Center, NSFC (60402029) and China 863 program (2006AA01Z149).

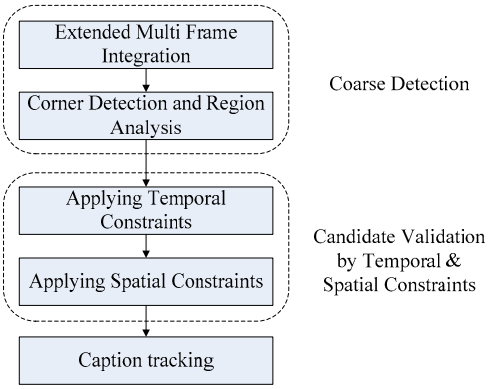


Fig.2. Flow chart of the caption detection algorithm

However, the time-stamps, extracted whether from soft captions or from automatically detected hard captions, are often not accurate enough for locating the speech regions. Captions are mainly created to help viewers to follow the dialog, rather than to accurately indicate the begin/end points of speech. Most captions are marked only near the correct begin/end points of speech, as in Fig. 1(a). Captions may stay on the screen for a little while after the speech is over, or appear earlier than the beginning of the speech, in order that the viewers could watch the caption longer and understand the dialogue better, as in Fig. 1(b) and Fig. 1(c) respectively. As a result of the inaccurate positions for the captions, the captioned-region may include non-speech audios (e.g. music, laugh, etc.), as in Fig. 1(c).

With these observations, we propose a caption-aided speech detection approach, which makes use of both caption information and audio information. The inaccurate positions of the captions are refined through using audio features (pitch and MFCCs) and BIC-based acoustic change detection.

The paper is organized as follows. Section 2 details the proposed approach, including caption detection and the refinement using audio information. Section 3 introduces three other approaches for speech detection. Section 4 presents experimental results, followed by conclusions in the last section.

2. CAPTION-AIDED SPEECH DETECTION

The proposed caption-aided speech detection approach has three steps. First, captions are detected with the time-stamps. Next, pitch segments are extracted from the audio stream using the pitch feature. Finally, a refinement step via collaboration between caption and audio information is performed as follows. 1) Captions are used to rule out non-speech audio from the pitch segments. 2) Pitch segments are used to align the caption begin/end points. 3) Bayesian information criterion (BIC) using MFCC feature is introduced to accurately locate the acoustic changing points.

2.1. Caption detection

The caption detection algorithm includes two stages, as shown in Fig. 2. First, the video is scanned to collect text-like regions, applying an extended multi-frame integration (MFI) [9] technique on consecutive frames to produce caption candidates and their temporal duration. The text-like regions are extracted based on corner detection and region analysis [8]. In the second stage, the temporal and spatial constraints are applied to the caption candidates to filter out false alarms [9, 10].

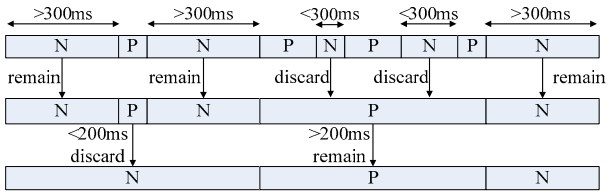


Fig.3. Smoothing of the pitch segments (plotted as 'P') and non-pitch segments (plotted as 'N')

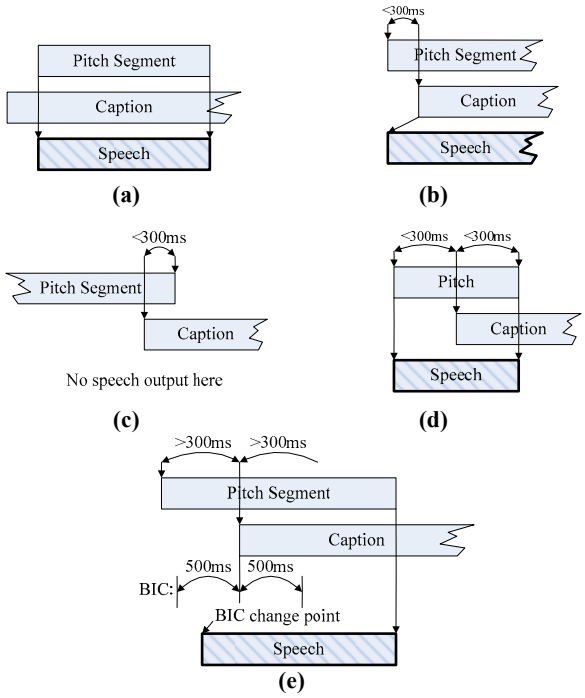


Fig.4. Illustration of the refinement step

2.2. Pitch-based audio segmentation

The ESPS tool *get_f0* [11] is used to estimate the pitch values of the audio stream. Using the pitch feature, the audio is divided into pitch segments and non-pitch segments. Noting that some resulting short segments (typically short noise or short break between two sentences) contain little information, we smooth the segmentation result as illustrated in Fig.3. Non-pitch segments shorter than 300ms and pitch segments shorter than 200ms are smoothed out.

2.3. Refinement via collaboration between caption and audio information

After caption detection and pitch-based audio segmentation, there is a refinement step. The main idea is to adjust the caption begin/end points according to the boundaries given by the pitch segments and the BIC-based change detection. The pitch segments are processed sequentially as follows.

For current pitch segment, we look for possible caption begin/end points in its time range. If no caption points are found, we then check whether the pitch segment falls entirely in any caption's time range. If not, we consider it as useless and discard it. Otherwise, this pitch segment will be output as the speech, as shown in Fig. 4 (a).

If there are caption begin/end points in current pitch segment's time range, they are classified into four types:

1) Caption begin/end points which are near to the pitch segment's begin point (with distance less than 300 ms);

2) Caption begin/end points near to the current pitch segment's end point;

3) Caption begin/end points near to both begin and end point of the current pitch segment;

4) Caption begin/end points far from both begin and end point of current pitch segment.

For the first two types of caption begin/end points, as shown in Fig. 4(b) and (c), we adjust the caption begin/end points to the boundaries of the pitch segment to compensate for the inaccurate positions of the caption.

For the third type of caption begin/end points, which means that the pitch segment are relatively short (less than 600 ms), we consider it as covered by the caption and output it as the speech, as shown in Fig. 4(d).

The fourth type of caption begin/end points are far from both the begin and the end of current pitch segment. This occurs when the pitch segment includes not only speech, but also other sounds that have pitch values, like music. So we introduce BIC over MFCC features[1] to detect the acoustic changing point, as shown in Fig. 4(e). A window of one second centered at the caption begin/end point is searched for the change point. For each candidate change point within the window, a BIC value is computed.[2] The candidate with the maximized BIC value is chosen as the final change point.

3. OTHER SPEECH DETECTION APPROACHES

To evaluate the performance of the proposed caption-aided approach, three other speech detection approaches are introduced for comparison.

The first approach operates in two stages. First, a robust voice activity detection using frame-energy is applied to obtain the voiced segment candidates as in [12]. Second, we use a speech discriminator based on MLER (Modified Low Energy Ratio) [5] and PR (Pitch Ratio) [6] feature to discard non-speech segments from the first stage.

$$\mathrm{MLER} = \frac{1}{2N}\sum_{n=1}^{N}\left[\mathrm{sgn}\left(\mathrm{lowthres} - E(n)\right)+1\right]$$

$$lowthres = \delta \cdot \sum_{n=1}^{N} E(n) / N$$

where N is the total number of frames in the segment, $E(n)$ is the short-time energy of the nth frame, δ is a control coefficient and is empirically set (δ=0.1).

$$PR = PitchNum / FrameNum$$

where *PitchNum* is the total number of frames in the segment that have a pitch and *FrameNum* is the total frame number of the segment. The pitch estimation algorithm is same as in section 2.2. A segment with its MLER value larger than a fixed MLER threshold (0.7) or its PR value smaller than a fixed PR threshold (0.37) will be marked as non-speech and discarded.

[1] A 20-ms frame length and a 10-ms frame shift are used to extract 14-dimension MFCCs from the audio stream.

[2] The BIC value at candidate change point t is calculated between two adjacent 500ms windows: [t-0.5s, t] and [t, t+0.5s].

Name	Language	Captions	Detected	False	Precision	Recall
Full House	Korean	500	492	4	99.2	98.40
DCJ	Chinese	285	281	3	98.94	98.60
Harry Porter	English	407	403	2	99.50	99.01
Sum		1192	1176	9	99.24	98.66

Table 1. Experimental results on caption detection. "Detected" means the number of correctly detected captions. "False" means the number of non-captions which are wrongly detected as captions; "Captions" means the number of captions in ground truth. The recall rate is (Detected/Captions), and precision rate is 1 – (False/(Detected+False)).

The second approach starts from the pitch segments obtained in section 2.2. A speech discriminator using only MLER feature is then applied to reject non-speech pitch segments. The PR feature is not used here in the speech discriminator since pitch information is already used to obtain the pitch segments.

The last approach also starts from the pitch segments. For each segment obtained in section 2.2, DISTBIC is first used to detect acoustic change point within the pitch segment's time range [7]. After that, the segment is divided into several shorter segments. Finally, the same speech discriminator as used in the second approach is applied to reject non-speech segments from these shorter segments.

4. EXPERIMENTAL RESULTS

Different speech detection approaches are evaluated on 7 excerpts from 3 different TV series and 1 excerpt from a movie:

- DCJ: 2 excerpts from Korean TV series "Dae-Jang-Geum" (episode 42 and 44), in Chinese, about 80 minutes in total.
- Friends: 4 excerpt from American TV series "Friends" (episode 1, 2, 3, 4 in season 3), in English, about 90 minutes in total.
- FH: 1 excerpt from Korean TV series "Full House" (episode 2), in Korean, about 30 minutes.
- HP: 1 excerpt from American movie "Harry Porter and the Sorcerer's Stone", in English, about 30 minutes.

The test data are selected to cover a wide range of audio conditions. There is much laughter in Friends series. Background music, song and voiced non-speech sounds are more observed in DCJ and FH series. The Harry Porter movie contains various kinds of complicated sound effects.

For these video data, Friends series contain soft caption, so we can obtain caption information directly. Captions for the other three videos are hard caption, so caption detection algorithm in section 2.1 is applied. The caption detection result is given in Table 1, which shows a precision rate of 99.24% and recall rate of 98.66% on average. As expected, by taking advantage of the uniform property of captions in font, size, alignment, and position, the performance of caption detection is much better than general-purpose text detection.

Different speech detection approaches are evaluated by miss rate, alarm rate and correct rate. When comparing the detected speech with the manually-labeled speech, there are two types of errors: miss (speech in reference but not in hypothesis) and false alarm (speech in hypothesis but not in reference). A forgiveness collar of 0.25 seconds (both + and -) will not be scored as error around each boundary. The correct rate is defined as (total time – miss time – false alarm time)/(total time).

		Friends	DCJ	FH	HP
Caption only	Miss %	0.2	2.4	0.3	1.8
	False alarm %	34.8	12.6	19.7	19.3
	Correct %	**65.0**	**85.0**	**80.0**	**78.9**
VAD _MLER _PR	Miss %	8.5	9.2	3.7	9.0
	False alarm %	13.1	3.0	11.3	8.2
	Correct %	**78.4**	**87.8**	**85.0**	**82.8**
PitchSeg _MLER	Miss %	2.7	6.2	9.1	6.2
	False alarm %	6.8	12.3	9.5	19.6
	Correct %	**90.5**	**81.5**	**81.4**	**74.2**
PitchSeg _BIC _MLER	Miss %	3.5	7.5	3.9	6.7
	False alarm %	5.8	3.1	9.2	7.6
	Correct %	**90.7**	**89.4**	**86.9**	**85.7**
Caption -aided	Miss %	2.4	4.4	1.5	6.3
	False alarm %	5.2	2.6	2.0	7.1
	Correct %	**92.4**	**93.0**	**96.5**	**86.6**

Table 2. Results for various speech detection approaches

Experimental results for various speech detection approaches are given in Table 2. The three other detection approaches introduced in section 3 are denoted as VAD_MLER_PR, PitchSeg_MLER and PitchSeg_BIC_MLER respectively. The speech detection result from using only caption time-stamps is also shown.

When comparing the three other approaches, VAD_MLER_PR approach has a high false alarm rate in Friends data while a low false alarm rate in DCJ. This suggests that VAD_MLER_PR can identify and reject music, but fails in removing laughter and other non-speech noises. PitchSeg_MLER approach is just the opposite: it performs better in Friends because this approach can discard most non-pitch segments such as laughter. But it is not as good as the VAD_MLER_PR approach in DCJ, FH, and HP data, which contain a lot of non-speech sound effects with pitch values. In these three approaches, PitchSeg_BIC_MLER algorithm achieves better performance, with both lower miss rate and lower false alarm rate. MLER speech discriminator is useful for these three approaches. As shown in Fig.5, the correct rate is greatly improved with the use of MLER in Pitch_BIC_MLER approach.

Caption-aided approach achieves the best detection performance with the lowest miss rate and false alarm rate. Note that the detection result from using only caption time-stamps is bad due to the inaccurate positions of the captions. However, captions indeed provide lots of useful information. Segments not in the caption's time regions can be regarded as non-speech with confidence, and the caption's begin/end points suggest potential acoustic changing points. With the help of pitch feature to accurately locate the speech begin/end points and the BIC criterion to accurately find the acoustic changing points, the caption-aided approach achieves the best performance over the other three approaches for real-world movie/TV videos.

5. CONCLUSION

In this paper, we propose a caption-aided speech detection approach, which makes use of both caption information and audio information. The inaccurate positions of the captions are refined through using audio features (pitch and MFCCs) and BIC-based acoustic change detection. Comparison experiments against several other traditional speech detection approaches are conducted in real-world movie/TV videos, showing that the proposed approach improves the speech detection performance greatly.

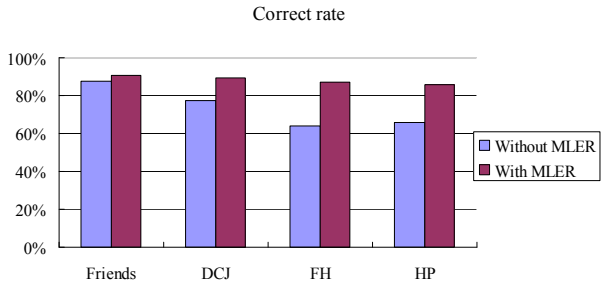


Fig.5. Results for BIC_Pitch_MLER approach with/without MLER speech discriminator.

6. REFERENCES

[1] Y. Li, S. Narayanan, C.J. Kuo, "Content-Based Movie Analysis and Indexing Based on AudioVisual Cues", IEEE transactions on circuits and systems for video technology, vol. 14, no. 8, pp.1073-1085, 2004.

[2] J. makhoul, et. al., "Speech and Language Technologies for Audio Indexing and Retrieval". Proceedings of the IEEE, Vol.88, No.8, pp.1338-1353, 2000.

[3] Q. Huang, A. Puri, Z. Liu, "Multimedia Search and Retrieval: New Concepts, System Implementation, and Application", IEEE transactions on circuits and systems for video technology, vol. 10, no. 5, pp.679-692, 2000.

[4] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition" , IEEE Trans. Acoust., Voice, Signal Processing, v29, pp. 777–785, Aug. 1981.

[5] W.Q. Wang, W. Gao and D.W. Ying, "A Fast and Robust Speech/Music Discrimination Approach", ICICS-PCM 2003, pp. 1325-1329, 2003.

[6] Ahmad R. Abu-El-Quran and Rafik A. Goubran, "Pitch-Based Feature Extraction for Audio Classification", ICASSP 03, pp. 43-47, 2003.

[7] P. Delacourt and C.J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing", Speech Communication 32, pp. 111-126, 2000.

[8] Xiansheng Hua, Xiangrong Chen, Liu Wenyin, Hong-jiang Zhang. "Automatic location of text in video frames", Proceedings of ACM Multimedia, workshop on Multimedia information retrieval, 2001.

[9] Rongrong Wang, Wanjun Jin, and Lide Wu, "A novel video caption detection approach using multi-frame integration", In Proceedings of IEEE International conference on Pattern Recognition, pp. 449-452, August 2004.

[10] A. Wernicke, R. Lienhart, "On the Segmentation of Text in Videos", IEEE Int. Conf. on Multimedia and Expo (ICME 2000), pp. 1511-1514, New York, USA, July 2000.

[11] "ESPS with waves", Entropic Research Laboratory, Inc. AT&T Bell Laboratories, 1993.

[12] Ye Tian, Ji Wu, Zuoying Wang and Dajin Lu, "Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection", ICASSP 2003, pp. 444-447, 2003.