ICA 2019 AACHEN

PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Detection of anchors' utterances in broadcast news using i-vector-based speaker similarity and temporal information

Daichi NOZAKI; Masaru YAMASHITA; Hiroyuki TAKADA; Shoichi MATSUNAGA

Nagasaki University, Japan

ABSTRACT

Accurate speech recognition of anchors' speech in broadcast news programs is essential for indexing tasks such as topic extraction, summarization, and so on. Consequently, the number of anchors and all speech segments for each anchor are important aspects that must be evaluated correctly. This paper discusses the effectiveness of these two items based on speaker similarity using i-vector. In our proposed approach, first, each input audio source was segmented into four types of sound: speech, noise, silence, and music. Next, a clustering was performed using all speech segments to collect segments from the same speakers. In this clustering, cosine similarity from i-vector was used. However, it was difficult to extract sufficient speaker characteristics from short utterances. To address this problem, we considered the time intervals between successive pairs of speech segments and the existence of background noises, adding to the information obtained from i-vector. Among the detected speaker clusters, anchor clusters were obtained based on the heuristic that the ratio of the number of speech segments from anchors was much higher than that of other speakers. It was experimentally shown that the detection method of anchors' speech using the proposed approach outperformed the method using conventional k-means clustering.

Keywords: Anchor, Speech, I-vector, Clustering

1. INTRODUCTION

Today, consumers of multimedia demand increasing access to multimedia out of the wide array of available contemporary and archival content. A typical example of such content is broadcast news. Accessing this content effectively requires useful descriptions (referred to frequently as metadata), which are made based on the verbal information extracted from speech and text data, including nonverbal sound and visual information. For the producer, however, creating metadata manually is expensive and time-consuming. Automatic speech recognition, speaker detection, and audio source detection including voice activity detection, are possible solutions to these problems. With the development of machine learning techniques, state-of-the-art technologies in these areas have increased recognition and detection performance drastically. Unfortunately, the presence of various competing sound sources common to most multimedia content significantly degrades the performance of metadata producing algorithms

Recently, many studies have been conducted on indexing multimedia content for information retrieval, focusing on areas such as topic detection and summarization [1-3]. Our paper focuses on the efficient indexing of broadcast news using audio information. In broadcast news content, speech comes from many sources such as anchors, commentators, interviewers, and interviewees. We focused on what anchors said because anchors direct the flow of topics as such include most of the relevant keywords for the development of metadata. In our research, we assumed that there was no preexisting information regarding the anchors' acoustic information or the number of anchors, instead relying on general assumptions about the timing and environment around their speech. Generally, we expect that the proportion of anchors' speech to the total amount of speech in a broadcast is much higher than any other speakers. This allows us to separate their speech from other speakers by performing voice activity detection based on the frequency of speech and similarity to other sounds from the speaker. If some clusters consist of speech segments which seem to be spoken by the same speaker, and the amount of the speech segments is proportionally great, we can assume that a single individual said those speech segments and that the speaker is an anchor. Likewise, these patterns of speech can be used to identify different anchors within a broadcast.

I-vector has been widely used for speaker verification and recognition [4, 5]. We have used it in our study to detect and assign speech to individuals. Detection of long segments of speech was successfully carried out; however, for shorter segments, the detection rates became lower, because it was difficult to extract sufficient speaker characteristics [6-8]. In broadcast news programs, there are several phrases, which although are short, vital for indexing.

To address this problem, we specifically focused on short time intervals between successive segments of speech and the lack of background noise and combined these with the information acquired from the i-vector. Trained anchors in a studio environment use the cadence of their speech and a quiet background to enhance the listening environment for the audience [9]. Through our research of Japanese broadcast news, we found that the time between speech segments uttered by the same individual is shorter than that between speech segments uttered by two people. If the time interval between successive segments of speech is short enough, these are connected into one long segment and the i-vector is calculated again using this newly segment of sound. Another way to distinguish anchors' speech is to compare lack of background noise, because they usually speak in a studio. If background noise is not detected around the successive segments of speech and the i-vector of one segment is similar to that of another segment, these can be added onto the segment of speech and the i-vector re-calculated. In this paper, we evaluated the effects of these two methods for improving i-vector on short segments of speech during the detection of anchors' speech segments in broadcast news content.

2. ANCHORS' SPEECH IN BROADCAST NEWS

2.1 Manual Tagging for Audio Broadcast News Programs

We prepared tagged data for five broadcast news programs to evaluate the detection performance of anchors' speech. The length of each program was about 30 min, with either one (Data I) or two (Data II-V) anchors. We manually tagged the data by segment according to the four types of sound sources: speech, music, noise, and silence segments and determined the beginning and ending times for each segment.

Speech segments consisted of words from anchors, reporters, interviewees, and all other transcribable speech. These segments included very short pauses, such as when the speaker took a short breath. However, background voices, which were usually picked up from outdoor locations, were tagged along with traffic and other general sound sources as noises. We identified individual speakers with a numerical tag for each speech segment to compare how well our methods were able to detect each anchor's speech. Jingles were tagged as music. Some tagged segments overlapped, such as an interview in a crowd, which would contain both speech and noise.

2.2 Detection Flow of Anchors' Speech

Figure 1 illustrates how we detected anchors' speech. First, we applied sound source segmentation to the evaluation audio data and subsequently extracted speech segments. Then, we performed speech clustering by finding the speaker similarities between speech segments to assign speech to individual speakers based on how these assignments were clustered. Finally, clusters in which the ratio of segments to total speech segments was larger than the predefined thresholds were assigned to the anchors. By counting the number of clusters that exceeded the thresholds, we estimated the number of anchors.

The sound source detection procedure detected four types of sound segments: speech, music, noise, and silence. We used seven acoustic features: signal energy, pitch frequency, peak-frequency centroid, peak-frequency bandwidth, spectral cross-correlation, temporal stability, white-noise similarity, and spectral shape [10,11]. In our speech detection procedure, we obtained the detection results for each timeframe for the four sound segment types and the likelihood being of each type for each feature parameter. The likelihood was calculated using the acoustic Gaussian mixture models (GMMs) of each type. The results for each timeframe were smoothed using a longer window to avoid inaccurate detection of very short segments. The details of this procedure are described in our reference [11].

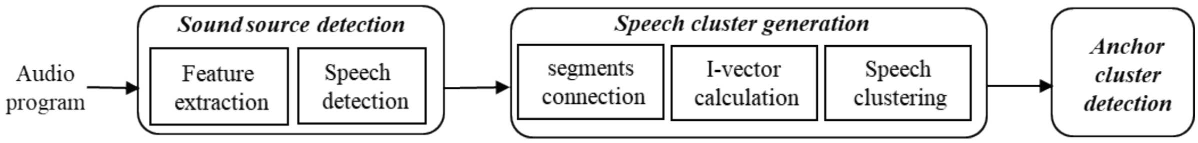


Figure 1 – Anchor speech detection procedure

2.3 Speaker Similarity using I-vector

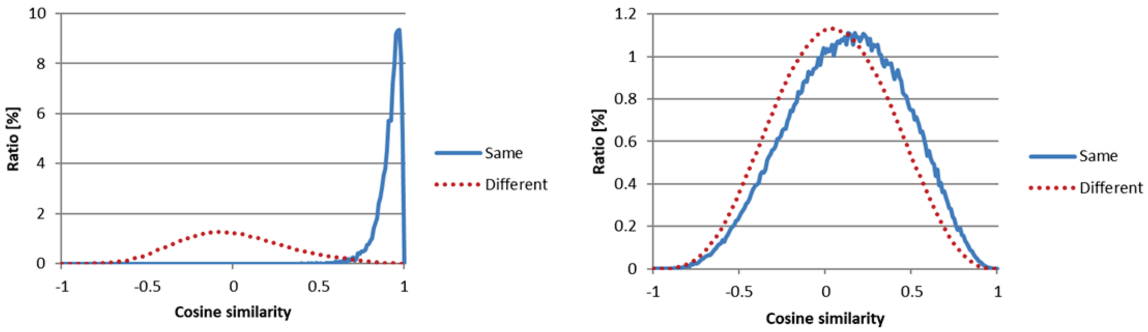
To correctly assign speech segments to an individual speaker, we needed to exactly capture speaker-dependent acoustic features in each segment of speech. To do this, we measured the difference in speaker features. We used i-vector to describe speaker characteristics and to measure the speaker similarity between two segments of speech by using the cosine similarity between two segments' i-vectors. This measure using i-vector usually achieves good performance for longer segments of speech but is inaccurate for shorter ones, as demonstrated in Table 1, which shows the mean value and standard deviation of cosine similarity for long speech segments (> 20 s) and short speech segments (< 300 ms) using speech data read by 110 speakers in Japanese. Figure 1 also shows the distributions of both lengths of speech segment. Importantly, there were few differences in the ratio of short speech segments between different speakers and a single speaker. This indicates that information is needed to detect the anchors' short speech segments in addition to the cosine similarity of i-vectors.

To address this problem, we focused on using the time interval between successive speech segments from a single speaker and different speakers. Table 2 shows the mean value and standard deviation for two interval lengths in our tagged data. Their distributions are shown in Figure 2. These indicates that the interval between successive speech segments from the same speaker is generally shorter compared to that between one speaker and the other. In this paper, we experimentally verify the effectiveness of utilizing this information.

Another way in which the issue of i-vector can be solved for short lengths of speech is by using background noise. In most cases, anchors work in a silent studio that typically results in low levels of background noise. Higher levels of noise most often indicates that the person speaking is not an anchor. We experimentally evaluated this method in section 4.

Table 1 – Value of cosine similarity between read speech segments from same/different speakers

	Long speech segments (> 20 s)		Short speech segments (< 300 ms)	
	Same speakers	Different speakers	Same speakers	Different speakers
Mean	0.91	0.00	0.12	0.03
S.D.	0.07	0.31	0.33	0.32



(a) Long speech segments

(b) Short speech segments

Figure 2 – Cosine similarity distribution for long and short read speech segments

Table 2 – Interval between a pair of successive speech segments (in sec)

Interval	Same speakers	Different speakers
Mean	0.68	2.44
S.D.	1.15	1.0

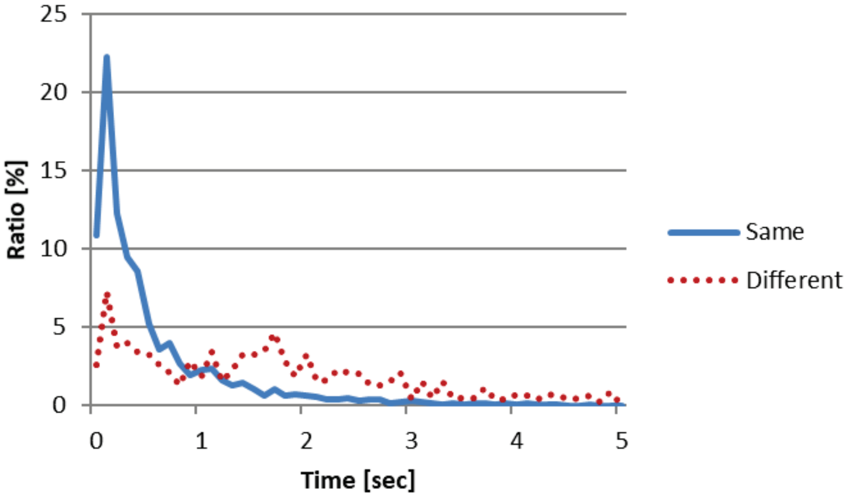


Figure 3 – Interval distributions between successive speech segments from the same speaker vs. different speakers in broadcast news

3. DETECTION OF ANCHORS' SPEECH

3.1 Detection of Speech from an Individual Speaker based on Speaker Similarity

After the detection of speech segments in the news content, we gathered speech segments from individual speakers based on speaker similarity. One of the most conventional methods to group speech is by using k-means clustering based on speaker-dependent characteristics, such as the i-vector. As before, we found this method unsuitable due to the several short speech segments in which the i-vector did not prove to be good acoustic feature. Instead, we focused on the larger volume of anchor speech as compared to other speakers. Our detection approach was to find the speech clusters in which speaker characteristics for each speech segment were similar and in which the speech made up a larger proportion of the total segments. Our approach to detect speech segment clusters is as follows:

Step 1: Calculate the i-vector based cosine similarity $D\left(I(u_i), I(u_j)\right)$ between all pairs of all long speech segments $\left(u_i, u_j, 1 \leq i, j \leq N\right)$, where N is the number of the detected speech segments in the sound source detection.

Step 2: Determine the ratio $C(u_i)$ of speaker-similar speech segments for each speech segment u_i. The speaker-similar speech segment u_j were selected as members in the cluster $S(u_i)$ using a pre-defined threshold D_{Th1}: $D\left(I(u_i), I(u_j)\right) > D_{Th1}$, where u_i is a centroid in the cluster. If the ratio $C(u_i)$ is larger than a pre-defined threshold C_{Th}, the speech segments in this cluster was recognized as belonging to an anchor.

Step 3: Detect short segments of speech u_k belonging to the cluster $S(u_i)$ if $D\left(I(u_i), I(u_k)\right) > D_{Th2}$ to compensate for insufficient i-vector for short segments of speech. D_{Th2} is also a pre-defined threshold that must be larger than D_{Th1}.

Steps 2 and 3 were repeated for unselected speech segments for all segments until the ratio of selected speech segments in a cluster in step 2 was less than the pre-defined threshold C_{Th}. Finally, the number of generated clusters was recognized as the number of anchors within the news broadcast.

3.2 Speech Segment Connection for Accurate Detection of Anchors' Speech

In this section we develop our baseline method for detecting distinct speech for individual anchors. In this baseline, we perform the clustering process for short speech segments as described in step 3, however, this process is not sufficient to deal with all short speech segments. As a result, we tried to derive more reliable i-vector value by connecting successive segments of speech and building upon the algorithm described in section 3.1 using temporal information and background noise information with the following three methods:

Method I (temporal information): If the time interval between successive speech segments is less than the pre-defined time-length, and cosine similarity of i-vectors of these segments is also larger than the pre-defined threshold, these segments are connected and i-vector is re-obtained using the newly combined segment.

Method II (background noise information): If all neighboring acoustic segments to the successive speech segments are silence, as shown in Figure 4, and the cosine similarity of i-vectors of the segments is also larger than the pre-defined threshold, these segments are connected and i-vector is re-obtained using the combined segment.

Method III (both types of information): Methods I and II are combined by connecting the pair of successive segments if either condition (time interval length or silence) is met.

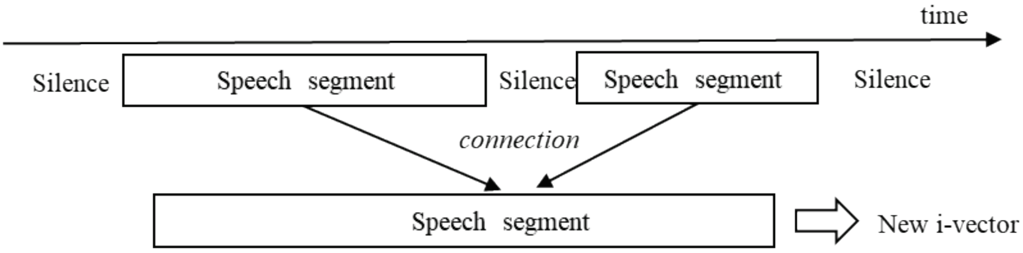


Figure 4– Recalculation of i-vector for successive speech segments in method II

4. Evaluation Experiments

4.1 Experimental Conditions

We used five broadcast news programs as described in section 2.1 for evaluation. In this data, speech and silence segments accounted for about 60% and 15% of the total audio, respectively. In the sound source detection process, the sampling rate of the audio data was 44.1kHz. Every 23 ms, the seven acoustic feature parameters in section 2.2 were determined using a 46-ms Hamming window. The acoustic GMM with a two-mixture Gaussian distribution was used for each sound source. The speech detection performance (F-measure) of this sound source detection process was 89.3%.

In the speech cluster detection process, acoustic features under 8 kHz were used to capture speaker characteristics. I-vector dimensionality was 10, and universal background model (UBM) size was 60. The UBM was trained using 50 sentences read by 55 male and 55 female speakers.

The thresholds $(D_{Th1},\ D_{Th2})$ for cosine similarity and that for the ratio of the number of speech segments C_{Th} to the overall speech segments was determined experimentally.

4.2 Detection Performance

Finally, we conducted detection experiments of anchors' speech from two points of view: speech segment clustering and speech segment connection. Through the lens of speech segment clustering, we compared the conventional k-means clustering based on the speaker-similarity between speech segments to the proposed clustering method as described in section 3.1. From the perspective of speech segment connection, the detection experiment without connecting speech segments (our baseline) was compared to methods I,II, and III, which connected speech segments based on the time interval, background silence, and both, respectively.

Table 3 shows the detection performance of each method. The detection performance (F-measure) of the baseline was higher than that of the k-means clustering. This indicates that clustering based on the ratio of anchor speech to total speech, grouped by speaker similarity, is more effective. Table 3 also shows that the performances of methods I, II, and III were superior to that of the baseline, and method I achieved the highest performance. These results show that the connection of speech segments prior to the clustering of speech segments, creating longer segments of speech to cluster, making it possible to more reliably obtain i-vector and more accurately detect an individual anchors' speech. It also shows that connecting speech segment based on background silence is useful, but not as much when connecting speech based solely on time interval between speech segments.

Every detection method except for k-means clustering correctly identified the number of anchors, however more evaluation data is needed to analyze their performance more thoroughly.

Table 4 shows the detection performance of anchors' speech for each broadcast analyzed. These results indicate that the speech detection and assignment in the case of two simultaneously reporting anchors in a broadcast is more difficult than that of one anchor.

Table 3 – Detection performance of anchors' speech

Method	Recall	Precision	F-measure
k-means clustering	0.961	0.495	0.645
Baseline	0.748	0.670	0.707
Method I	0.747	0.798	0.765
Method II	0.737	0.722	0.730
Method III	0.733	0.776	0.754

Table 4 – Detection performance of anchors' speech for each broadcast

Evaluation news	No. of anchors	F-measure
Data I	1	0.815
Data II	2	0.722
Data III	2	0.789
Data IV	2	0.711
Data V	2	0.765

5. CONCLUSIONS

This paper investigated the detection methods for effectively isolating and assigning individual news anchor's speech to automatically index broadcast news programs. In our approach, speech from an individual speaker was collected using the cosine similarity of i-vector for each segment of speech. News anchor's speech was separated from the speech of other speakers by establishing that the ratio of anchor's speech to total speech was higher than that for any other speakers. Since using i-vector on short speech segments is ineffective due to the lack of sufficient speaker characteristics, we connected successive pairs of speech segments into one speech segment. To do this, we used both the time intervals between speech segments and the lack of background noises, then added this to the information obtained from i-vector. We connected these shorter speech segments prior to clustering them to individual anchors to use more reliable i-vector for each segment of speech. The experimental results showed that the clustering method which considered the ratio of anchor speech to overall speech was effective. We also obtained useful results by connecting speech segments based on the time interval between segments.

Due to the time-consuming nature of this experiment, the quantity of data analyzed was small and further research is required to validate our findings. Since the thresholds used in our methods were determined experimentally, these may also benefit from refinement with larger data sets.

REFERENCES

1. Dharanipragada S, Franz M, Roukos. Audio indexing for broadcast news. Proc TREC-7, 1998, p. 115-119.
2. Johnson S, Jourlin P, Jones K, Woodland P. Audio indexing and retrieval of complete broadcast news shows. Proc RIAO, 2000, p. 1163-77.
3. Logan B, Goddeau D, Thong J. Real-world audio indexing systems. Proc IEEE ICASSP, vol.5, 2005, p. 1001-4.
4. Dehak N, Dehak P, Kenny P, Brummer N, Ouellet P, Dumouchel P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. Proc. Interspeech, 2009. p. 1559-62,
5. Dehak N, Kenny P, Dehak P, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Language Processing, 19, 2011. p. 788-798.
6. Kanagasundaram A, Dean D. Vogt R, McLaren M, Sridharan S, Mason M. Weighted LDA techniques for i-vector based speaker verification, Proc IEEE ICASSP, 2012, p. 4781-4.
7. Kanagasundaram A, Vogt R, Dean D, Sridharan S, Mason M. i-vector based speaker recognition, Proc Interspeech, 2011, p. 2341-4.
8. Sarkar A, Matrouf D, Bousquet P, Bonastre J. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification, Proc Interspeech 2012
9. Douglas G, The impact of interstimulus interval and background silence on recall. Journal of Consumer Research, Vol. 23, No. 4, 1997, pp. 295-303.
10. Matsunaga S, Mizuno O, Ohtsuki K, Hayashi Y. Audio source segmentation using spectral correlation features for automatic indexing of broadcast news. Proc EUSIPCO, 2004. p. 2103-6.
11. Matsunaga S, Yamaguchi M, Yamauchi K, Yamashita M. Sound source detection using multiple noise models. Proc IEEE ICASSP, 2008, p. 2025-8.