THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.26, NO.3E 2007. 12 pp. 74~79

Retrieval of Broadcast News Using Audio Content Analysis

Hyoung-Gook Kim*
*Intelligent Multimedia Signal Processing, Kwangwoon University
(Received July 18 2007; Revised September 10 2007; Accepted September 28 2007)

Abstract

In this paper, we report our recent work on a indexing and retrieval system of broadcast news using audio content analysis. Key issues addressed in this work are two major parts of the audio indexing system: anchorperson detection based on audio segmentation, and phone-based spoken document retrieval, developed in the framework of the emerging MPEG-7 standard. Experiments are conducted on a database of Britisch broadcast news videos. We discuss the development of the retrieval system, and the evaluation of each part and the retrieval system.

Keywords **:** *Indexing and retrieval system, Anchorperson detection, Audio content analysis*

I. Introduction

Video streams of multimedia documents such as broadcast news need to be indexed before they can be retrieved with content-based queries. To this end, methods for automatic content-based analysis are needed. The automatic extraction of high-level semantic information is very useful for the automatic content-based multimedia indexing and retrieval so that it can largely avoid manual annotation.

In this paper, we present a retrieval system of broadcast news which integrates two audio content-based approaches: Anchorperson (AP) detection and Spoken Document Retrieval (SDR). The broadcast news video is a sort of well structured video, because it is usually introduced and summarized by the AP prior to and following the detailed reporting conducted by correspondents and others. The AP segments provide the landmarks for detecting the semantic content boundaries. Therefore, the AP detection is useful for fast topic-oriented navigation within news by a viewer or for further content-based analysis. Various approaches have been suggested in the literature [1][2]. Qi et al. [1] suggested an audiovisual technique for AP detection. They used speaker change point detection and speaker clustering in the audio domain and key-frame clustering in the visual domain. Delacourt et al. [2] proposed a speaker-based segmentation technique which composes of a distance-based algorithm followed by a Bayesian information criterion (BIC) algorithm.

Spoken parts of the anchor segments enclose a lot of semantic information. This information, called *spoken content*, contains actual topic keywords so that it is important to perform automatic indexing of news stories based on automatic speech recognition (ASR) system. In the past decade, the extraction of spoken content metadata has become a key challenge for the development of efficient methods to index and retrieve audio-visual documents. Several works have proposed SDR approaches that use some ASR specific information [3][4]. In the same way, the SDR approach presented here use phonetic information only. Our indexing system does not require any *a priori* word lexicon. The use of phones as basic indexing terms restrains the size of the indexing lexicon to a few dozens of units. However, phone recognition

Corresponding author: Hyoung-Gook Kim (hkim@kw.ac.kr)
Dept. of Wireless Communication Engineering, Kwangwoon University, 447-1, Wolgye-Dong, Nowon-Gu, Seoul, 139-701

systems have to contend with high error rates. For this, we consider two different ways of coping with the problem of phone recognition inaccuracy. The first one consists of indexing the spoken documents with MPEG-7 [4] phone lattices rather than 1-best transcriptions. This allows taking several competing phone hypotheses into account. The other one takes advantage of some ASR specific information, the phone confusion statistics [4][5], to compensate for the recognition errors. The SDR system presented here exploits the phone confusion matrix enclosed in the MPEG-7 spoken content descriptors for expanding the phonetic representation of documents.

II. Retrieval System Framework

In Figure 1, the structure of a retrieval system is described.

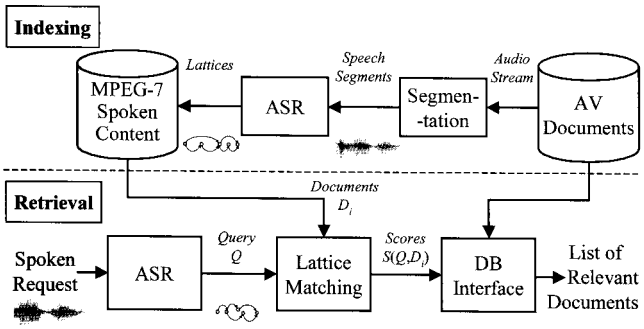

Figure 1. System structure for the indexing and retrieval of broadcast news.

It mainly consists of the following steps:

- The TV broadcast news received from a television satellite are stored as MPEG-compressed files. The audio data are compressed with MPEG audio layer 2 compression at a data rate of 192 kbit/s and a sampling rate of 44.1 kHz. The recorded audio signal is sampled down to a 22.05 kHz/16 bit PCM format.
- During the indexing phase the segmentation of the audio streams are applied to identify spoken parts and discard non-speech signals (or non-exploitable speech signals, e.g. if too noisy). The spoken part is fed into AP detection, where the speech recording is segmented into regions spoken by AP. The AP speech segments are useful for the topic boundaries or for spoken content analysis. A document representation D is the spoken content description extracted through ASR from the corresponding AP speech segments. When phonetic indexing is used, the typed word has to be preprocessed in order to get its phonetic transcription.

Once indexed, the request is spoken input to the system. Depending on the retrieval scenario, whole sentences or single word-requests may be used. After the query Q has been formed from a spoken request, it is compared to each archived document D_i (i.e. the associated metadata). This is based on a score $S(Q, D_i)$ reflecting how *relevant* is D_i with respect to Q (i.e. how likely will D_i satisfy the user's request). These *relevance* scores are finally used to rank the document and output the most relevant ones.

2.1. Audio Segmentation and Anchorperson Detection

The input audio signal is transformed into a feature vector sequence, which consists of the logarithmic frame energy and the first 12 Mel-frequency cepstral coefficients (MFCCs). The segmentation uses these features and performs a maximum likelihood (ML) classification of half a second sub-segments. Here, the segmentation distinguishes between two categories. The first category is the speech without noise or other environmental sounds in the background. The second category is the noisy speech or other possible audio signals included in broadcast news. The temporal segments with clean speech have a direct relation with the well-defined recording settings in a studio environment. Therefore, these segments are referred to as studio segments. As we expect that the anchor person is only included in these studio segments, the further analysis process regards only these segments. Hence, only sub segments classified as the category studio are used to build smoothed studio segments and all other sub-segments are ignored. The ML-classification uses Gaussian Mixture Models (GMMs). As it is also possible that in one studio segment different speakers are included, pause detection determines additional potential segment boundaries. These are possible speaker change points.

The cluster analysis is performed in three steps. At

first, the distances between the segments are computed for speaker change detection. The principle behind speaker change detection is to measure a dissimilarity value between two consecutive parts of the parameterized signal, assuming the each of these parts is related one speaker only. Afterwards, a Hierarchical Cluster Analysis (HCA) is carried out in order to merge speech segments containing the same speaker. Finally, the AP cluster is selected.

To determine the distances between studio segments, the low-level features from the feature extraction are used to determine the distances between studio segments. Inspired from the speaker segmentation [2], Kullback-Leibler divergence (KL2), Generalized Likelihood Ratio (GLR), and the Bayesian Information Criterion (BIC) are compared as distance measures, since KL2, GLR, and BIC are among the most commonly used criteria for the audio change detection. The single, complete, and average linkage methods were compared for the HCA. The HCA results in clusters A_k, where the index k identifies the k-th cluster. These clusters of studio segments are used as input for the cluster selection, where the AP cluster will be chosen. Assuming that each cluster includes segments with similar speech utterances from one speaker, the AP is identified by the verification of three cluster selection criteria for the temporal structure. First, the person has the appearance with the longest duration as he/she dominates the broadcast. Second, the appearance of the person is close to the center of the broadcast as the person introduces topics repeatedly throughout the broadcast. Third, the person has the highest variance of time points as he/she is usually present at the start and the end of the broadcast. Thus, the mean $\mu_{a,k}$ and the variance $\sigma_{a,k}$ of the time points of frames are determined for each cluster. The mean value is used to compute the closeness $c_{a,k}$ to the center t_μ of the video with

$$c_{a,k} = 1 - \left| t_\mu - \mu_{a,k} \right| / t_\mu \,. \tag{1}$$

$c_{a,k}$ lying in the range [0, 1]. With the center closeness, the variance and the total frame number $N_{a,k}$ for the studio segments contained in the k-th cluster, the vector

$$a_k^T = \left(c_{a,k}, \sigma_{a,k}^2 / \max_{\forall k}\left(\sigma_{a,k}^2\right), N_{a,k} / \max_{\forall k}\left(N_{a,k}\right) \right). \tag{2}$$

can be defined for each cluster. The I-th cluster is chosen, if it maximizes the Euclidean norm ($L2$) of all obtained vectors, so that

$$I = \arg\max_{1 \le k \le M_a} \left(\left\| a_k \right\|_2 \right). \tag{3}$$

identifies the index of the AP cluster.

Once the AP cluster is chosen, a post-processing removes the segment boundaries introduced by pause detection between AP segments by merging such segments with a distance below 6 seconds. At the end of the audio segmentation, a result list of start and end time points of AP segments is found.

2.2. Spoken Content Indexing

The MPEG-7 spoken content tool [4] defines a standardized description of the lattices delivered by a phone recogniser. The lattice consists of an oriented graph whose paths represent different possible transcriptions. Each node in the graph represents a time point between the beginning and the end of the speech signal. A link between two nodes corresponds to a recognition hypothesis (e.g. a word). The recogniser used for indexing performs phone recognition without any lexical constraints. The 46 context independent Markov models are looped, according to a bigram language model (LM). Given a spoken input, our ASR system produces an output phone lattice containing several hypothesized phonetic transcriptions. The recogniser has been tested on the WSJ0/WSJ1 corpus which had not been used for training. This provided the phone confusion probabilities. The resulting confusion matrix is enclosed, along with the phone lexicon and the phone lattices, in the MPEG-7 spoken content descriptions.

2.3. Retrieval

Our retrieval technique is based on the Vector Space Model (VSM). The model creates a space in which both documents and queries are represented by vectors. Given a query Q and a document D, two T-dimensional vectors

q and d are generated, where T is the predefined number of indexing terms. Each component of q and d represents a weight associated to a particular indexing term. The most straightforward is a binary weighting, in which a vector component is simply set to "1" if the corresponding indexing term is present. For a given term t, the corresponding components in q and d are:

$$q(t)=\begin{cases}1 & \text{if } t\in Q\\ 0 & \text{otherwise and}\end{cases}$$

$$d(t)=\begin{cases}1 & \text{if } t\in D\\ 0 & \text{otherwise.}\end{cases} \quad (4)$$

The indexing terms used in this study are phone 3-grams, i.e. the sequences of 3 successive phones are extracted from the transcriptions or lattices used for indexing the queries and the documents. In that case, the indexing terms t mentioned in equation (4) are all 3-phone sequences extracted from the phone transcriptions or the phone lattices used to index the queries and the documents.

In this study, we compare three approaches as following:

1) *Binary weighting and a simple inner vector product to compute the information retrieval (IR) similarity*: a measure of similarity between Q and D is then estimated by using the inner product of q and d:

$$S(q,d)=\frac{\sum_{t\in Q}q(t).d(t)}{\sqrt{\sum_{t\in Q}q(t)^2}\cdot\sqrt{\sum_{t\in Q}d(t)^2}}=\frac{1}{\|q\|\cdot\|d\|}\sum_{t\in Q}q(t).d(t) \quad (5)$$

It allows to create a list of relevant documents, ordered according to their relevance scores, which can be returned to the user.

2) *Combination of 3-grams lengths*: For a given document, the retrieval scores obtained using each set separately can be combined to get a single score. A simple combination of monogram (N=1), bigram (N=2) and trigram (N=3) indexing terms can be defined by using the following relevance score:

$$S_{1,2,3}(q,d)=\frac{1}{6}\sum_{N=1}^{3}N\cdot S_N(q,d) \quad (6)$$

where S_N represents the relevance score of equation (5), obtained with the set of 3-gram indexing terms. This combination allows to take short indexing units into account. At the same time, it gives more weight to the longer ones, which are more sensitive to recognition errors but contain more information.

3) *Expansion based on confusion probabilities*: the phone confusion probabilities to compensate for the recognition errors. The elements $P(\varphi_1|\varphi_2)$ of this matrix represent the probability that the phone φ_1 is recognized instead of φ_2. The diagonal of the matrix consists of the probabilities $P(\varphi|\varphi)$ that a phone is correctly recognized. The probability of confusion between two 3-grams t and u is roughly estimated by:

$$P(u|t)=\prod_{i=1}^{N}P(\beta_i|\alpha_i) \quad (7)$$

where α_i and β_i are the ith phones of t and u respectively.

In particular, $P(t|t)$ is an estimation of the probability that 3-gram t has been correctly recognized. These probabilities can be used as weights in the calculation of the relevance score. Equation (5) is rewritten as follows:

$$S_{conf}(q,d)=\sum_{t\in Q}P(t|t)\cdot q(t)\cdot d(t) \quad (8)$$

Reliably recognized terms thus receive higher weights. We proposed to refine Equation (8) as follows:

$$S_{Exp}(q,d)=\sum_{t\in Q}P(u_t|t)\cdot q(t)\cdot d(t) \quad (9)$$

where

$$u_t=\begin{cases}1 & \text{if } t\in Q\cap D\\ \arg\left[\max_{t'\in D}P(t'|t)\right] & \text{if } t\in Q\cap D\end{cases} \quad (10)$$

Contrary to the definition of equation (8), the terms t that are present in Q but not in D are taken into account. In that case, we first determine D in which term t' has the highest probability of confusion with t. In the calculation of the relevance score, we then use $d(t)=1$ instead of $d(t)$ (zero in this case) and $P(t'|t)$ instead of $P(t|t)$.

III. Experiment

Experiments have been conducted with video material of twenty broadcast news sequences from public British TV with a total duration of ten hours. They consist of 125 stories. To form a set of evaluation queries, we chose 20 news keyword queries.

The anchorperson detection was evaluated in terms of well-knowm *recall* (RCL), and *precision* (PRC). To evaluate the ranked list by phone-based spoken document retrieval, we evaluated the retrieval performance by means of a single performance measure, called *mean average precision* (mAP), which is the average of precision values across all recall points. A perfect retrieval system would result in a mean average precision of 100% (mAP = 1).

First, we evaluated the effect of the anchorperson detection. Table 1 shows results for three approaches.

Table 1. Results for AP Detection.

Methods	Anchorperson (AP) segments	
	PRC	RCL
KL2	0.94	0.80
GLR	0.95	0.82
BIC	0.95	0.81

All three distances reached the same results in terms of precision and recall in combination with the average linkage method. The other cluster methods achieved less accurate detection results. As the BIC and KL2 measures are slightly more sensitive to the cutting threshold, our technique uses only GLR and average linkage for the cluster analysis of the studio segments with a global common threshold for dendrogram cutting.

The table 2 represents the mAP values obtained with four different retrieval methods for each of the 20 topic keywords used as queries.

Table 2. mAP values (%) obtained with different retrieval methods.

Retrieval methods	Phone N-grams	
	N=3	N=(1,2,3)
Vector Space Model	41.93%	46.31%
Expansion	52.83%	53.63%

The baseline performance was obtained using trigrams as indexing terms and relevance scores computed as described in equation (5). The second mAP value results from the combination of monogram, bigram and trigram indexing terms described in equation (6). The third performance measure corresponds to use of trigrams with the expansion technique presented in the equation (9). The last case combines the 2 previous ones. The combination of 3-grams increases the mean average precision from mAP=41.93% to mAP=46.31%. The expansion method brings much better results. It outperforms the baseline system as well as the 3-gram combination approach. This expansion method cpmpensates for certain recognition errors by taking into account some document indexing 3-gram terms that, although not contained in the query, are closed to them in terms of confusion probability. Even in the case of poorly performimg queries (e.g. the short, three-phone-long query), the retrieval performance is significantly improved in comparison to both base line and 3-gram combination approaches. The combination of the two previous approaches yields a mean average precision of 53.63%. This represents only a very slight improvement (+0.8%) over the average retrieval performance obtained with the expansion technique alone (mAP=52.83%). Applying the docuemt expansion approach to the 3 sets of indexing terms (1-, 2- and 3-grams) simultaneously can impair the retrieval efficiency in some cases.

Although our search accuracy is below 54%, these experiment results show that sub-word units are able to capture enough information to perform effective retrieval, provided that adequate expansion techniques are applied in order to compensate for the high phone error rate of the indexing engine. According to the experiment, we confirm that phonetic search tends towards low miss rates with many false alarms, while word-level search towards few false alarms but high miss rates. This gives rise to a hybrid approach.

IV. Conclusions

This paper described a indexing and retrieval system of broadcast news using audio content analysis. The system is based on anchor person detection and phone-based spoken document retrieval. Using GLR and hierrahical

cluster analysis of the detected studio segments the anchorperson technique obtained highly accurate results. The method for spoken document retrieval used a simple vector space model with phone 3-grams as indexing units. In order to compensate for the inaccuracy of the phone recognition system, we proposed a technique to expand the representation of documents by menas of phone confusion probabilities. Compared to the baseline system, it improved the retrieval performance by about 11% on average.

Further work should prove the systemon a larger amount of common used news test videos. The more consequent usage of the compressed domain data can increase the computational efficiency. The effects of the N-gram length should be further studied in the future. As query lengths have an impact on the retrieval performance, another perspective would be to consider different N-gram lengths for different query lengths.

References

1. W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, "Integrating Visual, Audio and Text Analysis for News Video," Proc. IEEE Int. Conf on Image Processing, 2000, 3, 520~523, 2000.
2. P. Delacourt, and C. J. Wellekens, "A Speaker-Based Segmentation for Audio Data Indexing," Speech Communication, 32, 111~126, 2000.
3. K. Ng, and V. W. Zue, "Subword-Based Approaches for Spoken Document Retrieval," Speech Communication, 32 (3), 157~186, 2000.
4. H.-G. Kim, N. Moreau, and T. *Sikora, MPEG-7 Audio and Beyond: audio content indexing and retrieval*, (Wiley, 2005)
5. TREC, *Common Evaluation Measures*, (Wiley10th Text Retrieval Conference, A-14, 2001)

[Profile]

• **Hyoung-Gook Kim**