An Online Audio Indexing System

Jitendra Ajmera, Iain McCowan and Herve Bourlard

IDIAP, Switzerland
{jitendra,mccowan,bourlard}@idiap.ch

Abstract

This paper presents overview of an online audio indexing system, which creates a searchable index of speech content embedded in digitized audio files. This system is based on our recently proposed offline audio segmentation techniques. As the data arrives continuously, the system first finds boundaries of the acoustically homogenous segments. Next, each of these segments is classified as speech, music or *mixture* classes, where mixtures are defined as regions where speech and other non-speech sounds are present simultaneously and noticeably. The speech segments are then clustered together to provide consistent speaker labels. The speech and mixture segments are converted to text via an ASR system. The resulting words are time-stamped together with other metadata information (speaker identity, speech confidence score) in an XML file to rapidly identify and access target segments. In this paper, we analyze the performance at each stage of this audio indexing system and also compare it with the performance of the corresponding offline modules.

1. Introduction

With the advent of unlimited storage capabilities and the proliferation of the use of internet, it has become necessary to store information in such a way that any part of it can be accessed with minimal keystrokes. Since significant fraction of this data is in the form of audio, it is important to develop techniques necessary for indexing and browsing such data based on its content. The techniques highlighted in this paper are mostly audio segmentation techniques like acoustic change detection, speech/non-speech classification and speaker clustering. These techniques extract characteristic information or metadata which is very useful for such indexing of audio data. These audio segmentation techniques also make useful pre-processing for Automatic Speech Recognition (ASR) system, which is also an integral part of any audio indexing system. For example, identifying non-speech segments in the audio stream and preventing them from recognition would save computation time in ASR as well as result in more meaningful transcription. Moreover, researchers have clearly shown the positive impact of further clustering of identified speech segments in terms of speakers (speaker clustering) on the transcription accuracy [1]. In the present system, all this structural information together with the ASR output is time-stamped and written in the form of an XML file, which can be used to construct highly selective search queries for retrieving specific content from large audio archives.

A number of similar systems have been previously explained in the literature [2, 3]. The speaker segmentation modules in these systems are often based on Log-Likelihood Ratio (LLR) or Bayesian Information Criterion [4] and depend on an adjustable threshold value, which leave the system less robust to unseen data conditions. Moreover, signal processing (cepstral) based methods are generally applied for speech/non-speech classification, which do not correlate directly with the recognizability of a segment for a given speech recognizer and rather classify the signal based on its acoustic behavior.

Recently, we have proposed novel techniques for speaker change detection [5], speech/music discrimination [6] and speaker clustering [7]. These works addressed the three problems individually. However, in an application, such as the proposed audio indexing system, the three modules and their performance are clearly related. Moreover, our earlier work on speech/music discrimination [6] and speaker clustering [7] make use of all the data (hence are referred in this paper as offline) to make global decisions, which is not possible in the proposed online framework where decisions have to be made using the data that has arrived so far.

A block diagram of the audio indexing system is shown in Figure 1. The audio data[1] is first segmented in terms of acoustically homogenous segments in the segmentation module using exactly the technique proposed in [5] and this is further explained in Section 2. Each of these homogenous segments is classified as speech, non-speech or mixture class, where the mixtures are defined in this paper as regions where both speech and non-speech are present simultaneously and noticeably. This is done using the technique based on offline speech/music segmentation framework proposed in [6], however, modified to be used in the present case of online processing. This modules is explained in Section 3. The speech segments are further grouped or clustered together to provide consistent speaker labels in the speaker clustering module. Every speech segment detected by the segmentation module is compared with all the previous clusters using a merging criterion. The decision making or merging criterion used in this module was first proposed in the offline speaker clustering framework [7] and is further explained in Section 4. This also provides an efficient alternative to offline speaker clustering approach [7], where we start from uniformly segmenting the data in large number (heuristically determined) of clusters. Our results (presented in Section 5) show that the performance of the two approaches (online and offline) are very similar, however the advantage of the online approach is a much reduced computational complexity.

This paper analyzes the performance of the audio indexing at every stage of audio segmentation in Section 5. This paper also highlights the differences between the modules in this indexing system and corresponding offline systems and presents a comparative study of their performance.

2. Segmentation

This module segments the audio stream in terms of acoustically homogenous segments, employing exactly the same technique as proposed in [5]. We briefly review this technique here: to

[1] A sequence on characteristic acoustic feature vectors, MFCC

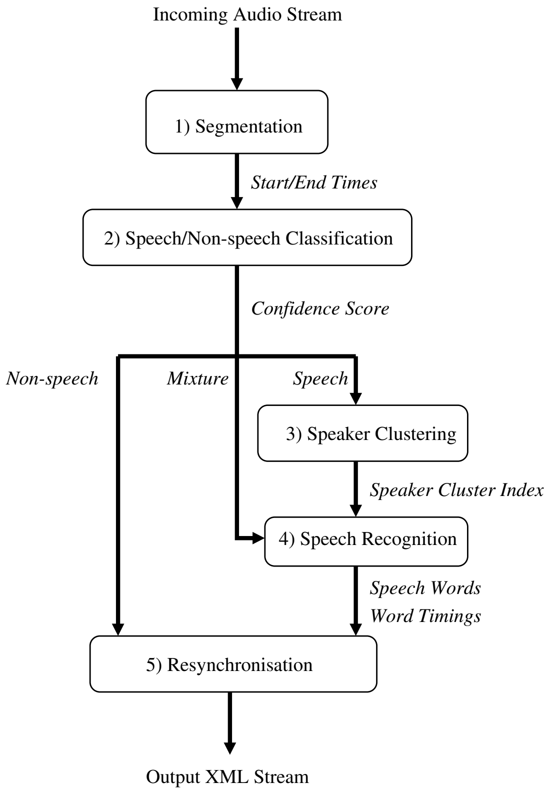

Figure 1: Block diagram of the online audio indexing system.

decide if a speaker change point exists at time t or not, two neighboring windows of relatively small size are considered as shown in Figure 2. The datasets in these windows are denoted as $X = \{x_1, x_2, ..., x_{N_x}\}$ and $Y = \{y_1, y_2, ..., y_{N_y}\}$, where N_x and N_y are the number of data points in the two windows respectively. Let Z denote the union of the contents of the two windows having $N = N_x + N_y$ data points. A decision about a change point at time t is made if:

$$\sum_{n=1}^{N_x} \log p(x_n|\theta_x) + \sum_{n=1}^{N_y} \log p(y_n|\theta_y) \geq \sum_{n=1}^{N_x} \log p(x_n|\theta_z) + \sum_{n=1}^{N_y} \log p(y_n|\theta_z) \quad (1)$$

where θ_x, θ_y are the parameters of Gaussian densities (means and variances) estimated over dataset X and Y respectively. On the other hand, θ_z are the parameters of a Gaussian Mixture Model (GMM) (weights, means and variance) of two Gaussian components.

It was mentioned in [5] that although this technique efficiently finds speaker change points during speech regions, it also results in detecting multiple change points within non-speech regions. This can be explained with the arguments that the "acoustics" during these regions (especially during music) are constantly changing. We expect to alleviate this problem in the next module where the segments are classified as speech, non-speech and mixture classes and two consecutive non-speech or mixture segments are combined back to form one segment, eliminating false alarms or unnecessary segment boundaries.

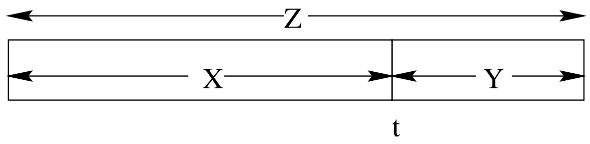

Figure 2: Two neighboring windows with acoustic vector sequences X and Y around time t, when we want to decide if a change point exists or not.

3. Speech/Non-Speech Discrimination

We first present a brief review of the offline technique proposed in [6]. The technique is based on the functioning of an HMM/MLP hybrid ASR system [8] where an MLP estimates the posterior probabilities of the phonemes used in the recognition of a language. We extract *entropy* (H_n) and *dynamism* (D_n) features from these probabilities as follows:

$$H_n = -\frac{1}{N} \sum_{t=n-N/2+1}^{n+N/2} \sum_{k=1}^{K} P(q_k|x_t) \log_2 P(q_k|x_t) \quad (2)$$

$$D_n = \frac{1}{N} \sum_{t=n-N/2+1}^{n+N/2} \sum_{k=1}^{K} [P(q_k|x_t) - P(q_k|x_{t+1})]^2 \quad (3)$$

where $P(q_k|x_n)$ is the posterior probability of k^{th} phoneme q_k, given acoustic feature vector x_n at time n.

These features are used in a two-state (speech and music) HMM, where the emission probabilities of the states of the HMM can be estimated by secondary MLP[2]. A confidence score based on the output of the secondary MLP (real posterior probabilities) was also proposed in [6] which basically provides an indication of "amount" of speech in a given audio segment. This confidence score is computed as:

$$Conf(Speech) = \frac{1}{N} \sum_{n=1}^{N} P(Speech|y_n) \quad (4)$$

where y_n is the two-dimensional feature vector composed of entropy, H_n (2) and dynamism, D_n (3), $P(Speech|y_n)$ is the posterior probability of speech, given feature vector y_n and N is the total number of such vectors in a homogenous acoustic segment.

In the proposed audio indexing system, we use this confidence score to classify segments as speech, music or mixtures. If the value of the confidence score for a segment is above an upper threshold, it is considered clean speech. If the value is below a lower threshold, it is considered non-speech or highly degraded speech. Segments with confidence score between these two regions are classified as mixtures.

For estimating the upper and lower threshold values, performance of a two-class speech/non-speech classifier on a development dataset as a function of a threshold value was analyzed. We note that for a very low value of the threshold value, the frame accuracy for the detection of speech class is very high but the frame accuracy for the non-speech class is poor, resulting in a low total frame accuracy. On the other hand, a very high value of a threshold value would result in very good non-speech

[2]referred to here as secondary MLP to avoid ambiguity with the MLP used for recognition of phonemes.

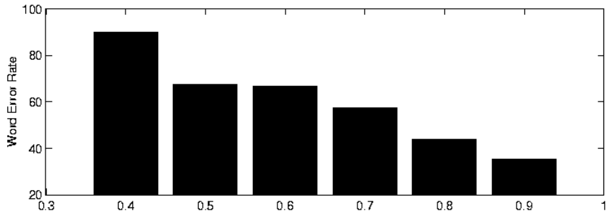

Figure 3: Word error rate as a function of confidence score. Values given are word error rates for all segments in the different confidence score ranges (i.e. $x > 0.9$, $0.8 < x >< 0.9$, ..., $0.4 < x < 0.5$).

accuracy but a poor speech accuracy, again resulting in low total accuracy. However, we note that across a range of threshold values around 0.5 (0.35 to 0.65), the performance of the two-class classifier is relatively stable, indicating mixture like situation.

Two consecutive non-speech or mixture segments are combined to form a single segment and no speaker identity is associated with these segments. In parallel, mixture and speech segments are also sent to an ASR system which generates text transcripts for these segments. While the ASR system is not explained, as it is not the focus of this paper, we present an interesting study (Figure 3) of how the confidence score of each segment (4) relates to the performance of the ASR system in terms of Word Error Rate (WER). Figure 3 shows that the segments with high confidence score (amount of speech) also result in better ASR performance and vice-versa. This can be further useful for practical applications where depending on the task, segments below a particular confidence score can be prevented from recognition. The speech segments are clustered in terms of speaker identities as explained in the following section.

4. Speaker Clustering

We first present a brief overview of the offline speaker clustering algorithm proposed in [7]. This technique is based on a HMM framework where each state represents a cluster and the PDF of each state is modeled by a GMM. The parameters of the PDF are trained via the Expectation-Maximization (EM) algorithm. Starting from a large number of clusters, most similar clusters are found according to (5) and merged in successive iterations. Since this merging criterion does not involve any threshold or heuristics, the algorithm is stopped when there are no more clusters left for merging according to (5).

Since, we do not have access to whole data to make global decisions in the present case, the clustering algorithm is modified in the following way: the latest identified segment is compared with previously existing speaker clusters according to the following merging criterion.

Let $C_i, i = 1, .., K$ be previously defined clusters with datsets D_i and parameters of the PDF θ_i. Let the latest segment be denote as C_j having dataset D_j and PDF parameters θ_j. For the purpose of deciding if C_j belongs to any of $C_i, i = 1, .., K$ (that is if the two clusters should be merged), we hypothesize a cluster C having dataset $D = D_j \cup D_i$ and model the PDF of this cluster by number of parameters equal to the sum of the parameters in the individual clusters C_i and C_j. For example, if clusters C_i and C_j are modeled by GMMs with M_i and M_j number of Gaussian components, respectively, we model the PDF of C by a GMM having $M_i + M_j$ number of Gaussian components. We denote the parameter set of PDF of cluster by θ. The merging criterion to decide if clusters C_j, C_i can be merged (i.e. if C_j should be given the identity of a previous speaker), we employ a merging criterion first proposed in [7] and as follows:

$$\log p(D|\theta) \geq \log p(D_i|\theta_i) + \log p(D_j|\theta_j) \tag{5}$$

If this segment cannot be merged with any of the previously found clusters, a unique speaker identity is assigned to this segment and a new cluster is created.

This solution can also be looked as another way of agglomerative clustering, where we start from number of clusters equal to the number of segments identified by the segmentation module. This is different from the offline scheme presented in [7], where we start from a large number of heuristically determined clusters and uniformly assigning equal amount of data to each cluster individually. Considering this, we expect the online clustering to perform better as we start from acoustically homogenous segments. However, we note that in this case, we compare the latest segment with only the clusters created so far as opposed to the offline scheme, where a cluster is compared globally with all the clusters. This is going to result in much lower computational complexity compared to the offline scheme but at the same time may lead to an inferior performance compared to offline scheme.

5. Experiments and Evaluation

For the purpose of evaluating the performance of the proposed system, we used a BBC broadcast news dataset of duration 1510 seconds. The data has been labeled in terms of time tags of speaker changes, and non-speech segments (marked as "excluded_regions"). The duration of these non-speech segments is 360 seconds.

Segmentation module was evaluated in terms of F-measure as follows:

$$F = \frac{2 \cdot RCL \cdot PRC}{RCL + PRC} \tag{6}$$

where RCL and PRC are recall rate and precision rate, respectively, calculated as:

$$RCL = \frac{correct\ system\ changes}{total\ reference\ changes} \tag{7}$$

$$RCL = \frac{correct\ system\ changes}{total\ system\ changes} \tag{8}$$

where "system" refers to the output of the segmentation module and reference refers to the groundtruth.

The performance of speech/non-speech discrimination is evaluated in terms of speech, music and total frame accuracy which is the percentage of speech, music and total frames that are classified correctly by the system.

The performance of speaker clustering is measured in terms of K-measure, introduced in [9]. First we define: n_{ij}: Total number of frames in cluster i spoken by speaker j; N_s: Total number of speakers; N_c: Total number of clusters; $n_{.j}$: Total number of frames spoken by speaker j; $n_{i.}$: Total number of frames in cluster i; and N: Total number of frames.

Using these notations, the Average Speaker Purity (asp) and Average Cluster Purity (acp) are defined as:

$$asp = \frac{1}{N} \sum_{j=1}^{N_s} \{ \sum_{i=1}^{N_c} n_{ij}^2 / n_{.j}^2 \} \cdot n_{.j} \tag{9}$$

Module	Metric	Online Performance	Offline Performance
Segmentation	PRC	0.55	0.46
	RCL	0.87	0.89
	F	0.67	0.61
Speech/ Non-Speech Accuracy	Speech	91%	99%
	Music	79%	82%
	Total	88%	95%
Speaker Clustering	asp	0.79	0.85
	acp	0.66	0.61
	K	0.72	0.72

Table 1: Summary of performance for different modules of the proposed online audio indexing application. The table also presents the performance of corresponding offline system where all the data is available for processing and hence global decisions can be made.

$$acp = \frac{1}{N}\sum_{i=1}^{N_c}\{\sum_{j=1}^{N_s} n_{ij}^2/n_{i.}^2\} \cdot n_{i.} \tag{10}$$

K-measure is then calculated as:

$$K = \sqrt{acp \cdot asp} \tag{11}$$

Using these evaluation criteria (6 - 11), the performance of the proposed online and corresponding offline systems for different modules are summarized in Table 1.

There are 44 true reference change points (from one speaker to another and from speech to music) in the test data. The Offline system detects a total of 85 changes. Out of these, 39 changes correspond to the true reference change points, resulting in 46 insertions. In our analysis we found that most of these insertions were made during non-speech regions. In the online system, as expected, many of these insertions were eliminated by the following speech/non-speech and speaker clustering modules, resulting in an improved precision rate (from 0.46 to 0.55).

In order to assess the performance of the system in terms of two-class (speech and non-speech) classification problem and also to compare this with that of offline system, all the segments with confidence score (4) above 0.5 were considered as speech and non-speech otherwise. Table 1 shows that the performance of the proposed system is inferior to the corresponding offline system. However, we note that the proposed system in fact classifies the data in terms of three classes, speech, non-speech and mixtures. Totally 181 seconds of audio data was classified as mixtures out of which 87 seconds corresponded to speech and 94 seconds corresponded to "excluded_regions" in the reference. Moreover, considering that mixture segments have speech activity, the proposed system also tries to recognize these segments. If we consider mixture segments as speech, the speech accuracy of the proposed system is 98% indicating that we do not discard any speech segment which is very important from information extraction point of view.

The performance of the proposed system for the speaker clustering task is very similar to the offline algorithm presented in [7]. This suggests that online clustering done in this way is a good alternative to the offline speaker clustering in [7], where we start from uniform initialization in large (heuristically determined) number of clusters. This is even more advantageous because, as mentioned earlier, we note that the computational complexity of this online clustering is much lower than the offline scheme. However, we note that the clustering in the presented work does not deal with non-speech and *mixture* (degraded speech), whereas it offline system deals with these segments.

6. Conclusions

This paper presented an online indexing system which labels the audio segments in terms of speech, non-speech and speaker identities. The system first segments the audio data in terms of acoustically homogenous segments. These segments are marked as speech, non-speech and mixtures and a confidence score is also computed representing amount of speech in that segment. Speech segments are further clustered in terms of speaker identities. All this metadata information together with the ASR output is properly time-tagged and written in the form of an XML file. This paper explained the functioning of each of the individual modules, highlighted the difference (or modifications) from our recently proposed offline algorithms and compared performance of the two schemes for each module.

Acknowledgements

This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2 and MULTI projects.

7. References

[1] B. Ramabhadran, J. Huang, U. Chaudhari, G. Iyengar, and H. J. Nock, "Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives," in *EUROSPEECH*, 2003, pp. 2589–2592.

[2] I. Magrin-Chagnolleau and N. Parlangeau-Vall, "Audio indexing: What has been accomplished and the road ahead." [Online]. Available: citeseer.ist.psu.edu/570010.html

[3] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, no. 8, August 2000, invited Paper. [Online]. Available: www.bbn.com/speech/docs/papers/Audio-Indexing-Retrieval.pdf

[4] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *IBM Technical Journal*, 1998.

[5] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *IEEE signal processing letters*, to appear. [Online]. Available: ftp://ftp.idiap.ch/pub/reports/2002/rr02-39.ps.gz

[6] ——, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, pp. 351–363, 2003.

[7] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition Understanding Workshop (ASRU)*, 2003, pp. 411–416.

[8] H. Bourlard and N. Morgan, *Connectionist Speech Recognition*. Kluwer Academic Press, 1994.

[9] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 573–576.