Impact of Audio Segmentation and Segment Clustering on Automated Transcription Accuracy of Large Spoken Archives

Bhuvana Ramabhadran, Jing Huang, Upendra Chaudhari, Giridharan Iyengar, Harriet J. Nock

IBM T. J. Watson Research Center,
Yorktown Heights, NY 10598. USA.
{bhuvana,jghg}@us.ibm.com

Abstract

This paper addresses the influence of audio segmentation and segment clustering on automatic transcription accuracy for large spoken archives. The work forms part of the ongoing MALACH project, which is developing advanced techniques for supporting access to the world's largest digital archive of video oral histories collected in many languages from over 52000 survivors and witnesses of the Holocaust. We present several audio-only and audio-visual segmentation schemes, including two novel schemes: the first is iterative and audio-only, the second uses audio-visual synchrony. Unlike most previous work, we evaluate these schemes in terms of their impact upon recognition accuracy. Results on English interviews show the automatic segmentation schemes give performance comparable to (exhorbitantly expensive and impractically lengthy) manual segmentation when using a single pass decoding strategy based on speaker-independent models. However, when using a multiple pass decoding strategy with adaptation, results are sensitive to both initial audio segmentation and the scheme for clustering segments prior to adaptation: the combination of our best automatic segmentation and clustering scheme has an error rate 8% worse (relative) to manual audio segmentation and clustering due to the occurrence of "speaker-impure" segments.

1. Introduction

There has been considerable success in creating technologies and infrastructures to enable access to digital archives in recent years, including projects by Informedia [3] and the National Gallery of the Spoken Word (NGSW) [7]. However, automatic technologies for search and exploration in spoken materials still have relatively limited capabilities, capabilities that must be dramatically enhanced if the full potential of digital archiving is to be realized. The ongoing MALACH project [6] aims to achieve a quantum leap in our ability to access the contents of large, multilingual, spoken archives by advancing the state of the art in automated speech recognition (ASR), information retrieval (IR) and other component technologies, by utilizing the world's largest digital archive of video oral histories collected by VHF[1]. The VHF corpus is an interesting and highly challenging digital archive from a research perspective: unique characteristics of the corpus include unconstrained natural speech, massive quantities of multilingual audio (thousands of hours) and an extensive set of labeled training data.

Efficient access to digital archives requires descriptions of their contents, through some combination of human effort and automation. Given the volumes of information involved and the cost of human labour, we are investigating techniques to automate as much of this process as possible without degrading performance below that achievable using humans throughout. Speech and language technologies form the basis for doing so; more specifically, ASR transcripts will be used in all text processing steps since word-level manual transcription would be exhorbitantly expensive. The automatic transcripts, annotated with confidence scores, boundaries, emotion, etc., will be used for subsequent classification into concept and topic metadata.

Our current large vocabulary speech recognition system requires a segmentation of audio prior to recognition: this is a prerequisite for practical, rather than theoretical, reasons. Firstly, our acoustic models are not robust to the often high level of background noise during interviews, which can cause high insertion errors; automatic identification and (successful) removal of non-speech segments prior to decoding improves performance. Secondly, the decoder benefits from segmentation in two ways: short segments reduce per-segment computational load for our current decoder implementation and eliminating non-speech segments reduces the overall computational load. Thirdly, speaker labelling of segments allows adaptation to be performed on speaker-coherent clusters, which may further improve performance. However, obtaining manual segmentation of audio into coherent speaker turns prior to recognition is time-consuming and expensive[2]. Therefore this paper discusses our progress towards identifying automatic segmentation schemes giving equally good end-result recognition performance. This is not a trivial problem: use of imperfect automatic rather than manual segmentations potentially impacts the recognizer in several ways. Segments may not be linguistically coherent, potentially hindering the language model[3]. Imperfect silence removal may lead to insertion errors in retained silence regions and deletion errors in incorrectly removed regions. Further, automatic segmentation raises the challenge of later automatically grouping (possibly speaker-impure) segments into speaker-specific clusters prior to adaptation [4]; poor automatic groupings may impact gains from speaker adaptation. Automatic schemes for audio segmentation[4] and for segment clustering prior to adaptation thus need careful evaluation to ensure that performance is not degraded below human-annotated performance.

The paper is organized as follows. Section 2 gives an overview of the VHF English corpus and Section 3 reviews IBM's ASR system. Sections 4 and 5 discuss schemes for automating the segmentation and for clustering segments for adaptation. Section 6 presents experimental results. The paper ends with conclusions and possible future work.

[1] VHF, or The Survivors of the Shoah Visual History Foundation.

[2] Though much less so than the exhorbitant costs of manual word level transcription.

[3] Since short-span (eg. trigram) language models are used in our current system, this may not be a serious problem.

[4] See [11] for a survey of audio segmentation schemes based on audio content analysis.

2. VHF English Corpus

VHF was created to record the firsthand accounts of Holocaust survivors, liberators, rescuers and witnesses and to disseminate that information to future generations [6]. This section gives a brief overview of this corpus, comprising a total 180 terabytes of MPEG1 video; see also [9]. Approximately 25000 of the collected testimonies are in English; the average duration of each interview is 2.5 hours. Recording conditions varied widely from quiet to noisy conditions such as background conversations or airplane, wind or highway noise. Human transcribers require an average 8 to 12 hours to transcribe an hour of speech from the English interviews, slightly higher than reports for transcribing spontaneous speech [10], illustrating the difficult speech seen in VHF and emphasising the reasons why at least partially automatic methods will be important if the entire corpus is to be catalogued. The difficulty for humans lies in understanding the unfamiliar names, places, multiple languages encountered during a single interview, age-related coarticulations, and heavily accented speech.

3. IBM Speech Recognition System

The English ASR system uses acoustic models constructed using 65 hours of English interviews from 260 speakers in the VHF corpus. The compressed audio signal from the MPEG1 videos is down-sampled to 16KHz; 24-dimensional Mel frequency cepstral coefficients (MFCC) and 60-dimensional transformed features [9] are then extracted. The 60000 word lexicon, built from existing cataloging information and study of frequency of occurrence of uncommon words, has good coverage of names and places likely to be mentioned during interviews. The language model was built by interpolating the 1.7M words from the MALACH corpus with data from Broadcast News (50M words) and Switchboard (3M words) corpora.

4. Audio Segmentation

Four schemes for automatic segmentation are investigated here. Each interview is divided into several 30-minute tapes. Transcribers have annotated these with speaker turns and organized them into shorter segments. We assume manual segmentation represents a “gold standard”, ie. a first recognition pass using these segments will yield good results. Our goal is then to identify an automatic segmentation scheme giving (at least) comparable first-pass transcription accuracy.

4.1. Speech vs Non-speech Segmentation

We adopt the broad approach of [5], described only briefly here. We train an HMM-based segmentation procedure with two models, one for speech and one for non-speech. Each model is a five-state, left-to-right HMM with no skip states. The output distributions are tied across all states in each HMM, and are modeled with a mixture of sixteen diagonal-covariance Gaussian densities. The segmentation is performed using a log-space Viterbi decoding algorithm that can operate on very long audio. A segment-insertion penalty is used during decoding to control the number and duration of the hypothesized speech segments. After decoding, hypothesized segments are extended by an additional 20 frames to capture any low-energy, unvoiced segments at the boundaries and to provide sufficient acoustic context for the recognizer. The audio features used incorporate information about degree of voicing (computed in a 25-ms frame using a normalized auto-correlation function computed on the mean-removed data [5]) and frame-level log-energy, computed from 25ms, mean-normalized frames of data each weighted using a Hanning window. To incorporate temporal context, 17 frames are spliced together and projected to two dimensions via an LDA+MLLT transformation [5].

4.2. BIC Segmentation

BIC segmentation is one example of a statistical model-based segmentation scheme. In these schemes, a sliding window(s) approach [2] is used where the window size is set equal to the lower bound mentioned above. At a given time index t, three windows are defined: two that are contiguous and have a common boundary at t, and one that encompasses both of these (i.e. a combination of the two). Each window is generally modeled with a multivariate Gaussian density with maximum likelihood parameters. At t, a maximum likelihood boundary decision is made [2] based on the log likelihood ratio comparing the maximal value of the likelihood with the two window model to that of the one window model. However, it has proven more effective to compare the penalized log likelihoods [1], where added to the log likelihood is a term that penalizes the model order. By appropriately defining this penalty, one can generate decisions based on the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Consistent AIC (CAIC), the Minimum Description Length (MDL) principle, and the Minimum Message Length (MML) principle. It has been found that BIC, MDL, and CAIC give the best results and with proper tuning all three produce comparable results (Cettolo and Federico 2000). This work uses BIC as described in [1]; once segmented, silence detectors (for example, simple Gaussian Mixture Models) can be used to identify purely silent segments for removal.

4.3. Audio-Visual Synchrony Segmentation

In the present corpus which comprises interviews of people, we notice the visual part of an interview is predominantly focused on the interviewee. Given this observation, we hypothesize that using the notion of *audio-visual synchrony* [8] we can disambiguate between interviewee speech and interviewer speech. Specifically, the synchrony between audio and video channels will be higher when the interviewee speaks. Based on a previous study of definitions and implementations of audio-visual synchrony measures [8], we adopt a scheme that models audio and video features as locally Gaussian distributions [8]. In order to search for synchronous segments, we observe the audio and video frames in a sliding window of 2 seconds (with a shift of 0.5 seconds) by computing the mutual information between each pixel in the video and the corresponding audio frame (parameterized by 24-dim MFCC coefficients) by modeling them using both individual and joint Gaussian distributions. To get a synchrony score from such a mutual information image, we compute the ratio between the mutual information of the face region and the average mutual information across the entire image. Intuitively, the higher this score the greater the synchrony between audio and video. Figure 1 shows example Mutual In-

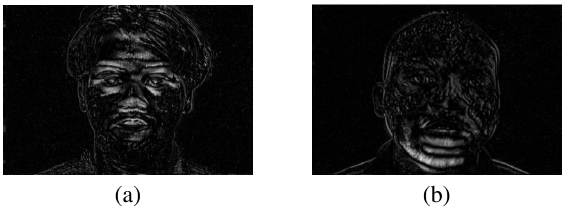

Figure 1: Mutual Information Faces

formation images, where brighter pixels correspond to higher mutual information. Note that face regions are successfully segmented from the background. Our experiments indicate that the face location information from the face detector is error-prone and varies considerably across the video. Hence, rather than rely on the face location estimates from the face detector, we compute a ratio between the best $m \times m$ pixel region in the mutual information plot and the background, where m is chosen empirically from a validation set. The search for the best region begins at the top left pixel and proceeds through the entire image in raster order. To speed up the search, we only consider regions whose center pixel is at least 80% of the maximum mutual information value in the plot. We obtain a synchrony score every 0.5 seconds. This score profile is then used to segment the audio by thresholding the score. In order to make the experiments comparable, we use synchrony score threshold values such that final number of segments is comparable to the other techniques investigated.

This approach has a practical limitation in silence regions, where a mutual information score between audio and video cannot be computed reliably because facial movements are limited and there is no speech audio. One simple solution incorporates additional silence information from audio-only schemes, e.g., the Speech vs Non-speech Segmentation scheme above.

4.4. Iterative Segmentation Scheme

Iterative segmentation begins with some initial set of boundary points, eg. from any scheme above, and then applies a refinement step wherein the defined segments are partitioned into a set of $N(i)$ (i is the iteration number) subsets. The partition is the result of a bottom up clustering. For every segment j, a diagonal covariance GMM model $M(j)$ is built by adapting a background model to the segment data via MAP. The same is done for every pair of segments j,k producing $M(j,k)$. Let $s(j)$ be the score of the data for segment j w.r.t. $M(j)$ and $s(j,k)$ the score of the data for segments j and k w.r.t. $M(j,k)$. Then the bottom up clustering proceeds by joining the two segments j,k that maximize $s(j,k) - s(j)s(k)$ until $N(i)$ sets remain or the criterion max $s(j,k) - s(j)s(k)$ falls below a threshold. Then models are adapted to the $N(i)$ sets. Subsequently, for each time index and each model, a likelihood score is computed over a fading window that captures the neighboring data. Segmentation chooses the label maximizing the likelihood for each frame.

5. Segment Clustering for Adaptation

Adaptation is unsupervised and performed per 30 minute tape; our goal is to produce high accuracy transcripts at this stage since we will ultimately perform a further decoding pass using MLLR adaptation. We adopt as our target the performance given when adaptation is performed on two data clusters independently, one corresponding to interviewer speech and one corresponding to interviewee speech. We use two transforms for the following reasons. Firstly, while we find two transforms improve significantly over one transform, no more than two transforms are used for computational efficiency (important since ultimately thousands of hours of data must be processed) and because gains from multiple transforms are relatively small. Secondly, because background channel conditions are fairly stable throughout each 30 minute segment and therefore our primary goal is to adapt to speaker properties (rather than some combination of speaker and channel). We seek an automatic scheme yielding (at least) comparable performance to this baseline scheme[5]. We use a simple bottom-up clustering scheme, modelling segments by a single multivariate Gaussian and using a Kullback-Leibler divergence metric. The two cluster stage gives the clusters needed for adaptation.

6. Experimental Results

6.1. Data Sets and Usage

Our test set comprises 5 different interviews, each of duration at least 1.5 hours and stored as 30 minute "tapes", for a total 9 hours of data. We select one tape per interview as adaptation data for the whole interview, together with the transcript from first-pass speaker-independent (SI) system decoding of the corresponding audio segmentation: specifically, we adapt a speaker-adaptive-training (SAT) system on a per-interview basis, using unsupervised adaptation with one global feature-space MLLR transform estimated per cluster of segments. Results are assessed using Word Error Rate (WER).

6.2. Audio Segmentation Results

Table 1 shows the WER for each interview for the 4 different segmentation schemes ("Speech vs Non-speech", "BIC", "I-seg" or iterative, "AV" or audio-visual), compared with the results from human segmentation ("Human"). None of these schemes remove silence segments prior to this first pass of decoding. The Speech vs Non-speech segmentation results are surprisingly good, even a little better than those from the human segmentation. We hypothesize this may be because the human transcribed segments tend to be shorter than the automatically produced segments: since the goal of human transcribers was not to segment the data per se, but rather to produce accurate transcriptions, and given that the data often requires repeated plays before it can be transcribed, transcribers often prefer to use short segments. Thus the "human" segmentation, while marking speaker turns accurately, is not always consistent in reflecting natural semantic boundaries, though it is the best (and only) available starting point for assessing recognition accuracy. However, it is possible that short segments break the semantic phrases or sentences and lose context for acoustic and language modeling more frequently than occurs in the Speech vs Non-speech segmentation. We observe the BIC, I-seg and AV schemes also tend to produce shorter segments on average.

	Speaker				
Technique	1	2	3	4	5
Human	30.6	60.6	47.3	56.3	49.2
Speech vs Non-speech	29.2	59.5	46.3	56.1	49.1
BIC	38.7	64.2	50.6	57.9	52.4
I-seg	38.0	61.7	46.3	58.9	53.5
AV	34.9	62.0	46.7	59.0	52.5

Table 1: Pass 1 Decoding WERs

One reason to perform segmentation prior to recognition is to remove unnecessary silence segments, which may cause insertion errors. Table 2 shows the results of removing segments labeled as silence by the Speech vs Non-speech segmentation scheme before recognition; the effect is only marginally beneficial. This trend is unlikely to hold across the entire VHF corpus, where background "silence" is occasionally very noisy.

[5] The reader will observe that an automatic scheme might be expected to beat this baseline, rather than simply match it, if automatically generated clusters better reflect ambient channel conditions: many popular speaker adaptation techniques adapt to some combination of speaker and channel, rather than speaker alone.

	Speaker				
Technique	1	2	3	4	5
No Sil	28.9	59.5	46.3	56.2	49.0
With Sil	29.2	59.5	46.3	56.1	49.1

Table 2: Silent Segment Removal WERs

6.3. Segment Clustering Results

Our baseline is the performance obtainable using manual audio segmentation plus a segment clustering based on human-assigned speaker labels prior to adaptation. In contrast to these manually-marked speaker turns and labels, automatic segmentation schemes yield "speaker-impure" segments eg. the best-performing scheme (Speech vs. Non-speech segmentation) gives 2422 segments of which 65 are speaker-mixed segments (2.5%). Of the speaker-mixed cells, there is an average mix of 14.2% vs 85.8% between speakers. This speaker impurity may reduce the gains expected from clustering based on speaker-id. Our experiments are therefore of two types. The first set of experiments examines whether clustering segments based on human-assigned speaker labels ("Human Segment Ids", i.e. the "true" speaker clustering) is any better (or worse) than using the bottom-up clustering scheme ("BUC") or a random clustering scheme ("Random Speaker Ids"). (In part this experiment investigates whether we will benefit from true speaker adaptation rather than to combinations of speaker and environment.) For these first experiments we start from the reference human segmentations. The results of Table 3 show best performance is obtained when adaptation uses "true" speaker-based clustering of segments, and as expected performance exceeds that achievable using the "Speaker-Independent Only" baseline models (no adaptation) and using a single cluster ("Single Transform") for adaptation.

	Speaker				
Technique	1	2	3	4	5
Speaker-Independent Only	26.4	61.3	44.9	57.6	79.2
Single Transform	23.9	49.2	41.4	44.9	71.4
Human Speaker Ids	22.0	47.3	37.2	44.6	67.6
BUC	22.9	50.2	40.6	46.3	72.4
Random Speaker Ids	28.3	47.9	43.5	57.9	71.0

Table 3: Pass 2 Decoding WERs (Human Segmentation)

In practice we often start from automatically-produced, often speaker-mixed segments, precluding a "true" speaker-based clustering of segments. Further experiments investigate whether this impacts performance achievable using existing automatic clustering schemes (eg. "BUC"). Two comparisons are provided: a clustering of segments based on the dominant speaker using information from the human-assigned speaker ids ("Human Speaker Ids") and random assignment ("Random Speaker Ids"). Table 4 shows that, when starting from speaker-mixed segments, the clustering scheme has relatively little effect on performance. Performance is below that obtained when starting from "pure" speaker segments, particularly when clustering uses the true speaker ids. Although performance exceeds that of "Speaker-Independent Only" baseline models (no adaptation) it is not significantly better than using a single cluster ("Single Transform") for adaptation.

7. Conclusions

Results show that for a single recognition pass, automatic schemes with silence removal give segmentation performance as good (or even better) than expensive manual segmentation. However, for multiple pass recognition incorporating speaker adaptation, the overall best performance is obtained when adaptation uses pure speaker segments and labels rather than the mixed-speaker segments typically derived from an automatic scheme (as we might hope). Preliminary analysis suggests this is because the limited interviewer speech (typically less than 20%, even including crosstalk) is dominated by interviewee speech in the resulting clusters. This impact upon interviewer transcription accuracy would pose a serious problem for cataloging applications since many spoken archives are recorded in the form of interviews, for which interviewer promptings are at least as important as interviewee responses for subsequent information access. Future work must therefore develop a scheme for (a) producing speaker-pure segments or pruning speaker-impure segments and then (b) grouping those segments into speaker-pure clusters prior to adaptation. This could be done directly or via an iterative scheme that refines the initial segmentation based on subsequent speaker information. Future work will also incorporate topic boundaries derived using NLP/IR techniques to refine initial audio segmentations.

	Speaker				
Technique	1	2	3	4	5
Speaker-Independent Only	27.2	60.1	45.6	57.5	79.2
Single Transform	24.4	49.5	41.6	47.9	74.7
Human Speaker Ids	23.9	49.5	41.2	47.9	73.8
BUC	23.9	49.1	41.4	48.4	74.0
Random Speaker Ids	24.1	49.6	41.6	48.5	74.9

Table 4: Pass 2 Decoding WERs (Automatic Segmentation)

8. Acknowledgements

This project is part of an on-going effort funded by NSF under the Information Technology Research (ITR) program, NSF IIS Award No. 0122466. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

9. References

[1] S. S. Chen and P. Gopalakrishnan. Speaker, Environment and Channel Change Detection And Clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[2] H. Gish, M.-H. Siu, and R. Rohlicek. Segregation of Speakers for Speech Recognition and Speaker Identification. In *Proc ICASSP*, volume 2, pages 873–876, Canada, May 1991.

[3] http://www.informedia.cs.cmu.edu.

[4] S. E. Johnson and P. C. Woodland. Speaker Clustering Using Direct Minimization of the MLLR-Adapted Likelihood. In *Proc. Int. Conf. on Spoken Language Processing*, 1998.

[5] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya. Robust Speech Recognition in Noisy Environments: The 2001 IBM SPINE Evaluation System. In *ICASSP*, 2002.

[6] http://www.clsp.jhu.edu/research/malach.

[7] http://www.ngsw.org.

[8] H. J. Nock, G. Iyengar, and C. Neti. Assessing Face and Speech Consistency for Monologue Detection in Video. In *Proc. ACM Multimedia*, 2002.

[9] B. Ramabhadran, J. Huang, and M. Picheny. Towards Automatic Transcription of Large Spoken Archives - English ASR for the MALACH Project. In *Proc ICASSP*, Hong Kong, 2003.

[10] http://www.isip.msstate.edu/projects/switchboard/doc, 2002.

[11] T. Zhang and C.-C. J. Kuo. Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. In *IEEE Transcations on Speech and Audio Processing*, 2001.