Automated Transcription and Topic Segmentation of Large Spoken Archives

Martin Franz, Bhuvana Ramabhadran, Todd Ward, and Michael Picheny

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.
{franzm, bhuvana, toddward, picheny}@us.ibm.com

Abstract

Digital archives have emerged as the pre-eminent method for capturing the human experience. Before such archives can be used efficiently, their contents must be described. The scale of such archives along with the associated content mark up cost make it impractical to provide access via purely manual means, but automatic technologies for search in spoken materials still have relatively limited capabilities. The NSF-funded MALACH project will use the world's largest digital archive of video oral histories, collected by the Survivors of the Shoah Visual History Foundation (VHF) to make a quantum leap in the ability to access such archives by advancing the state-of-the-art in Automated Speech Recognition (ASR), Natural Language Processing (NLP) and related technologies [1, 2]. This corpus consists of over 115,000 hours of unconstrained, natural speech from 52,000 speakers in 32 different languages, filled with disfluencies, heavy accents, age-related coarticulations, and un-cued speaker and language switching. This paper discusses some of the ASR and NLP tools and technologies that we have been building for the English speech in the MALACH corpus. We also discuss this new test bed while emphasizing the unique characteristics of this corpus.

1. Introduction

With recent advances in information technology, digital archiving has emerged as an important and practical method for capturing the human experience. But, before these archives can be used efficiently, their contents must first be described, through some combination of human effort and automation. Automated technologies for the cataloging and indexing of spoken materials presently have relatively limited capabilities; capabilities that must be dramatically enhanced if the full potential of digital archiving is to be realized. The MALACH project seeks to improve the ability to access the contents of large, multilingual, spoken archives by advancing the state of the art in automated speech recognition (ASR), Natural Language Processing (NLP) and other component technologies by utilizing the world's largest digital archive of video oral histories collected by VHF. The unique characteristics of this corpus, including massive quantities of multilingual natural speech and an extensive set of labeled training data, serve to accomplish this goal. In the past, there have been several research efforts, such as Informedia [3], and the National Gallery of the Spoken Word (NGSW) [4], that have focused on the creation of technologies and infrastructures to improve access to digital archives. However, none of these projects have had to address the magnitude and complexity of issues raised by the MALACH archive.

2. Characteristics of the archive

The Shoah Visual History Foundation was created to record the firsthand accounts of Holocaust survivors, liberators, rescuers and witnesses and disseminate that information to future generations [2]. The 52,000 testimonies, with an average duration of 2.5 hours, amount to 180 terabytes of digital video (MPEG1). Approximately 10% of the collection has been manually cataloged with the segment-level description (Table 1), using a domain-specific thesaurus containing 21,000 places and concepts. Names of people and places mentioned in the course of an interview are also stored separately in a database populated from pre-interview questionnaires (PIQ).

The cataloging process is the division of an interview into small segments (three to six minute user-defined as well as one-minute segments) that reflect natural topic boundaries in the interview. For each user-defined segment, the cataloger prepares a short description of the content of that segment and selects appropriate VHF Thesaurus terms to describe experiences and geographical locations associated with that segment. Often thesaurus terms assigned to a segment are not explicitly spoken in the segment. Table 1 provides an example of the words actually spoken in one segment and the corresponding summary and thesaurus terms.

While this kind of extensive cataloging supports search very well, the costs and the language skills needed to catalog multilingual materials impose severe limitations. Thus automation of the cataloging process is absolutely essential if effective access to collections of this scale is needed. The combination of the cataloged data, thesaurus, and the PIQ database with the spontaneous speech from non-professional speakers define the unique features of this spoken archive.

Speech and language technologies are the main sources of information for cataloging this archive. ASR is the basis of all text processing steps. The output of the recognizer, annotated with confidence scores, boundaries, emotion, etc., will be used for subsequent text classification based on VHF's thesaurus. New metadata may be added as more data is processed. In this paper, we will only discuss the ASR and NLP components of MALACH in English. Section 3 and 4 describe the ASR and NLP components while presenting results.

3. Automated Speech Recognition

3.1. Training and Test Data

This section gives a brief overview of the corpus that was created from VHF's archive. More details on the English and Czech automated transcription systems that have been developed to date are presented in [5] and [7] respectively.

The English corpus was generated using 15-minute segments of an interview from 800 randomly selected speakers. Thus, a total of 200 hours of data was selected for manual tran-

Spoken words	It wasn't everybody living in one in one one ghetto you know was a little like the in this street a was a house ghetto in this street it had ghetto but people couldn't people wasn't allowed to go out in the streets when they came in the Nazis came in he wanted they made a Jewish committee the Jewish committee have to help him take where to live and took out the furniture from from the from the Jewish people and so and Jewish committee had eighteen people with me also I helped the Jewish committee I mean the reason is they had eighteen people we walked the street everyday two two people two friends
Thesaurus terms	Conditions under German Occupation, Ghetto Procedures, Jewish Committee
Summary	Recalls living in the ghetto. He tells of the formation of a Jewish Committee. He remembers working for the committee and recalls patrolling the streets. He explains why he tried to prevent Jews from going onto the streets.

Table 1: Spoken words, metadata and summary of a typical segment.

scription that would subsequently serve as training material for ASR systems. The male and female speakers in this corpus were more or less equally distributed and covered a wide range of accents, namely, Hungarian, Italian, Yiddish, German, Polish, etc. It should be mentioned here that this is truly the only corpus of its kind filled with unconstrained natural speech from a wide-variety of accents. The data was recorded under a wide variety of conditions ranging from quiet to noisy conditions such as airplane noise, wind noise, background conversations, highway noise, etc. Human transcribers needed about 8 to 12 hours to transcribe an hour of speech. A significant fraction of the data is obtained under noisy conditions with an energy level below 10 dB. The average speaking rate of the interviewees is 146 words/minute with a dynamic range of 100 to 200 words/minute.

In this paper, we report all results on ASR systems with acoustic models constructed using 65 hours from the 200 hour corpus and language models constructed using both, 65 hours (LM1) and the entire 200 hours (LM2) of the corpus. Two test sets were built on this corpus. The first one (I) consists of 30-minute segments of interviews from 30 randomly selected speakers. This test set will serve as a good representative set for studying ASR performance across speakers [5]. The second test set (II) consists of full testimonies from four speakers (approximately 10 hours of speech). This test set will not only provide vast quantities of data from a single speaker for speaker adaptation purposes, but will also serve as a good test set for retrieval purposes by spanning a wider set of thesaural keywords and providing continuous segments across full testimonies. The ASR results presented in this paper are on both test sets. The document segmentation results are presented on the second test set only.

3.2. Acoustic Modeling

This section describes the ASR system including feature extraction, acoustic models and speech recognition results obtained on the MALACH corpus.

The compressed audio signal from the MPEG1 video files was extracted and down-sampled to 16KHz and subsequently used to produce 24-dimensional mel frequency cepstral coefficients (MFCC) and 60-dimensional transformed features [5] for the acoustic models. The transcriptions contain a good number of foreign words, names, places and sequences of words uttered in a foreign language (such as German, Yiddish or Hebrew) that the transcribers were unfamiliar with. This presents one of the main difficulties of this database. The first step was to clean-up these transcriptions with the aid of the thesaurus, the PIQ and any other related resource. *Mordoh, Schacter, Kerolchikha, Shamway, Juci* represent examples of some of the names that are seen in this corpus. While some of these words could be corrected with the aid of the PIQ and the cataloged information, many of the words required multiple passes at listening to the audio. This is indicative of the difficulties in processing the speech in this archive. The second major difficulty arose from the nature of the speech itself. This corpus consists of elderly speech, where the interviewee's age ranges from 56 years to 90 years. The heavy accents and noisy background combined with poor articulations of phonetic sounds make it difficult for even human transcribers to understand the audio correctly.

	Test Set I WER (in %)
Baseline on Malach	54.3
Baseline + Malach LM1 (A)	51.3
(A) + MLLR +LM2	43.8

Table 2: Word Error Rates on Test Set I

The lexicon for the speech recognizer (60K words) was thus carefully selected to include good coverage on the names and places that were likely to be mentioned during the course of the interview. This was done using the PIQ and cataloging information and studying the frequency of occurrence of uncommon words. 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.

Pronunciations for the many unseen words in this corpus were derived with the help of existing dictionaries and tools using spelling-to-sound rules. In addition to the speaker independent models, we also built speaker adaptive models on this corpus (SAT) using a constrained maximum-likelihood linear regressing (MLLR) [5] transform on the features.

3.3. Recognition Results

The speech recognizer used for this task is described in [5]. Table 2 presents the speech recognition results on this new task. The speaker-independent system built on 65 hours of MALACH data produces a word error rate of 54.3% on this task on Test Set I . When this system is augmented with a language model that has been trained on the MALACH corpus, further improvements can be seen. Subsequent speaker adaptation using SAT and MLLR and an improved language model results in a word error rate of 43.8%.

A testimony (speaker-wise) breakdown of the speech recognition results on Test Set II using the above system showed that the word error rates vary widely across interviews, from 30% to 60%. These automated transcriptions subsequently served as a test bed for the NLP research.

3.4. Research Issues

The ASR research will focus on improving recognition accuracy, and robustness to the varied accents and noisy conditions, topics and spontaneous speaking styles found in the VHF archive. While the ultimate goal of ASR is to produce readable transcriptions, our more immediate goal is to produce transcriptions that support the development of the NLP components and is tightly integrated into their development.

4. Document Segmentation

The purpose of the document segmentation work is to partition the testimonies into shorter (a few minutes long), topically coherent segments. The segments can later be used as retrieval units in search, or to assist catalogers in selecting annotation intervals.

4.1. Statistical Models for Topic Segmentation

For the MALACH project we have extend our previous work done as part of the TDT task [8]. Our segmentation algorithm uses a combination of two probabilistic models, a Decision Tree and a Maximum Entropy model, to compute the probability of a segment boundary occurring at a given clause (sentence) boundary.

The Decision Tree model use a combination of features, including presence of key words and bigrams indicating segment boundaries (indicators are learned automatically from the training data by a mutual information criterion) and features comparing the distribution of nouns on the two sides of the proposed boundary. The Maximum Entropy model uses the feature used by the Decision Tree and individual words, bigrams and trigrams.

4.2. Ground Truth and Scratchpad Boundaries

To experiment with automatic document segmentation, we need to establish a set of segment-annotated data to serve as ground truth for training segmentation models and for judging the system performance. An important component of the VHF archive is a data structure known as *scratchpad.* It is created manually by the catalogers, and contains short summaries and VHF-specific thesaurus terms extracted from the corresponding intervals of the testimonies. The scratchpad interval boundaries are selected to divide the testimonies into short (typically three to six minutes long) blocks, based on the observed topicality changes. Table 1 shows an example of a scratchpad segment. Considering these characteristics, we decided to use the scratchpad segment boundaries to establish the ground truth applied in our segmentation work.

4.3. Training and Test Data

Our training data set consists of 15 minute intervals extracted from 710 testimonies, which represents 177.5 hours of speech. The test set is based on four full testimonies, representing 7.5 hours of speech. To measure the influence of noise introduced by ASR on segmentation performance, we used tree versions of the test set: manually transcibed, and automatically transcribed with WER's equal to 42% and 51%. Using timing information, we aligned the transcribed text with the scratchpad data to obtain the topic-based segment boundaries. Table 3 summarizes the data size statistics.

	Hours	Words	Clauses	segments
Training	177.5	1553914	210497	2856
Test (human)	7.5	58913	7427	168
Test (ASR)	7.5	57152	7772	167

Table 3: Document Segmentation: Training and Test Data Sizes (clause-like intervals were manually annotated)

4.4. Measuring Segmentation Performance

To measure segmentation performance we used an approach similar to the one applied in the TDT segmentation task [9]. The performance measure is based on determining the agreement between computed and reference boundaries, using an interval moved through the segmented data. At each position of the interval we declare correct segmentation if there is both computed and reference boundary or neither computed nor reference boundary found in the interval. Similarly, a *false alarm* is stated if there is a computed boundary and no reference boundary in the interval, or a *miss* is declared if there is no computed boundary and a reference boundary in the interval. In contrast to the technique used in the TDT project, where the above described computation is performed on every word position in the data, we move the interval so that the above described computation is done only with intervals centered at clause boundaries. The interval length is set to ten words.

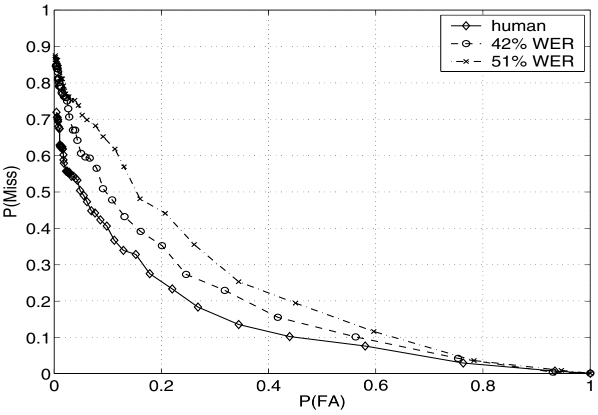

Figure 1: Document Segmentation: Human vs. ASR Transripts

4.5. Segmenting Human and ASR Transcripts

Fig.1 compares segmentation performance on manually and automatically transcribed data. The relative degradation caused by speech recognition errors is quite uniform over a wide range of operating points. Table 4 compares performance degradation caused by ASR errors at the Equal Error Rate operating point. Data with 41% WER suffer only less than half of the performance degradation observed on the data with 52%WER.

4.6. Training Data Size and Segmentation Performance

To investigate the influence of varying training data size on segmentation performance, we created a set of smaller training corpora by selecting in a uniform fashion from the original set of fifteen minute intervals. Fig.2 compares the baseline perfor-

WER[%]	EER[%]	Relative Degradation[%]
Human	23	-
42	26	14
51	30	33

Table 4: Segmentation: Human and ASR Transcripts

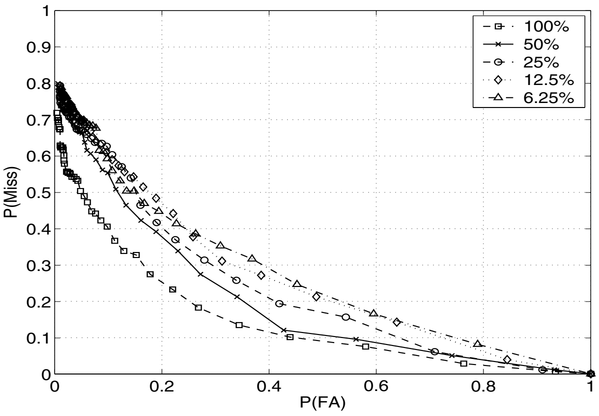

Figure 2: Segmentation Training Size

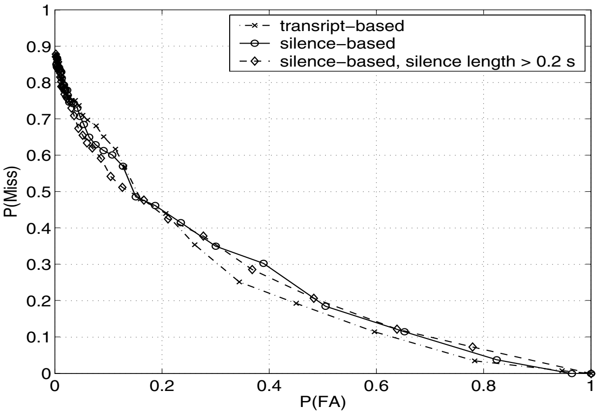

Figure 3: Document Segmentation: Transcript-based vs. Silence-based Clause Boundaries

mance with the results based on gradually reduced training sets. We observe that segmentation performance improves steadily with growing training size, with most of the improvement taking place in the experiments using the larger training sets. This suggest that, with more transcribed data available in the future, we should be able to increase the training data sizes and substantially improve segmentation performance.

4.7. Transcript-based vs. Silence-based Clause Boundaries

In the above experiments we computed the probability of a segment boundary occurring at clause boundaries marked-up in the data by human transcriptionists. For a segmentation system to work in an unsupervised fashion, the clause boundaries have to be established automatically.

Fig. 3 shows segmentation performance of a system where the "clause" boundaries were selected based on the occurrence of silence in the input data. We observe small performance degradation. We also experimented with thresholding the silence length so that a segment boundary is never proposed for a silence shorter then a given threshold. As shown on Fig. 3, the effect is barely measurable improvement in the low False Alarm area and degradation in the low Miss area.

4.8. Speaker Turns as Segmentation Feature

As described earlier in this paper, the VHF documents are based on dialog between an interviewer and an interviewee. Speaker turns are marked up in the manual transcripts and can be detected automatically in the ASR data. Examination of a few interviews suggests that an interviewer's prompt sometimes triggers a change of the topic of the testimony, and motivated us to experiment with using speaker turns as segmentation features. However, including the speaker turn feature has only marginal influence on the performance.

5. Conclusions and Future Work

In this paper, we have reported on the current state of IBM's ASR and NLP components for the English portion of the MALACH project. In the future, we plan to focus on improving these component technologies, which will clearly generalize to related applications, given the difficult and diverse nature of the data.

6. Acknowledgements

The authors would like to thank our partners for the many useful discussions. This work is part of a five-year joint effort between IBM, the Johns Hopkins University, the University of Maryland, and the Visual History Foundation which has been 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.

7. References

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

[2] Gustman, S., et al., "Supporting Access to Large Digital Oral History Archives,", *Joint Conference on Digital Libraries*, Portland, OR, July 15-17, 2002.

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

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

[5] Ramabhadran, B., Huang, J., Picheny, M., "Towards Automatic Transcription of Large Spoken Archives - English ASR for the MALACH project", submitted to ICASSP 2003.

[6] McCarley, J. S. and Franz, M., "Influence of Speech Recognition Errors on Topic Detection", *Proceedings of the 23rd ACM SIGIR Conference on Information Retrieval,* pp. 342-344, 2000.

[7] Psutka, J. et al., "Language Model Data Selection for Czech ASR in the MALACH Project", submitted to ICASSP 2003. award=012246

[8] Dharanipragada, S., Franz, M., McCarley, J. S., Ward, T., Zhu, W.-J., "Segmentation and Detection at IBM: Hybrid Statistical Models and Two-tiered Clustering", in *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Publishing, 2002.

[9] Fiscus, J. G. and Doddington, G. R., "Topic Detection and Tracking Evaluation Overview", in *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Publishing, 2002.