TOWARDS AUTOMATIC TRANSCRIPTION OF LARGE SPOKEN ARCHIVES - ENGLISH ASR FOR THE MALACH PROJECT

Bhuvana Ramabhadran, Jing Huang and Michael Picheny

Human Language Technologies
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598.
{bhuvana, jghg, 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 NSF-funded MALACH project aims to provide improved access to large spoken archives by advancing the state-of-the-art in automated speech recognition (ASR), Information Retrieval (IR) and related technologies [1, 2] for multiple languages. This paper describes the ASR research for the English speech in the MALACH corpus. The MALACH corpus consists of unconstrained, natural speech filled with disfluencies, heavy accents, age-related coarticualtions, uncued speaker and language switching, and emotional speech collected in the form of interviews from over 52000 speakers in 32 languages. In this paper, we describe this new testbed for developing speech recognition algorithms and report on the performance of well-known techniques for building better acoustic models for the speaking styles seen in this corpus. The best English ASR system to date has a word error rate of 43.8% on 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. Before archives can be used efficiently, their contents must first be described, through some combination of human effort and automation. Automatic technologies for search and exploration in 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 make a quantum leap in the 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 unique characteristics of this corpus, including unconstrained natural speech, massive quantities of multilingual audio 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 [4], and the National Gallery of the Spoken Word (NGSW)[5], that have focussed 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 VHF archive.

[1] Also known as The Survivors of the Shoah Visual History Foundation

The objectives of MALACH and various component technologies and their interactions are described in [2]. 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 classification into concepts. In this paper, we will only discuss the ASR components of MALACH in English. Issues related to the components of MALACH in Czech are addressed in [8].

The rest of the paper is organized as follows. Section 2 describes the generation of English training and test corpora from VHF's archive and their characteristics. Section 3 begins with an overview of IBM's speech recognition system and discusses a set of techniques for the construction of acoustic and language models for this corpus. Speech recognition results are reported in Section 4. The paper concludes with a summary and a discussion on future directions of research.

2. CREATION OF THE ENGLISH CORPUS

2.1. Data Description

The Shoah Foundation was created to record the firsthand accounts of Holocaust survivors, liberators, rescuers and witnesses and disseminate that information to future generations [2]. Approximately 25000 of the collected testimonies are in English and 575 are in Czech. The average duration of each interview is 2.5 hours. The entire collection amounts to 180 terabytes of digital video (MPEG1). A parallel paper describes the efforts in building automated Czech transcription systems [8]. Table 2.1 illustrates the spontaneous nature of the speech in this corpus with an example of the actual words spoken during the course of an interview.

2.2. Training and Test Corpora

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 transcription 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. In this paper, we report all results on ASR systems with acoustic models constructed using 65 hours from the 200 hour corpus and

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 ...

Table 1. Example of the spontaneous nature of the speech in the MALACH corpus

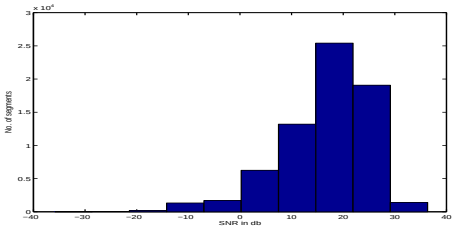

Fig. 1. SNR computed over the training data

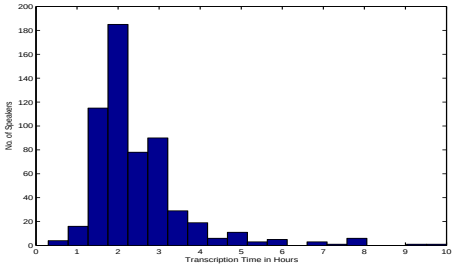

Fig. 2. Manual Transcription Times for 15 minute segments

language models constructed using both, 65 hours and the entire 200 hours of the corpus. At the time this work was performed, all 200 hours had not yet been transcribed. Therefore, we commenced the development of speech recognition systems of this task with 65 hours of training material. The test corpus consists of 30-minute segments of interviews from 30 randomly selected speakers. The results presented throughout this paper are on an hour of data from 20 speakers, selected from this test corpus, unless stated otherwise.

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 using "Transcriber", as the transcription tool [9]. The difficulty lies in understanding the unfamiliar names, places, multiple languages encountered during a single interview, coarticulations related to age, and heavily accented speech. A distribution of transcription times for 15 minutes of an interview is presented in Figure 2. These times are slightly worse than what has been reported in the past for transcribing spontaneous speech [10], illustrating the difficult speech seen here. Table 2 provides more details on the number of names, places and foreign words seen in this corpus. The average speaking rate of the interviewees is 146 words/minute with a dynamic range of 100 to 200 words/minute. The average speaking rate in the SWB corpus is 100 words/minute [10].

Hours	Names and Places (%)	Foreign Words (%)
65	7.2	4.1
200	10.6	5.3

Table 2. Distribution of Names, Places and Foreign Words

3. ASR SYSTEM OVERVIEW

This section briefly describes the IBM large-vocabulary speech recognition system. The various aspects of this system were detailed earlier in [3]. IBM LVCSR systems use context-dependent sub-phone classes and phonetic baseforms. Words are represented as a sequence of phones and each phone is modeled with a 3-state left-to-right HMM. Each of the states roughly correspond to the beginning, middle and end of each phone. A decision tree is constructed for every sub-phonetic unit that corresponds to a state of the three state HMM [3] by querying the surrounding phonetic context. The feature vectors used to parameterize the speech signal are produced at a 10ms frame rate from 16-bit PCM sampled at either 16KHz or 8KHz. The feature vectors at each terminal node (leaf) are modeled using a Gaussian mixture density with each Gaussian having a diagonal covariance matrix. Output distributions on the state transitions are expressed in terms of the rank of the leaves. The systems used in this paper have approximately 3000 leaves and anywhere between 50000 and 300,000 Gaussian distributions. A simple N-gram language model is used to compute the language model probabilities. The decoder is a single-pass decoder which employs the rank-based decoding strategy and the envelope search algorithm [3] to hypothesize a sequence of words corresponding to the utterance.

4. ACOUSTIC MODELING

4.1. Feature extraction

This section describes the construction of the acoustic models for this task. The video interviews were obtained from VHF in MPEG1 format. The compressed audio signal in MP3 format was stored at a sampling frequency of 44.1KHz. This signal was extracted from the video and down-sampled to 16KHz or 8 KHz in accordance with the recognizer that was used. The original recordings were done in stereo with the interviewer and interviewee in separate channels. However, very often, the microphones were placed such that they recorded both the speakers with equal intensity or the interviewee was recorded louder than the interviewer in one channel and vice versa. Also, on many occasions, both the interviewer and interviewee were connected to the same channel, or the interview was conducted using a far field microphone with a noisy background, as reflected by the SNR distribution given in Figure 1. It can be seen that a significant fraction of the data is noisy with an energy level below 10 db. The channel in which the interviewee's speech was the loudest was selected for subsequent processing. While it is important to transcribe the interviewer's questions equally well, we decided to use the channel in which the interviewee was the loudest as the bulk of the data is from the interviewee. In the future, we plan to explore algorithms that will use information from both channels to select the channel that produces the best word error rate for each of the two speakers.

The 16-bit down-sampled PCM signal was used to produced 24-dimensional mel frequency cepstral coefficients (MFCC). The MFCC features were computed from a 24-filter Mel filterbank spanning the 0 Hz - 4.0 kHz frequency range for th 8KHz sys-

tem and 0Hz - 8.0 kHz for the 16KHz system. All feature sets use 25-ms. frames with a 10-ms. step, perform spectral flooring by adding the equivalent of one bit of additive noise to the power spectra prior to Mel binning, and use periodogram averaging to smooth the power spectra. Every 9 consecutive cepstral frames are spliced together and projected down to 60 dimensions using a linear discriminant feature space transformation to ensure maximum phonetic discriminability. The range of these transformations is further diagonalized by means of a maximum likelihood linear transform (MLLT) to decorrelate dimensions.

4.2. Acoustic Segmentation

The data obtained from each interview was organized in 30-minute interview segments. In the process of transcription, the transcribers also annotated the corpus with speaker turns and organized the corpora into smaller segments. Although, this was done to make the transcription of this difficult speech easier, these turned out to be very useful segments for bootstrapping initial acoustic models and subsequently for exploring various automatic acoustic segmentation algorithms. Throughout this paper we report word error rates on acoustic models constructed using the manual segmentations.

4.3. Acoustic Models

The first step in the construction of acoustic models is the construction of the decision trees to model context-dependent variations of this speech. The trees are built from Viterbi alignments of the speech signal in the training data with the manual transcriptions at the context-dependent state level. The initial transcriptions that we used had a fair number of transcription errors. Many of these errors were due to the good number of foreign words, names, places and sequences of words uttered in a foreign language (such as, German, Yiddish or Hebrew) that the transcriber was unfamiliar with. The percentage distribution of foreign words and names in this corpus is given in Table 2. The clean-up of these transcriptions was aided by the use of a thesaurus obtained from VHF that contained frequently used names and place names and a pre-interview questionnaire. Many of these place names constitute cities, streets and names of concentration camps. This presents one of the main difficulties of this database.

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, noisy backgrounds that include airplane, road, construction and wind noise, and background conversations combined with poor articulations of phonetic sounds make it difficult for even human transcribers to understand the audio correctly. In order to obtain initial alignments, the average log-likelihood of each segment in the training data conditioned on the alignments was used to reject the segments that had transcription errors and or incorrect pronunciations in the lexicon. Pronunciations for the many unseen words in this corpus were derived with the help of existing dictionaries and tools using spelling-to-sound rules. The data at the leaves of the decision were modeled with Gaussian distributions via a BIC-based procedure [6] and trained using multiple iterations of the EM algorithm.

In addition to the speaker independent models, we also built speaker adaptive models on this corpus (SAT). The training was done via a feature space maximum likelihood linear transforms, i.e. fMLLR, for each training speaker. The canonical model was first initialized as the speaker independent model. After fMLLR transforms for training speakers were computed against the canonical model, the canonical model was then re-estimated using the affinely transformed features. This method is based on the SAT [11] principle, but differs slightly from SAT in that the normalization is applied to the features. This corresponds to using a constrained maximum-likelihood linear regressing (MLLR) [12] transform instead of a mean-only MLLR transform.

LM Corpora	Perplexity	WER (%)
SWB + BN (LM0)	180	57.3
65 hours of MALACH interpolated with SWB and BN (LM1)	95.1	54.1
200 hours of MALACH	86.9	53.3
200 hours of MALACH interpolated with SWB and BN (LM2)	72.3	53.1

Table 3. Perplexity and Word Error Rates for various Language Models

4.4. Language Modeling

Two language models were trained on both 65 hours and 200 hours of data from this corpus. The technique that was used to compute and smooth the n-gram counts was the modified Kneser-Ney algorithm [13]. A challenge in language model training and lexicon design for this corpus was that a large portion of personal names and places were not covered by this data. To enhance these language models, data from the Switchboard and Broadcast News corpora were added and interpolated with the data from the MALACH corpus. The interpolated weights were optimized to achieve minimum perplexity on the held-out data from the MALACH corpus. The effect of an increase in the in-domain material and the interpolation across other speaking styles such as those seen in Broadcast News (BN) and Switchboard (SWB) tasks are illustrated in the Table 3. The 65-hour and 200-hour MALACH corpora contain about 320K and 1.7M words respectively. The BN and SWB corpora contain 158M and 3.4M words respectively. The decoding lexicon consists of 30K words. The average OOV rate on the test set ranges from 3.2% to 11.6% with an average OOV rate of 8.2%. The percentage of trigram counts used from the SWB and BN corpora decreased from 66% to 26% with the addition of more in-domain data from the MALACH corpus.

5. RECOGNITION RESULTS

This section presents the first set of recognition results on the English portion of the MALACH corpus. A baseline word error rate was first computed using a speaker independent, MFCC system, which served as the very first baseline system for the SWB task [6]. This system comprising of 300K diagonal Gaussians and a lexicon of 64K words was trained on the Switchboard, CallHome, CTIMIT, and the National Cellular Corpora and has an error rate of 47.3% on the spontaneous conversations in the Switchboard 1998 Evaluation task. However, the speaker independent performance of this system was as high as 85.6% on this test data (Table 4) compared to the baseline system trained on in-domain data. It is interesting to note that even though both corpora are recorded conversations between two individuals, the drastic differences in the nature of the speech coupled with the high OOV rate results in poor speech recognition performance.

	System I (8 Khz)	System II (16 Khz)
SWB system + LM0	85.6	NA
Baseline on Malach + LM0	57.3	54.3
Baseline + Malach LM1	54.15	51.3
SAT + LM1 (A)	49.5	46.8
(A) + MLLR	–	45.4
(A) + MLLR +LM2	46.1	43.8

Table 4. Word error rates using two different systems

It can be seen from Table 3 that both, increasing the in-domain data as well as incorporating data from other corpora produces a reduction in perplexity and the overall word error rates. A gain of 4% absolute can be obtained when the in-domain is tripled and augmented with similar data from other related corpora. Table 4 presents the speech recognition results on this new task. The two systems were built by down-sampling the original signal to 8KHz (System I) and 16KHz (System II). While both systems are comparable in performance, the wider bandwidth system has a relative 6% performance improvement over the bandlimited system.

The speaker-independent system built on 65 hours of MALACH data produces a word error rate of 57.3% on this task. This reiterates prior work in the literature that significant improvements, such as halving the word error rate can be obtained when the acoustic models are trained using in-domain data. When this system is augmented with a language model that has been trained on the MALACH corpus, further improvements can be seen. The SAT models reduce the error rate further to 46.8 %. Subsequent adaptation using MLLR and an improved language model results in a word error rate of 43.8%.

6. SUMMARY AND ANALYSIS

As demonstrated in Section 4, training ASR systems with in-domain data and conventional adaptation techniques improve the performance significantly, bringing the word error rate down to 43.8%. This high word error rate, despite matched training and test conditions, illustrates the difficult and diverse nature of the VHF corpus. In this section, we will analyze the many factors that contribute to the degradation in performance.

In order to isolate the errors made by the acoustic and language models, an experiment was conducted, wherein, two native speakers were asked to read an hour of transcripts from the test set. The word error rate from an ASR system trained on broadcast news data (without any MALACH data in the acoustic or language models) on this task was 11.89%, clearly implying that most of the errors made by the ASR systems could be attributed to the acoustic models. OOVs were a major source of errors, introducing many insertion and substitution errors. The disfluencies present in this spontaneous speech pose significant challenges for the recognizers as well. There are sections of frequent interruptions by the interviewer, sometimes to assist the interviewee along, and these rapid speaker changes and cross talk pose problems for manual and automatic segmentation methods. There are also sections of with emotional, low-volume and whispered speech. All of these contribute to the overall word error rate.

7. RESEARCH ISSUES AND FUTURE WORK

Despite the considerable progress made in recent years in speech recognition, current technologies are sensitive to the acoustic and channel properties of the data, speaker variability and to mismatches between the training and real usage conditions, as can be seen from Section 4. Due to the nature of the material, a significant percentage of this corpus is comprised of highly emotional and sometimes whispered speech. Hence, modeling the disfluent, emotional and whispered speech from elders is crucial. Background noise and frequent interruptions pose problems for adaptation to the interviewee's speaking style. Very often, the speakers switch naturally to their native language or the languages they are used to, especially when describing cultural events and these uncued switches pose challenges to ASR systems. All the speakers in the corpus are non-native speakers with dialects from all over the world and hence pronunciations that capture heavily-accented speech need to be derived. OOVs (names, places, events) pose serious problems as well. The huge quantities of audio data renders this corpus useful for studies on the effect of the size of the adaptation data on the word error rate and optimal selection of data for training and adaptation. These are some of the issues that will be addressed in subsequent years, while efforts will be spent to reduce the overall error rates on this task using well-known techniques such as VTLN and MMI that we have not currently explored.

Acknowledgements

The authors would like to thank our partner, Josef Psutka from University of West Bohemia for his suggestions in setting transcription standards and transcription tools. This work is part of a five-year joint effort between IBM, JHU, UMD and VHF 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.

8. 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] Bahl, L.R., et al., "Robust Methods for using context dependent features and models in a continuous speech recognizer," *Proceedings of the ICASSP*, Vol. I, pp. 533, 1994.

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

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

[6] Huang, J. et al., "Large Vocabulary Conversational Speech Recognition with the Extended Maximum Likelihood Linear Transformation (EMLLT) model", *ICSLP*, pp. 2597-2600, Sept. 2002.

[7] Leggetter, C. J., et al., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, No. 2, pp. 171-186, 1995.

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

[9] http://www.etca.fr/CTA/gip/Projets/Transcriber

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

[11] Anastasakos,T., et al., "A compact model for speaker-adaptive training," *ICSLP*, 1996.

[12] Gales, M. J. F., "Maximum likelihood linear transformations for HMM-based speech recognition," *Tech. Rep. CUED/F-INFENG/TR291* , Cambridge University Engineering Department, 1997.

[13] Chen, S. F., et al., "An empirical study of smoothing techniques for language modeling", *Computer Speech and Language*, Vol. 13, No. 4, pp. 359-393, 1999.