Speech Recognition Error Analysis on the English MALACH Corpus

Olivier Siohan *Bhuvana Ramabhadran* *Geoffrey Zweig*

IBM T.J. Watson Research Center
1101 Kitchawan Rd., Rte 134/PO BOX 218
Yorktown Heights, NY 10598, USA
{siohan,bhuvana,gzweig}@us.ibm.com

Abstract

This paper presents an analysis of the word recognition error rate on an English subset of the MALACH corpus. The MALACH project is an NSF-funded research program related to the development of multilingual access to large audio archives. The archive of interest is a large collection of testimonies from 52,000 survivors, liberators, rescuers and witnesses of the Nazi Holocaust, assembled by the Shoah Visual History Foundation. This data has some unique characteristics that make it quite unusual in the speech recognition community such as elderly speech, noisy conditions, heavily accented speech. Hence, it is a challenging task for automatic speech recognition (ASR). This paper attempts to identify the factors affecting the ASR performance on that task. It was found that the signal-to-noise ratio and syllable rate were two dominant factors in explaining the overall word error rate, while we observed no evidence of the impact of accent and speaker's age on the recognition performance. Based on this evidence, noise compensation experiments were carried out and led to a 1.1% absolute reduction of the word error rate.

1. Introduction

The MALACH (Multilingual Access to Large Audio arCHives) project is an NSF-supported effort between the Survivors of the Shoah Visual History Foundation, IBM, Johns Hopkins University, the University of Maryland, Charles University (Prague, Czech Republic) and the University of West Bohemia (Prague, Czech Republic) to develop technology allowing multilingual access to large spoken archives [1]. The project fosters research in several areas, notably including automatic speech recognition (ASR) and natural language processing (NLP) techniques for automated creation of metadata for document retrieval purposes. The archive used in this project is a unique collection assembled by the Shoah Visual History Foundation, the world's largest coherent archive of videotaped oral histories. The collection consists of 116,000 hours of digitized interviews in 32 languages from 52,000 survivors, liberators, rescuers and witnesses of the Nazi Holocaust.

This paper focuses on the some of the issues related to ASR on a subset of the English part of this archive. In Section 2, we first describe some of the characteristics of this data. We then study in Section 3 the different factors affecting the word error rate of our ASR system, and carry out statistical analyses to identify the dominant factors contributing to the relatively poor ASR performance on this data, compared to other tasks such as broadcast news or conversational speech recognition. Similar analyzes have been carried out on other tasks (e.g. Switchboard) in the literature [2, 3]. The unique aspect of the MALACH domain is that the sources of variability are much larger than in other corpora and include heavy accent, acoustic noise, elderly and disfluent speech, all at the same time. As the signal-to-noise (SNR) ratio is identified to contribute to the overall performance, we carry out in Section 4 some noise compensation experiments to attenuate the impact of the background noise on the recognition performance. Last, we suggest in conclusion some future research directions to improve ASR performance on this difficult data.

2. Characteristics of the MALACH archive

The MALACH archive presents numerous challenges for speech recognition systems. The database was collected by video-recording testimonies in each of the interviewees' home. The acoustic environment was not fully controlled and is typical of in-the-field recordings, as opposed to studio conditions, and presents a highly variable level of background noise across testimonies. Figure 1 represents the distribution of the segmental signal-to-noise (SNR) ratio on a 65 hours subset of English testimonies, derived using the Mississippi State tool [4]. It clearly appears that it is a highly skewed distribution with a long tail towards low SNR and an average SNR of 23dB.

Another characteristic of this data is the distribution of interviewees' age. Given the nature of the data, interviewees were all in their early 60's to late 90's at the time the data was collected. Clearly, this is a much older population with a wider age range compared to the data commonly available in the speech community, such as Broadcast News or Switchboard. The age distribution is shown in Figure 2. Existing results in the literature suggest that ASR systems built on adult speakers may perform poorly on elderly speech [5, 6, 7, 8], leading system designers to explicitly build acoustic model for elderly speakers. Our setup falls into such an approach and can be considered as a multi-style training, with relatively matched training and test conditions. However, we will further study whether age variability within the elderly population impacts the recog-

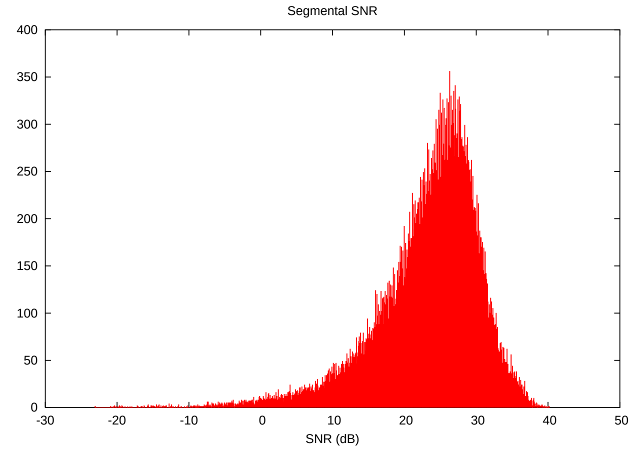

Figure 1: Distribution of the segmental SNR on 65 hours of English testimonies.

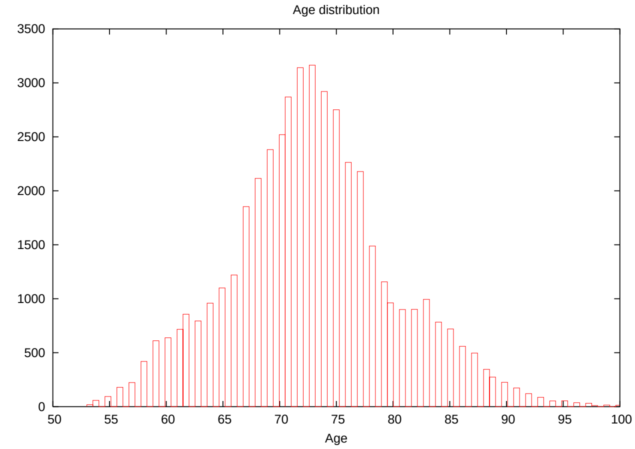

Figure 2: Age distribution.

nition performance.

We also computed the average speaking rate in terms of syllable per seconds on about 200 hours of English testimonies for 265 speakers. The audio was aligned against its transcription and the number of syllables per second over a speech "spurt" (between pause speech segment) was computed. The NIST syllabification software [9] was used to generate the syllabic transcription. The distribution of the syllabic rate is shown in Figure 3, which exhibits an average syllabic rate of 4.0 syllable/second. In contrast, the average syllable rate on Switchboard was found to be close to 4.8 syllable/second. As the literature suggests that the speaking rate can have a significant effect on the word error rate (WER) [10, 11], we will study whether any correlation can be observed between speaking rate and WER.

While these testimonies were recorded in English, English was often not the native language of the interviewees. Indeed, a very large portion, if not all of the database can be considered as non-native speech. In an attempt to correlate the WER to the degree of accent, we labeled some of the data into some subjective accent classes, labeled as"light", "medium" and "heavy" accent, as evaluated by our transcriber. Again, we will study whether the degree of accent can affect the WER.

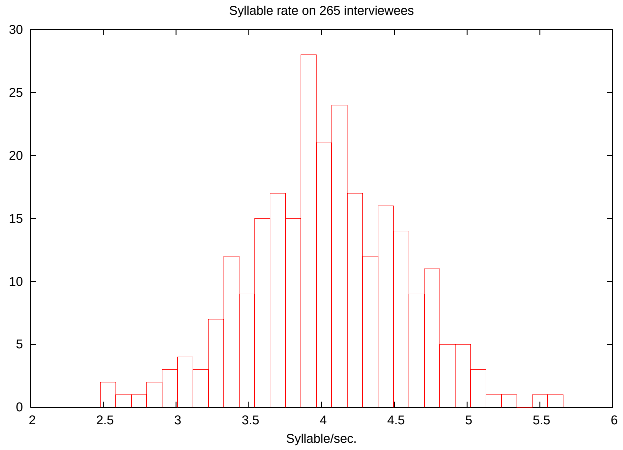

Figure 3: Syllable rate distribution across 265 interviewees. English testimonies.

3. Word Error Rate Analysis

3.1. ASR setup

We used 65 hours of transcribed testimonies to build a set of acoustic models. The setup was similar to the one described in the English acoustic modeling section in [1]. The audio signal was down-sampled to 16kHz, and 24-dimensional MFCC features were computed. Every 9 consecutive MFCC frames were stacked together and projected down to a 60-dimensional space using Linear Discriminant Analysis, followed by a maximum likelihood linear transform [12] (MLLT) to decorrelate the feature components. A speaker independent model consisting of about $3,000$ states was built using a decision-tree state tying procedure for a total of $40,000$ Gaussians. The recognition was carried out using a $30K$-word vocabulary.

Two different test sets were used for our experiments. The first one consists of 943 utterances extracted from a set of 20 testimonies, for a total of about 2 hours of data. The second one was much larger, consisting of about 62 hours of data, and was required to get statistically significant results when correlating WER and speaker characteristics. The average WER was 45.8% on the small test set, and 38.9% on the large one.

Given the difficulty of this task, due to highly variable acoustic conditions, elderly and emotional speech, non-native speakers, it is important to identify which sources of variability contribute the most to the ASR performance.

3.2. Analysis of the WER

In an attempt to identify factors affecting the word error rate, we conducted a statistical test of hypothesis to study whether the WER is correlated with SNR, speaker age, and syllable rate. A test of hypothesis was constructed with the null hypothesis, $H0$, defined as $H0 : R = 0$, where R is the

WER vs. SNR
$R = -0.3989 \quad t = -5.166 \quad DF = 141$
$p(\lvert t\rvert >= 5.166) = 8.282e-07$

WER vs. Syllable rate
$R = -0.2292 \quad t = -2.806 \quad DF = 142$
$p(\lvert t\rvert >= 2.806) = 0.005727$

WER vs. age
$R = 0.0707 \quad t = 0.9327 \quad DF = 173$
$p(\lvert t\rvert >= 0.9327) = 0.3523$

Table 1: Level of significance for the test $H0 : R = 0$. t is modeled by the t-student distribution with DF degree of freedom.

correlation coefficient between WER and SNR (and respectively WER with age and WER with syllable rate). Between 143 and 175 interviewees were used for this analysis, depending on the scenario under consideration. Results are given in Table 1, where t is the level of significance of the test (probability of rejecting $H0$ when it is true). It appears that both the SNR, and to a lower extent the syllable rate are connected to the WER, as the probability of rejecting $H0$ when it is true is low, while it does not show any evidence to claim that the speaker's age and WER are connected.

One characteristic of this data is that a large quantity of meta information is available about each interviewee, such as their country of origin, the languages they speak, their level of education, the places they lived, their religion, and so on. It is believed that some of these features could help characterizing the speaker's accent. In order to look for a finer level of predictability, we used C4.5 [13] to relate such factors to the average WER. In spirit, this is similar to the error analysis that was carried out in [3] on Switchboard.

To test this, we compute the average WER for each speaker and to build a classifier to attempt predicting the observed WER based on the speaker characteristics. These characteristics included the speaker's age, average syllabic rate, average SNR, country of birth, spoken languages, level of education, religion, etc. The word error rates were divided into 5 groups, defined by dividing the word error rate distribution into 5 quantiles. C4.5 was used to build such a classifier, hoping that it would identify some of the features that could help predicting the word error rate.

We believe that the differences in WER across speakers may be explained by some of the underlying characteristics of the speakers (eg. accent, age) or of the recording conditions (eg. SNR). We repeated the training of the decision tree on randomly selected held-out data sets, hundreds of time for various pruning configurations of C4.5. It turned out that C4.5 was not able to consistently predict the 5 WER classes based on the features we selected. Indeed, the WER classification error on the held-out test set was always very close to a random guessing, illustrating the very high degree of variability in the data and the lack of a strongly dominant factor contributing to explaining the performance. Informal analysis of the decision trees generated during the held-out experiments showed that the SNR feature would typically occur at some of the highest levels in the tree, rather than near the leaves, though this was highly variable across the held-out test sets. No trend was observed related to the other features, including degree of accent. Our results are consistent with the findings in [3], where it was found to be difficult to predict the WER using a variety of acoustic-related features, and where the most important features were syllable rate and SNR.

4. Noise compensation in the log-spectral domain

As the SNR ratio is shown to be the largest contributor to the to WER, we carried out some preliminary experiments using a simple noise compensation technique. The approach we followed essentially corresponds to a minimum mean square error (MMSE) estimation of the clean speech, carried out in the log-spectral domain. This is similar to the approach described in [14]. The basic idea is to define a mismatch function representing the non-linear relation between noisy speech, noise and clean speech in the log-spectral domain. Then, given an existing clean speech model (represented as a Gaussian mixture model), it becomes possible to estimate the distribution of the noise (represented as a single Gaussian distribution) given the observed noisy signal, and hence the distribution of the noisy speech. Once the distribution of the noisy speech is available, a MMSE estimate of the clean speech can be derived. The major issue in this approach is due to the non-linear nature of the relation between noisy speech, noise and clean speech, calling for some simplifying assumptions in order to carry out the estimation. One of such assumption, used in this work, is to use the vector Taylor series approach to linearize the mismatch function [14], while some authors have recently used a numerical integration approach with some success [15, 16]. While this type of filtering is typically applied by mapping the cepstral coefficients back into the log-spectral domain using an inverse discrete cosine transform [15, 16], therefore leading to a highly smoothed version of the original FFT spectrum, we applied the noise compensation right after the FFT analysis, prior to the Mel binning used to derive the MFCC coefficients.

The clean speech model needed for the noise compensation is a 64 Gaussian mixture model estimated on about 6 hours of the high SNR sentences extracted from the training data. The mean and variance of the noise Gaussian distribution are blindly estimated on the first 10 frames of each sentence, and the mean is further refined by running several iterations of the EM algorithm. Once the mean of the noise has been estimated, the distribution of the noisy speech is derived and the clean speech signal is estimated, frame by frame. The entire process is conducted in the log-spectral domain and the noise-compensated spectrum is then used instead of the original noisy speech spectrum in the subsequent steps of the feature extraction. Note that similar processing was used both in training and test. Experiments were run on the small test set, and results are given in Table 2 for both the original system (without noise compensation in training nor test) and the noise-compensated system.

Original system	Noise-comp. system
45.8%	44.7%

Table 2: WER on original system and noise compensated system (small test set).

A small absolute improvement (1.1%) of the WER rate is observed and we believe that further improvement can be obtained by using an exact numerical integration to derive the mean of the noisy speech. We also observed some numerical instabilities during the noise estimation in some of the low-energy spectral bins, and are planning to reduce the spectral resolution to limit this problem, by carrying out the noise compensation on Mel-binned log-spectrum.

5. Conclusions

We presented an analysis of the word error rate on an English subset of the MALACH database. This is a difficult task as it encompasses many sources of variability, related to accented speech, noisy conditions, wide speaker's age range, and elderly speech. Our analysis shows that the SNR and syllable rate are the most dominant factor in contributing to the ASR performance on that task, in terms of acoustic variability. This is consistent with the findings reported in [3] on Switchboard, despite the fact that our corpus is heavily accented. In addition, a noise compensation technique is shown to provide a moderate reduction of the word error rate. Another aspect of this database that was left out in this paper resides in the difficulty of coming up with a good language model and lexicon on this task, as the testimonies are full of named entities and disfluencies. A companion paper addresses some of these issues.

6. Acknowledgments

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.

7. References

[1] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajič, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 2004. To appear.

[2] R. Rosenfeld, R. Agarwal, B. Byrne, R. Iyer, M. Liberman, E. Shriberg, J. Unverferth, D. Vergyri, and E. Vidal. Error analysis and disfluency modeling in the switchboard domain. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, USA, 1996.

[3] M. Ostendorf, B. Byrne, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. Technical report, The Center for Language and Speech Processing, John Hopkins University, 1996. 1996 LVCSR Summer Research Workshop.

[4] Mississipi State Univ. ISIP. Signal to noise ratio estimation software. http://www.isip.msstate.edu/projects /speech/software/legacy/signal_to_noise/index.html.

[5] N. Minematsu, M. Sekiguchi, and K. Hirose. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume I, pages 137–140, Orlando, Florida, USA, 2002.

[6] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano. Elderly acoustic model for large vocabulary continuous speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2001.

[7] J. Wilpon and C. Jacobsen. A study of speech recognition for children and the elderly. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, Georgia, USA, 1996.

[8] S. Anderson, N. Lieberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson. Recognition of elderly speech and voice-driven document retrieval. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Phoenix, Arizona, USA, 1999.

[9] W. Fisher. NIST Syllabification software. Available at ftp://jaguar.ncsl.nist.gov/pub/.

[10] M. A. Siegler and R. M. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 612–615, Detroit, Michigan, USA, 1995.

[11] N. Mirgafori, E. Fosler, and N. Morgan. Towards robustness to fast speech. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume I, pages 335–338, Atlanta, Georgia, USA, 1996.

[12] J. Huang, V. Goel, R. Gopinath, B. Kingsbury, P. Olsen, and K. Visweswariah. Large vocabulary conversational speech recognition with the Extended Maximum Likelihood Linear Transformation (EMLLT). In *Proc. Int. Conf. on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[13] J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, 1993.

[14] P. J. Moreno, B. Ray, and R. M. Stern. A vector Taylor series approach for environment-independent speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 733–736, Atlanta, Georgia, USA, 1996. ICASSP'96.

[15] M. Afify. An accurate noise compensation algorithm in the log-spectral domain for robust speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, pages 3037–3040, Geneva, Switzerland, 2003.

[16] T. A. Myrvoll. Optimal filtering of noisy cepstral coefficient for robust ASR. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, US Virgin Islands, 2003.