USE OF METADATA TO IMPROVE RECOGNITION OF SPONTANEOUS SPEECH AND NAMED ENTITIES

Bhuvana Ramabhadran, Olivier Siohan, Geoffrey Zweig

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

Abstract

With improved recognition accuracies for LVCSR tasks, it has become possible to search large collections of spontaneous speech for a variety of information. The MALACH corpus of Holocaust testimonials is one such collection, in which we are interested in automatically transcribing and retrieving portions that are relevant to named entities such as people, places, and organizations. Since the testimonials were gathered from thousands of people in countries throughout Europe, an extremely large number of potential named entities are possible, and this causes a well-known dilemma: increasing the size of the vocabulary allows for more of these words to be recognized, but also increases confusability, and can harm recognition performance. However, the MALACH corpus, like many other collections, includes side information or metadata that can be exploited to provide prior information on exactly which named entities are likely to appear. This paper proposes a method that capitalizes on this prior information to reduce named-entity recognition errors by over 50% relative, and simultaneously decrease the overall word error rate by 7% relative. The metadata we use derives from a pre-interview questionaire that includes the names of friends, relatives, places visited, membership of organizations, synonyms of place names, and similar information. By augmenting the lexicon and language model with this information on a speaker-by-speaker basis, we are able to exploit the textual information that is already available in the corpus to facilitate much improved speech recognition.

1. INTRODUCTION

In a recent report, an international digital library working group called for the creation of systems capable of providing access to an estimated 100 million hours of culturally significant spoken word collections [12]. Achieving this will require two fundamental advances over the present state of the art: (1) the degree to which existing LVCSR and NLP techniques can be adapted to provide access to spontaneous conversational speech and (2) the robust ability to identify spoken words and other useful features such as named entities in many types of collections. Several narrow-band and broadband speech collections are currently available [1, 13, 4], and carefully tuned Automatic Speech Recognition (ASR) systems are now able to achieve word error rates between 10% and 40%, depending on the difficulty of the collection. The Spoken Document Retrieval (SDR) track of the Text Retrieval Conferences (TREC) has demonstrated the feasibility of subject-based searching in non-spontaneous broadcast news collections in the presence of such word error rates [13]. However, none of these spontaneous speech corpora were designed to contain a substantive discussion of the same topic by multiple speakers or mentions of several named entities that could subsequently be searched for. The MALACH corpus described in [7, 14, 15] naturally lends itself as an excellent testbed for both LVCSR as well as NLP and search applications.

Apart from spoken archives, many practical applications such as name-dialing and call-center applications require not only accurate recognition of spontaneous speech but accurate recognition of names, places, digits, foreign words, and acronyms. Building such a system is complex due to the very large number of such entities that can occur, some more frequently than the others. Name recognition has been the focus of many researchers, particularly in the context of directory assistance applications [3, 2, 5, 6], where ASRs are designed to recognize between 200K and 2M names. A significant decrease in recognition performance has been noted when increasing vocabulary size in [3]. One of the techniques proposed to counter the adverse effects of a large lexicon is to include a diverse set of pronunciations to cover the acoustic variability [3]. While this helps, it has been shown that focusing on the discriminative segments in a multi-pass approach reduces the effect of confusability [2]. Approaches to include confidence measures and rejection thresholds have shown to be useful in the accurate recognition of names [5].

ASR systems, including the ones mentioned above typically focus on short-time information distributed over periods of 10-20 ms. It has been shown in [6] that capturing information distributed over longer periods of time, such as syllabic or word level time span, can lead to substantial gains in name recognition accuracy. The number of different acoustic units required for a given recognition task is a function of the vocabulary size and the nature of the underlying acoustic units. For phonemes the number of basic models (without context modeling) is fixed for a given language. However, when using syllable or word size units, the number increases in general with the vocabulary size. Many of these units are pronunciations of words which are not used frequently and will have poor coverage in the training data. For small vocabulary tasks such as alphabet or digit recognition, longer units (typically word level units) have been used successfully. Sparsity of training data has been the main hindering block in using longer acoustic units for LVCSR tasks. However, in [8] it was demonstrated that an LVCSR system which uses competing phonetic and mixed syllabic-phonetic paths in parallel can be built to improve recognition of names and concepts (by 17% relative). In the MALACH project this is particularly important for the search and retrieval of segments of speech relevant to the mention of a name, place or a concept [14].

In this paper we report on the use of metadata to improve recognition of named entities while reducing the overall word error rate (WER). The use of metadata in the form of a caller ID string associated with an incoming call to aid name recognition in a voice mail transcription task has been presented in [9]. The metadata for the MALACH corpus, as is the case for any other oral history archive is available in the pre-interview questionnaire (PIQ) completed by the interviewees. This includes biographical data, person names, family relationships, locations and extensive demographic data. All of these named entities are not equally likely to occur during every speaker's testimony. Therefore, if the metadata can be used to select the subset of words that can occur with the highest probability on a per-interview basis, this can lead to significant improvements in recognition accuracy.

Section 2 describes the MALACH database and the difficulties associated with recognizing named entities for this data. Sections 3 and 4 describe the metadata and the technique used to condition the ASR system with the available metadata. Section 5 presents the experimental setup and improvements in recognition accuracies obtained when dynamically adapting the lexicon. Section 6 discusses the implications of better recognition on subsequent search and retrieval. The paper concludes with a summary and potential applications for this work in other collections.

2. MALACH

MALACH (Multilingual Access to Large Spoken Archives), is an ongoing effort that aims to improve 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] [7]. The MALACH corpus consists of unconstrained, natural speech filled with disfluencies, heavy accents, age-related coarticulations, uncued speaker and language switching and emotional speech collected in the form of interviews from over 52000 speakers in 32 languages. Approximately 25000 of these testimonies are in English, spanning a wide range of accents, such as Hungarian, Polish, Yiddish, German, Italian, French, Czech, Hebrew, Croatian, Spanish, Ukrainian etc. A good number of words uttered in this corpus are foreign words or sequences of words spoken in a foreign language, unfamiliar names and places. The corpus consists of elderly speech, where the age of the interviewees range from 56 years to 90 years. In order to obtain training data for acoustic and language models, approximately 200 hours of the English portion of the MALACH corpus was manually transcribed and annotated with named entities. Transcription is challenging even for skilled annotators and they typically required 8 to 12 hours to transcribe a single hour of an English interview. The difficulties arise from unfamiliar names and places, multiple languages encountered during a single interview, coarticulations related to age, highly variable speaking rates, and heavily accented speech.

3. METADATA

This metadata from the PIQ is available on a per-speaker basis and therefore serves as the name and place authority for the mentions in the interview. Many of the place names constitute cities, streets and names of concentration camps as well as their synonyms that may be mentioned during the course of an interview. For illustration, we include here an example of the actual words spoken by different speakers during several segments, annotated with the appropriate named entities (shown in bold face).

> *because there was no normal teacher in* **Blashova** *so there was a teacher that ran more or less like a high school teacher ...*
>
> *okay well we got to the point where I was in* **Peterboro** *on* **Flaten** *near near* **Peterboro**
> ...
>
> *I was no longer able to stay in* **Flaten** *and neither was my f- the son of the* **Cookland** *who was the same age as me and we both all came and went back to live in* **Rectory Road** *in* **Hackney** *then the question came as to what I was going to do with my life ...*

Moreover, these names occur in many variations (Hebrew names, Yiddish names, diminutives, first names only, nicknames, etc.). Named entities have proven to be important to searchers of this collection [14] and hence it is important that the ASR systems hypothesize these words correctly. The MALACH data offers an opportunity to study this problem through its large database of personal identities (approx. 2.5 million names) that is populated with information taken from survey forms filled out by the subjects (PIQs), additional names of topics and concepts assigned by catalogers, and a large list of place names and their synonyms (over 20,000 locations). An important challenge in this work is that many key search terms comprising of these named entities will be found only among the infrequently occurring words and phrases, and rare terms are inevitably modeled less well than more common ones.

The metadata also contains synonyms for named entities. Many street and city names have changed over a period of time, for example, St. Petersburg was formerly known as Leningrad and Petrograd. Every interviewee provided their current name, name at birth, release name, Hebrew name, Yiddish name, nicknames and any other false names they worked under during their lives. All of these were also included in the dynamic graph that was built. For example, a person with the first name, Alicia, has Alicja, Chana, Alice, Jadwiga, Alushia, and Alla as possible variations of the first name. Table 1 illustrates the distribution of names, places and foreign words on the English portion of the MALACH corpus as a function of hours of speech.

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

Table 1: Distribution of Names, Places and Foreign Words

4. APPROACH: METADATA in ASR

Given the intended role of ASR to support information access, we are particularly interested in named entity recognition [10], especially the recognition of personal names and place names which are both important search criteria. Therefore, the ASR lexicon was carefully constructed using a large database of personal identities populated with information taken from survey forms filled out by the interviewees, additional names assigned by catalogers, and a large list of place names. However, many

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

of these entities are rare terms and therefore cannot be modeled very well.

The manual transcriptions of 180 hours of training data was used to build language models using the modified Kneser-Ney algorithm [7]. The training data is relatively small (1.7M words), therefore, the language models built from Broadcast News (BN) and Switchboard (SWB) corpora (158M and 3.4M words, respectively) were interpolated with the LM built from this collection. The interpolated weights were optimized to achieve minimum perplexity on the held-out data from this collection. The perplexity of this task on the held-out test set is 72.3. Although a lexicon can be built with the most frequently occurring words and by minimizing the OOV rate on a held-out set, as the number of interviews processed grows, many new words will need to be added. It is important for these words to be recognized accurately in order for subsequent search to be successful.

The PIQ database is indexed on the interview code. The approach presented in this paper includes the named entities contained in the metadata for the interview being decoded into the ASR's lexicon and replaces the static decoding graph (Section 5.2) with the new graph appropriately weighted with the language model probabilities seen in the training data. If a mention of a named-entity did not occur in the MALACH training material, its language model probability backs off to that of an unknown word. Many of the words added from the metadata occurred in our LM training material that had been derived by interpolating MALACH data with Broadcast News and Switchboard material. During test time, the identity of the interviewee as defined by the interview code is used to derive the dynamic graph using the pre-defined metadata available for that interview code. In [9], a class-based language model was built from the metadata defined names derived from the caller ID string and a name network was composed with finite-state transducers for this specific caller. Given the spontaneous speech in the MALACH data it is very difficult to derive a network for the usage of named entities, however, augmenting the ASR lexicon with possible realizations of names and places can be done.

5. EXPERIMENTS AND RESULTS

5.1. Training and test corpora

The English training corpus was generated using 15-minute segments of an interview from 720 randomly selected speakers. Thus, a total of 180 hours of data was selected for manual transcription to serve as training material for ASR systems. Male and female speakers in this corpus were more or less equally distributed and a wide range of accents were covered (e.g., Hungarian, Italian, Yiddish, German, and Polish). The ASR test set consists of 30 minute segments taken from 15 randomly chosen speakers. This test set was also appropriately annotated with named entities tags (illustrated in the example in Section 3).

The audio signal was down-sampled to 16KHz from 44.1Khz and parameterized using 24-dimensional mel frequency cepstral coefficients (MFCC). Final acoustic features were derived using linear discriminant (LDA) and maximum-likelihood based linear transformations (MLLT). Speaker specific transformations (SAT and MLLR) were used by the final system that gave the best WER. This ASR system had a WER of 35.2% on the test set described in [7].

5.2. Decoding Strategy

The decoder used in our experiments is a Viterbi decoder operating on a fully flattened state-level HMM. A traditional decoding setup operates on a single HMM constructed from an overall language model. In contrast to this, in our experiments, the states are built for each speaker using a speaker-specific lexicon and language model. Detailed descriptions of the Viterbi decoder used is presented in [11].

5.3. Results

The use of metadata was evaluated using the overall WER and the WER on named-entities for different vocabulary sizes. Table 2 illustrates the large gains obtained when incorporating metadata information into the recognizer. The maximum gains (51% relative) on the recognition of named entities was obtained with a 30K vocabulary that was specialized to include person-specifc metadata. On the other hand, the improvement in named entity recognition is fairly small if the size of the static lexicon is increased by a factor of three. If the names and places derived from the metadata were not added to the ASR's lexicon, the named entity WER decreases marginally when the vocabulary is expanded from 30K to 90K (66.4% to 61.4%). However, when adding speaker-specific information, the named-entity WER goes down from 66.2% (with a static 30K vocab) to 32.3%, almost reducing the WER by half. It can be seen that as the lexicon size increases, a small percentage of the gains obtained from the metadata is lost, probably offset by the added confusability, i.e with a 90K lexicon, the named entity WER increases to 38.8% from 32.3%. This is consistent with the degradation in performance seen in the literature with increased lexicon sizes. Table 3 shows the decrease in overall WER (relative 7%) obtained with the use of metadata. The reduction in the overall WER that is obtained when tripling the lexicon size without the use of metadata is much smaller (relative 2.5%).

Vocab	Static Vocab (%)	Metadata adapted Vocab (%)	Relative Gain (%)
30K	66.2	32.3	51.2
60K	62.1	36.8	40.7
90K	61.4	38.8	36.8

Table 2: WER computed on the named entities for different vocabulary sizes

Vocab	Static (%) Vocab	Metadata adapted (%) Vocab
30K	40.1	37.6
60K	39.4	36.7
90K	39.2	36.5

Table 3: Improvements in overall WER for different vocabulary sizes

Our goal is to select the vocabulary size that yields the best overall WER and named entity WER, and surprisingly, the 30K vocabulary coupled with the PIQ words (the best matching vocabulary) for an interview is the best choice. This is interesting because Table 4 shows that a 90K vocab actually has a much better overall OOV rate. However, the extra words increase confusability and our results show that this is detrimental.

Vocab	Named Entities (%) OOV rate	OOV rate (%) w/PIQ
30K	25.5	9.2
60K	16.4	8.9
90K	13.2	8.8

Table 4: OOV rate computed on the named entities

6. IMPLICATIONS FOR SEARCH AND RETRIEVAL

A test collection for the English data in the MALACH corpus was presented in [14]. The collection comprised of 404 full interviews comprising over 600 hours of speech with automated speech recognition transcripts and associate relevant judgments for 28 queries. These queries were built from over 600 written requests for materials from the collection from scholars, educators, documentary film makers and students. The mean average precision score obtained by the best search system was 0.09. An analysis of the retrieval systems indicated that for 25% of the queries, the keywords did not appear in the ASR transcripts and hence resulted in failure of the system to retrieve the relevant segments. In about 30% of the queries, the keywords were recognized at least once in a segment of speech even though there were several mentions of the same. All of these keywords were domain-specific names and places. This illustrates the importance of a high recognition accuracy on named-entities for search and retrieval tasks.

7. CONCLUSIONS

This paper presents a technique to incorporate metadata information into a speech recognition system. The results show that whenever available, adding domain-specific metadata not only provides substantial gains of the order of 50% relative to named-entity detection accuracies; but also provides a modest improvement in overall WER that otherwise cannot be achieved by simply increasing the size of the static lexicon. The large gains in the recognition of named entities is crucial to search and retrieval tasks, particularly in the MALACH project. Analysis of real-user requests for this corpus indicates that the topical requests account for approximately 53% of the requests while 89% of them are searches based on person names, organization, camp and city names. Therefore for search to be successful it is more crucial to recognize these terms accurately than it is to improve the overall recognition accuracy. The technique presented here on the use of metadata achieves both simultaneously. The use of metadata is extremely promising and we plan to explore its impact on search by redecoding the test collection described in [14] with a lexicon derived from the metadata for each interview. The proposed algorithm also has applications in searching other spoken word collections such as recordings of meetings, lectures and call center mining.

8. 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.

9. References

[1] Bacchiani, M., “Automatic Transcription of voice-mail at AT&T”, *ICASSP*, 2001.

[2] Junqua, J.-C, Valente, S., Fohr, D., abd Mari J.-F, “An n-best strategy, dynamic grammars and selectively trained neural networks for real-time recognition of continuously spelled names over the telephone”, *ICASSP*, pp. 852-855, 1995.

[3] Gao, Y., Ramabhadran, B., Chen, J., Erdogan, H., and Picheny, M., “Innovative approaches for large vocabulary name recognition”, *ICASSP*, pp. 53-56, 2001.

[4] Glass, J., Hazen, T. J., Hetherington, L., and Wang, C., “Analysis and Processing of Lecture Audio Data : Preliminary Investigations”, Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval, *HLT-NAACL04*, 2004.

[5] Liao, Y.-F. and Rose, G., “Recognition of chinese names in continuous speech for directory assistance applications”, *ICASSP*, pp. 741-744, 2002.

[6] Sethy, A., Narayanan, S. and Parthasarthy, S., “A syllable based approach for improved recognition of spoken names”, *Proceedings of the ISCA Pronunciation Modeling Workshop*, Denver, 2002.

[7] Ramabhadran, B., Huang, J. and Picheny, M., “Towards Automatic Transcription of Large Spoken Archives - English ASR for the MALACH Project”, *ICASSP* 2003.

[8] Sethy, A., Ramabhadran, B., and Narayanan, S., “Improvements in English ASR for the MALACH project Using Syllable-Centric Models,”, Proc. Automatic Speech Recognition and Understanding Workshop, *ASRU*, 2003.

[9] Maskey, S., Bacchiani, M. Roark, B., and Sproat, R. “Improved Name-Recognition with Meta-data dependent name networks”, *ICASSP*, 2004.

[10] 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.

[11] Saon, G., Zweig, G., Kingsbury, B., Mangu, L. and Chaudhari, U., “An Architecture for Rapid Decoding of Large Vocabulary Conversational Speech”, *Eurospeech*, 2003.

[12] EU-US Working Group on Spoken-Word Audio Collections, http://www.dcs.shef.ac.uk/spandh/projects/swag, 2003.

[13] Garofolo, S. J., Cedric, G., Auzanne, P. and Voorhees, E. M., “The TREC Spoken Document Retrieval Track: A Success Story”, The Eighth Text Retrieval Conference, TREC-8, 1999.

[14] Oard, D.W., Soergel, D., Murray, C. G., Doermann, D., Wang, J., Ramabhadran, B., Franz, M., and Gustman, S., “ Building an Information Retrieval test Collection for Spontaneous Conversational Speech”, to appear in *SIGIR*, 2004.

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