Lexicon Design for Transcription of Spontaneous Voice Messages

Michal Gishri, Vered Silber-Varod, Ami Moyal
ACLP – Afeka Center for Language Processing, Afeka College of Engineering
Tel Aviv, Israel
E-mail: michalg@afeka.ac.il, veredsv@afeka.ac.il, amim@afeka.ac.il

Abstract

Building a comprehensive pronunciation lexicon is a crucial element in the success of any speech recognition engine. The first stage of lexicon design involves the compilation of a comprehensive word list that keeps the Out-Of-Vocabulary (OOV) word rate to a minimum. The second stage involves providing optimized phonemic representations for all lexical items on the list. The research presented here focuses on the first stage of lexicon design – word list compilation, and describes the methodologies employed in the collection of a pronunciation lexicon designed for the purpose of American English voice message transcription using speech recognition. The lexicon design used is based on a topic domain structure with a target of 90% word coverage for each domain. This differs somewhat from standard approaches where probable words from textual corpora are extracted. This paper raises four issues involved in lexicon design for the transcription of spontaneous voice messages: the inclusion of interjections and other characteristics common to spontaneous speech; the identification of unique messaging terminology; the relative ratio of proper nouns to common words; and the overall size of the lexicon.

1. Introduction

Building a comprehensive pronunciation lexicon is a crucial element in the success of any speech recognition engine. The two main concerns of lexicon design are word list compilation, and the subsequent transcription (conversion to phonemic representations) of all lexical items on the list. Previous studies on large vocabulary lexicon design (Lamel and Adda, 1996; Bacchiani, 2001) focus mainly on transcription issues and their consequences on recognition performance. However, equally important is the pre-transcription phase of selecting which vocabulary items are actually represented in the lexicon. This stage of word list compilation is critical due to the fact that a word that does not appear in the lexicon will not be recognized by the speech recognition engine. In the recognition of *spontaneous* speech, the composition of the word list is particularly central as the predictability of vocabulary used by speakers is extremely limited. On the other hand, the size of the lexicon should be restricted so as not to adversely affect speech recognition performance, computational complexity, and consequently, system performance.

The main goal of the ACLP lexicon design is to customize a lexicon to be utilized by a speech recognition engine in the automatic transcription of spontaneous American English voice messages. Classical lexica are not fully suited to this task as they have been collected mainly from textual sources. Furthermore, existing lexica compiled using spontaneous speech sources do not necessarily meet all aspects of the target – namely voice messages. Thus, in order to customize a lexicon for voicemail transcription, an authentic database of voice messages has been analyzed.

This analysis has led to the identification of several factors that need to be addressed when designing a lexicon for the purpose of voicemail message recognition. First of all, voicemail messages are spontaneous spoken speech, which is quite different from written language, and this should be reflected in the lexical items selected for the word list. Moreover, there may be some differences between common lexical items found in voicemail messages as opposed to other forms of spontaneous speech. Any such terms must be included in the lexicon. Furthermore, due to the informative nature of voice messages, a large number of proper nouns ("names") is required. This means that typically implemented ratios of content words to names are not applicable and the overall size of the lexicon is affected.

The following paper describes the methodology and sources used in order to compile a word list that is suited to the transcription of spontaneous voicemail messages (§2). Our analysis has led to the identification of four characteristics of voicemail messages that should influence lexicon design (§3). The results presented raise further questions regarding later stages of lexicon design which will be discussed briefly in (§4).

2. Data Collection Methodology

For effective coverage, the Out-Of-Vocabulary (OOV) word rate should be kept to a minimum. However, when it comes to spontaneous speech, it is impossible to predict what any given speaker will say. Moreover, it is common knowledge that vocabulary used in spontaneous speech is quite different from vocabulary used in text. Thus the first step in spontaneous speech lexicon compilation should be to extract words from representative spontaneous speech databases.

For this purpose, IBM’s Voicemail I and Voicemail II (henceforth: VM) (Padmanabhan, 1998a; Padmanabhan 1998b, Padmanabhan, 2002) were licensed from LDC – Linguistic Data Consortium. To increase the size of the spontaneous speech corpus used, transcriptions of the Santa Barbara Corpus of Spoken American English (henceforth: SBCSAE) (DuBois and Englebretson, 2004; MacWhinney, 2007) were also analyzed.

In order to further ensure maximum coverage, textual sources were used to supplement the spontaneous speech lexicon. The use of textual sources also facilities the division of lexical items according to topic domains. A design goal of at least 90% coverage of words for each domain was set, and high frequency words from individual domain-relevant sources were extracted. Domain-specific retrieval is particularly essential when it comes to proper nouns (names). This is because the commonality of names cannot necessarily be determined based on their occurrence in a spontaneous speech database (§3.3).

To date, a large lexicon of 100,000 American English words has been successfully completed by the LC-Star project (Ziegenhain, 2004), with an attempt to cover the entire language vocabulary for a general set of applications. This format can be used as a basis for building almost any lexicon geared to automatic transcription in a specific domain or application, in our case – messaging. The ACLP domain design is based on the design determined by the LC-Star project (Ziegenhain, 2004) and includes the following six domains for common word retrieval: sports, news, business, culture, consumer information, and personal communications. Details regarding the extraction of common words will not be outlined in this paper, but will be shared in future publications. An underlying assumption is that the lexicon can be edited at a later stage to incorporate new words in the operational mode of the transcription engine.

3. Voicemail Characteristics

3.1 Spontaneous speech

It is common ground that spontaneous speech is very different from written text, read speech or otherwise elicited speech. These differences are particularly evident in the rampant number of disfluencies. These include mispronounced or partial words ("I'm not *rea-* I'm not really sure what's going on"), filled pauses (e.g. um, er), speaker noises (e.g. laugh, breath, sneeze) and background noises. While speaker and background noises are commonly dealt with via acoustic training, speech disfluencies must be incorporated into the lexicon when possible. This means that examples such as "um" "er," "uh-oh" etc. should be incorporated into any word list that is intended for spontaneous speech recognition. It will be left to be decided at a later stage whether or not to remove these disfluencies from the reported output of the transcription engine.

On the other hand, stutters and other unintentional partial utterances cannot be incorporated into the lexicon. This is because their occurrence is unpredictable and any attempt to include partial words will exponentially increase the size of the lexicon. However, their existence cannot be completely ignored as they are extremely common in spontaneous speech. In fact, there are 32,697 instances of this type of disfluency in VM (10%). These will need to be treated in later stages or at the algorithmic level. Since the optimal solution is not in the word list, this issue will not be discussed further here.

Additional fillers common to spoken language that should be incorporated include, discourse markers (e.g. *like*, *ya'know*), interjections (e.g. the *frickin'* door won't open) and profanities, among others. Colloquial word forms (e.g. yup for 'yes') and truncated contractions (e.g. *'em* for 'them', *'cause* for 'because') are also essential to a successful transcription lexicon.

3.2 Messaging Terminology

In order to determine whether the VM databases contain any terminology unique to messaging, a word frequency analysis of VM and SBCSAE was executed to establish a list of most probable words in spontaneous speech. Any terminology unique to messaging, if exists, can be determined by comparing the VM and SBCSAE word lists.

Of the 329,350 word tokens found in the combined VM databases, 9,908 unique word types[1] were identified (3% types per tokens ratio). Of the 238,626 word tokens found in the SBCSAE downloadable transcripts (DuBois and Englebretson, 2004; MacWhinney, 2007), 12,175 unique word types were identified (5% types per tokens ratio). The cumulative number of tokens comes to 567,976, while the collective number of types reaches only 16,713 (2.9% types per tokens ratio). The types per tokens ratio reflects thin vocabulary, rather than rich vocabulary found in academic articles, literature and other forms of written text.

Of the 4,493 lexical entries found in VM but not in SBCSAE, 3,478 are common words (not names). These may include special "messaging terminology" words. At a later stage in the ACLP messaging lexicon project, this list will be cross-checked against the six-domain database in order to ensure that words specific to messaging are incorporated in a separate domain.

3.3 Common Words vs. Names Distribution

A further analysis of voicemail messages indicates that it is also necessary to customize the distribution of common words versus proper nouns (names).

The LC-Star project (Ziegenhain, 2004) defined the optimal lexicon as containing 45% names, 50% common words and 5% special application words. However, preliminary research conducted at the ACLP indicated that this distribution may not be applicable to voice messaging. The results demonstrated that non-intimate messages contain at least two names per message as opposed to informal or friendly messages that may contain no names ("Hi, it's me, bye"). Analysis of the VM database reflects the same insight. Figure 1 illustrates that the number of names per message peaks at 2-3.

[1] These numbers may vary slightly based on the definition of "word type."

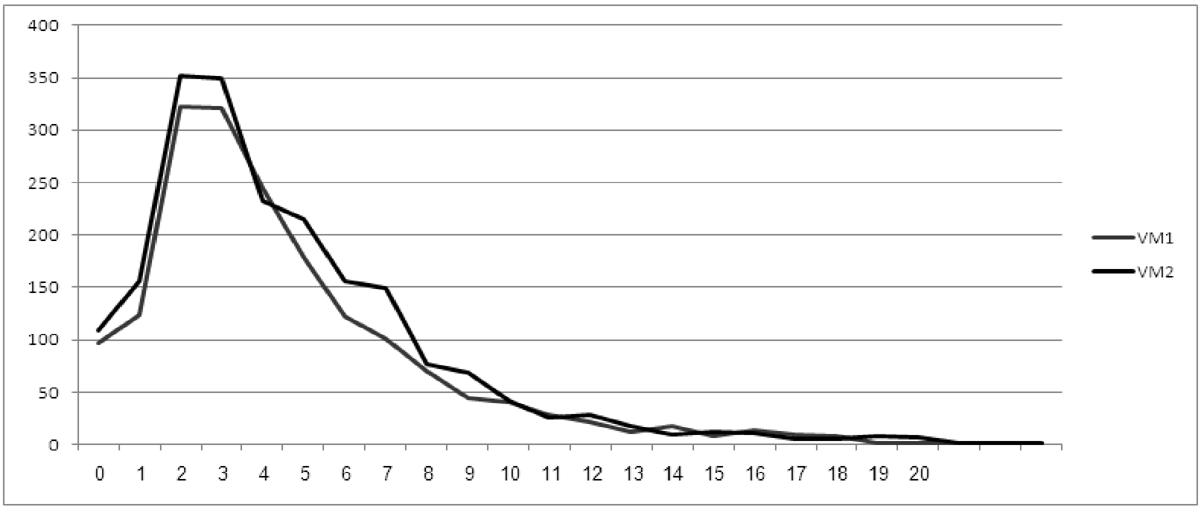

Figure 1: Name frequency per message in Voicemail I and II

This finding is also supported by the AT&T voicemail proprietary corpus, where out of ~10,000 voicemail message, 87.57% of all non-empty messages contain at least the caller's name (Jansche and Abney, 2002).

In order to test whether the distribution of names in voicemail messages varies from other forms of spontaneous speech, the VM and SBCSAE databases were again compared. In the VM transcriptions proper names are clearly indicated. Thus division of the word list into content words and names was relatively straightforward. However, it was necessary to ignore paralinguistic markings (indicated by triangular brackets, such as <INHALE> <LAUGH>), so that they would not be included in the word list. The SBCSAE transcriptions were also cleaned-up, so as to exclude markings irrelevant to our study. Furthermore, names in SBCSAE transcriptions were not marked other than via a word-initial capital, which was also used sentence-initially. It was therefore necessary to manually clean-up transcriptions of all sentence initial capitalized words that were not also names[2].

The results of the name distribution comparison revealed the following: As mentioned above, Voicemail I and II combined contain a total count of 329,350 word tokens and 9,908 word types. Among these 329,350 tokens, 18,100 are names. This means that 5.5% of the VM tokens are names. However, of the 9,908 word types found, a total of 3,770 are names, making up a total of 38% of the unique word types in voicemail messages. This means that 94.5% of words spoken in messages can be covered by approximately 6,138 lexical entries, but that over 3,700 lexical entries (i.e. names) are needed to cover the remaining 5.5%. The SBCSAE database contains a total of 238,626 word tokens compared to 12,175 types. Among these 238,626 tokens, 7,818 are names. Thus a relatively comparable 3.3% of SBCSAE tokens are names. On the other hand, of the 12,175 word types found in SBCSAE, only 2,687 are names. This is only 22% – relatively low compared with the nearly 38% found in VM. Moreover, in the SBCSAE database a larger number of common words (9,488) are needed to cover 96.7% of a smaller corpus, while a smaller number of 2,687 names are sufficient to cover the remaining 3.3%. The resulting conclusion is that a lexicon for the messaging transcription application should include a proportionately large number of names.

3.3.1. Name List Compilation

Given the fact that names are not only rampant in voice messages, but also extremely non-repetitious in comparison to common words, 90% coverage of names is a challenging task. Completing a comprehensive name list should include an in-depth analysis of common names for people (first and last), cities, countries, tourist locations, streets, organizations, companies, brand-names, web-sites and holidays, to name some.

As an example, the United States Census 2000 (U.S. Census Bureau) provides a list of all surnames registered in the US according to their frequency in the population (the list includes only names occurring 100 or more times). This list alone contains 151,671 entries of possible surnames. Only 18,839 are needed to cover around 77.5% of the population. However, while the remaining 69,960 surnames occur each in less than 0.001 percent of the population, together they make up the substantial 12.5%. This means that in order to reach 90% coverage for surnames, it is necessary to include all 151,672 entries on the list. The Census 1990 provides a list of 4,275 female first names and 1,219 male first names (U.S. Census Bureau), covering approximately 90% of the population. An additional 1,000 "most popular babies names" cover new names added since 1990. Thus nearly 160,000 lexical entries are needed in the word list in order to cover people names alone.

[2] Note that the determination of proper nouns in this case is subject to interpretation, and thus some level of leniency must be allowed regarding the accuracy of the results.

4. Lexicon Size

Because the ratio of name types is higher in voicemails compared to standard spontaneous speech (§3.3), it follows that the ratio of common words is actually lower (particularly in voicemails as shown above). Thus it may be possible to decrease the number of common words from the 50K recommended by the LC-Star Project (Ziegenhain, 2004). However, at this stage we will assume that to guarantee 90% coverage, the standard 50K is still warranted. Furthermore, following the results presented above, it is estimated that over 200,000 names will be needed to complete the lexicon. Thus the ratio of names to common words should be roughly 80% and 20%, respectively (cf. Figure 2).

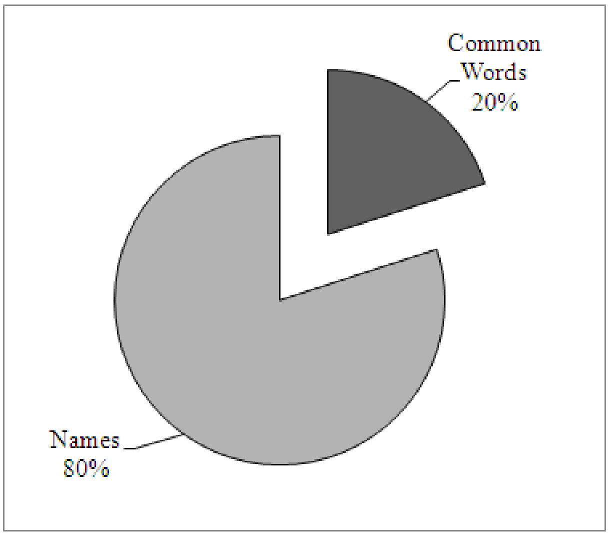

Figure 2: Lexicon distribution covering at least 90% of common words and names

According to these results, in order to achieve the design goal of 90% coverage an extremely large lexicon of 250,000 words is needed (50K common words and 200K names). This is much larger than classical lexicons used for LVCSR (Large Vocabulary Continuous Speech Recognition) engines. It is therefore necessary to perform a cost-benefit study taking into consideration the size of the lexicon, as well as computational complexity and memory requirements for the transcription task. This will allow us to conclude the optimal number of words needed to compile a lexicon for the messaging transcription task. Parallel to these activities, we are also researching ways to substantially reduce the computational complexity of the speech recognition engine in order to enable the use of such a large lexicon.

5. Discussion

The analysis presented in this paper has shown that names are frequent in voicemail messages but have relatively low recurrence rates. This has led us to conclude that a lexicon designed for the purpose of automatic voice message transcription should contain an unusually large number of names. This raises several problematic issues. First and foremost, an active lexicon with 250,000 entries is a challenge to the performance and required infrastructure (computation and memory) of current speech recognition technologies. This problem is increased multifold, when one considers that names are proper nouns, and that all nouns in English may be pronounced in either plural, possessive or plural possessive forms. Even if it is concluded that it is sufficient to duplicate each name entry with only the addition of a word-final [s] (ignoring the intended morpheme), the 200,000 names will become 400,000. To complicate matters even further, spontaneous speech also contains other contracted forms of names (e.g. "*John'll* [John will] call you tomorrow", "*Sara'd* [Sara would] like you to call her Monday"). Taking this into consideration would essentially quadruple the lexicon size. The ACLP is looking into other ways to solve this issue. One option is to add all contracted endings as separate lexical entries. However, this is likely to lead to a large number of insertions and thus decrease recognition results. Another innovative option is to ignore these endings in the lexicon and phoneme recognition stage, and then compensate later by adding them into a text based post processor stage operating semantic and syntactic analysis.

As mentioned, the second stage of lexicon design consists of providing phonemic representations of all lexical entries. This is a complex and tedious task which can be simplified somewhat by the use of existing transcriptions for common words and Automatic Pronunciation Generation (APG) tools. However, providing phonemic representations of names is much more difficult. APG tools are of little help, as names are often borrowed from other languages and their pronunciation is not necessarily consistent with the rules of the language. This obstacle can be partially overcome using manual transcription techniques (Huang, Zweig and Padmanabhan, 2001). However, this does not provide a complete solution as it is extremely time-consuming and unreliable due to the fact that the pronunciation of any given name may not be common knowledge – even to native speakers. A multi-transcription lexicon is also an option. Lamel and Adda (1996) suggest alternate pronunciations for 10% of the words. However, when it comes to names, the percentage of alternate pronunciations is likely to be much higher.

6. Summary

The analysis presented here has led to the identification of several elements in lexicon design which are vital to the task of voice message recognition: the lexicon must include content words that are common in spontaneous speech (as opposed to textual); the lexicon should include unique messaging terminology; the lexicon must incorporate disfluencies, contractions and truncations common to spontaneous speech; the lexicon must incorporate a much larger than usual proportion of names; and due to the overwhelming number of names needed, the resulting overall size of the lexicon is larger than usual.

From a linguistic perspective, the ideal lexicon will contain all possible words. However, given current requirements of speech recognition engines with respect to computational complexity, memory requirements and recognition performance with very large lexicons, this vision cannot be implemented. The conclusion is for a design goal of 90% coverage of words in each domain. This means that the lexicon should include at least 250,000 words with a 20/80 ratio of common words to names. However, this design also entails that reductions in computational complexity are achieved and assumes that noun-final morphemes and colloquial contractions can be dealt with in the textual analysis of the speech recognition results rather than in the speech recognition phase itself.

7. Acknowledgments

This research is part of a grant (#41914) provided by the Chief Scientist of the Israeli Ministry of Commerce for developing efficient algorithms for spontaneous speech transcription. The research was done as part of the Magneton program which encourages the transfer of research results from academic institutions to industrial companies – in this case Afeka Academic College of Engineering and SpeechModules LTD.

8. References

Bacchiani, M. (2001). Automatic Transcription of Voicemail at AT&T. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. Salt Lake City, UT, pp. 25-28.

DuBois, J. and Englebretson, R. (2004). Conversation SBCSAE Corpus. Retrieved on November 2009 from http://talkbank.org/data/CABank/

Huang J. Zweig, G. And Padmanabhan M. (2001). Information Extraction from Voicemail. In proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France. pp. 290-297.

Jansche M. and Abney S. P. (2002). Information Extraction from Voicemail Transcripts. Proceedings of the Conference on Empirical Methods in Natural Language (EMNLP), Philadelphia, July 2002, pp. 320-327. Association of Computational Linguistics.

Lamel L. and Adda G. (1996). On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition. Forth International Conference on Spoken Language, ICSLP 96. pp. 6-9.

LDC – The Linguistic Data Consortium. University of Pennsylvania. http://www.ldc.upenn.edu/

MacWhinney, B. (2007). The TalkBank Project. In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora: Synchronic Databases, Vol.1*. Houndmills: Palgrave-Macmillan.

Padmanabhan, M. et al. (1998a). Issues involved in voicemail data collection. *Proceedings ARPA Hub4 Workshop*. Lansdowne VA.

Padmanabhan, M. et al. (1998b). Voicemail Corpus - Part I (LDC98S77). CD-ROM. Philadelphia: Linguistic Data Consortium.

Padmanabhan, M. et al. (2002). Voicemail Corpus - Part II (LDC2002S35). CD-ROM. Philadelphia: Linguistic Data Consortium.

U.S. Census Bureau, Census 2000 Summary, 2000surnames file, California. http://www.census.gov/genealogy/www/data/2000surnames/index.html. Last retrieved March 9, 2010.

Ziegenhain U. (2004). Specification of corpora and word lists in 12 languages. Last retrieved March 9, 2010 from http://www.lc-star.com/archive.htm.