ISSUES IN GENERATING PRONUNCIATION DICTIONARIES FOR VOICE INTERFACES TO SPATIAL DATABASES

Julie Ngan, Joseph Picone
Institute for Signal and Information Processing
Mississippi State University
Mississippi State, Mississippi 39762, USA
Ph (601) 325-3149 - Fax (601) 325-3149
{ngan, picone}@isip.msstate.edu

Abstract **- In speech recognition research, increasing emphasis has been placed on generating pronunciation dictionaries for spontaneous human-computer interactions. We present here a review of our strategies for developing lexicons for three distinct voice interfaces: (1) recognition of proper nouns in data entry applications, (2) recognition and synthesis of spoken language for the navigation of a spatial information database using a speech-aware graphical user interface, and (3) interpretation of prosodic information to improve machine understanding performance. We discuss a broad range of issues in the development of such systems including surname pronunciations and automatic mapping of dictionary-style pronunciations into systems used to improve the performance of machine-based speech recognition and generation.**

INTRODUCTION

A long-term goal of speech recognition research has been to develop technologies that will allow vocal interaction between humans and computers. As a result, an interface that facilitates interpretation of human speech into information the computer can understand is needed. Since various software exist that will perform reliable speech-to-phoneme and phoneme-to-speech conversion, a significant challenge in speech research is to accurately identify the actual text associated with the phonemes. Pronunciation dictionaries are central to addressing this problem.

A pronunciation dictionary consists of words and how these words are pronounced through the use of phonemes. It serves as a reference for text-to-phoneme or phoneme-to-text conversion. Once a pronunciation dictionary is built, applications can be implemented to perform speech recognition and synthesis.

One of the goals at the Institute of Signal and Information Processing (ISIP) is to develop next generation software and techniques that will benefit speech recognition research, and to place them into the public domain. We are particularly interested in developing systems that are capable of intelligent interactions by the integration of a multiplicity of interface technologies including speech, natural language, database query, and imaging. In this paper, we report on several dictionaries and tools related to this mission that we have made available to the public.

DESCRIPTION OF PROJECTS

We present here a brief description of several projects recently completed at ISIP that were the motivation for the development of pronunciation databases for human-computer interaction.

Generation of Proper Noun Pronunciations

Speech recognition systems typically have trouble with recognition of surnames because these words can have numerous types of pronunciations, and often, not easily predicted from the spelling, thereby rendering conventional text-to-speech technology useless. We have developed a system that automatically generates an ordered list of the N-best possible pronunciations for surnames using an algorithm based on a Boltzmann machine neural network [1].

This network model transforms the spelling of a proper name to a network of phonemes that produces various pronunciations of the name using local and long-distance constraints involving n-tuples of the input string letters. The design is based on the use of a relatively small amount of contextual information to sufficiently narrow the range of possible sounds, and choosing a correct sound with information at more remote points in the name. The network is trained using a pronunciation dictionary that contains the surname spellings and letter-to-phoneme aligned pronunciation transcriptions. The generated pronunciation models can then be used for speech recognition and synthesis of proper nouns, with extensions to corpora involving directory assistance, corporate names, and general English.

Navigation of a Spatial Information Database

Voice interfaces are particularly useful for database query

applications. Humans can express complicated queries that are extremely difficult for program using standard query languages to interpret. One application we are developing is an information kiosk which uses speech recognition and synthesis interfaces to allow users to navigate a large industrial complex. This interface will particularly benefit users who are visually-challenged, but also provide other means of access to applications where a visual display is not possible or available, or applications in which the user's hands or eyes are occupied with other tasks.

The quality and effectiveness of a displayless interface for traversing a spatial information database is determined by its recognition accuracy. In such a system, human speech is broken down into phonemes; and the computer matches each phoneme in context of its adjacent phonemes to identify words. A pronunciation dictionary is used to map phoneme sequences into words. We have developed a tool that generates n-gram sequences of phonemes from our standard dictionaries. The goal of this conversion is to minimize the number of triphones required, yet maintain good coverage of the pronunciations.

We have utilized the Carnegie Mellon University (CMU) Pronouncing Dictionary [2] which supplies monophone pronunciations of approximately 116,000 commonly used English words as the source for monophones lookup. In a second stage of processing, we map all possible three phoneme sequences, referred to as triphones, into a set of 14,348 triphones used by our speech recognition system to provide good coverage of general English [3,4].

In our development of an efficient dictionary lookup and monophone-to-triphone mapping algorithm, we experimented with two different procedures:

- We created a one-to-one mapping of all CMU phonemes to the phonemes used in the triphone system. The monophone transcription of a word is generated from the CMU dictionary. Each phoneme was converted into its corresponding phoneme in the triphone set (which used a different transcription convention). The transcribed monophones were then converted into triphone format.
- We created a mapping of every possible triphone sequence in the CMU dictionary to the best available triphone sequence appearing in the recognition system. The monophone transcription of a word was generated from the CMU dictionary and converted into triphone format. Then each triphone sequence was converted into the best matching triphone.

Interpretation of Prosodic Information

Beyond the generation of accurate pronunciations, something that is crucial for a spontaneous speech interface, we are investigating the use of prosody in these command and control type interfaces. The use of prosody is crucial in speech understanding since it conveys much of the context and accounts for the elimination of many possible yet unreasonable variabilities of a speech.

In order to augment our dictionary with prosodic information, the Tones and Break Indices (TOBI) approach to prosodic analysis is used [5]. TOBI is a prosodic transcription system for marking the intonation patterns and other aspects of the prosody of English utterances. It has been developed by a diverse group of researchers with expertise in prosodic analysis and speech recognition technology to provide a standard for prosodic transcription of large speech corpora.

TOBI consists of parallel tiers reflecting that prosody has multiple components. The tone tier is used for intonational analysis whereas the break index tier is used for indicating strength of coherence or disjuncture between adjacent words. An example of a TOBI transcription is shown in Figure 1. With the use of TOBI, high quality standardized speech synthesis and recognition models can be implemented.

1. Anna will win, Manny.
   H* L- L* L-H%
2. Anna will win, Manny.
   H* H* L-L%

Figure 1. An example of a TOBI transcription for a sentence that has an ambiguous meaning. In the first case, the sentence meaning/reading is "Somebody is talking to Manny and telling him that Anna will win." In the latter case, the sentence meaning is "She won't lose him."

DATABASES AND DICTIONARIES

In order to support automatic generation of surname pronunciations, we have developed a dictionary containing approximately 20,000 surnames. Each entry contains all plausible pronunciations of the surname. A letter-to-phoneme aligned version is also available to facilitate the training of certain systems.

A major issue in accurate recognition of proper nouns is that a large number of proper nouns have no obvious letter-to-sound mapping rules. Most proper names have more than one probable pronunciation, and often, the correct pronunciation is independent of the context of the application. The construction of a representative database of surnames presents problems because a number of foreign names have both ethnic as well as anglicized pronunciations and individual pronunciations often defy any kind of typical

text-to-speech rules [6].

To facilitate our research in voice interfaces for spatial relational databases, we have developed a 5,076-word dictionary containing the most common words and phrases used in such applications. Each entry in the dictionary consists of a word, its monophone pronunciation, and a mapping to a set of triphones used by our speech recognition system. Such a system depends highly on the reference dictionary, and therefore requires an extensive phoneme mapping. In our implementation, an accurate one-to-one phoneme mapping is critical to the performance of the interface since some of the phonemes used in one dictionary do not directly map into another.

Even with a complete monophone mapping, the construction of an effective and representative triphone transcription presents problems because of the large number of possible triphones can be generated from a small number of monophones. Using such a large dataset of triphones in speech recognition research generally hinders the effectiveness of the system due to combinatorial problems. It is possible to generalize the system by limiting the number of triphones in the database to those that are more frequently used. In doing so, extensive care has to be taken to produce a representative triphone mapping reference and to maintain a good coverage of the pronunciations.

PERFORMANCE

We performed triphone transcription on the 5,076-word dictionary using the two approaches described above. We have found that the second approach is more effective. Table 1 provides an analysis of the coverage achieved:

Approach	# words in dictionary	# words transcribed successfully	overall performance
1	5076	3825	75%
2	5076	4651	92%

Table 1. Performance of the two transcription approaches.

Theoretically, approach 2 should give a 100% transcription rate since all the triphones used in this approach must appear in the triphone database. However, 8% of the words do not exist in the CMU dictionary; this shows that the approach is highly dependent on the dictionary. Another drawback of this approach is whenever the reference dictionary is updated, the triphone database will need to be augmented with new triphones created from the new words. In our attempt to limit the number of triphones used, only a subset of triphones that actually appear in the CMU dictionary is used in the mapping. This can be avoided if all triphone combinations are included in the mapping file.

CONCLUSION

Comprehensive pronunciation dictionaries are critical to speech recognition and speech synthesis applications. We have discussed various problems related to text-to-phoneme mapping for proper noun pronunciations. We have also discussed the problems involved in generating a pronunciation dictionary for the navigation of a spatial information database. The main issues included the construction of a representative database with a reasonable size that would provide good coverage of the pronunciations, the need to reformat dictionary database for accurate transcription, the need to use prosodic information to improve speech recognition, and the extension of domain-specific databases to general English.

We have implemented software and algorithms that would provide efficient and reliable n-gram generation, pronunciation dictionary lookup, and phonetic transcription, to aid researchers in the development of speech recognition technologies. The dictionaries, data, and software developed are available to the public at the URL: *http://www.isip.msstate.edu/software.*

REFERENCES

[1] N. Deshmukh, M. Weber, and J. Picone, “Automated Generation of N-Best Pronunciations of Proper Nouns,” *Proceedings of ICASSP’96*, pp. I283-I286, Atlanta, GA, May 1996.

[2] B. Weide, “The CMU Pronouncing Dictionary,” URL: *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

[3] D. Paul and J. Baker, “The Design for the Wall Street Journal-based CSR Corpus”, *DARPA Speech and Language Workshop*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[4] S. Young, P. Woodland, J. Odell, and V. Valtchev, “Large Vocabulary Continuous Speech Recognition Using HTK,” *Proceedings of ICASSP’94*, pp. 125-128, Adelaide, Australia, 1994.

[5] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. “TOBI: A Standard for Labeling English Prosody,” *Proceedings of ICSLP’92*, pp. 867-870, Banff, Alberta, Canada, October 13-16, 1992.

[6] N. Deshmukh, J. Ngan, J. Hamaker and J. Picone, “An Advanced System to Generate Pronunciations of Proper Nouns”, to appear in *Proceedings of ICASSP’97*, Munich, Germany, April 1997.