Speech Recognition Databases for Technological Applications

Vishwanath Mantha
Department for Electrical and Computer Engineering
Mississippi State University, Mississippi State, MS 39762
mantha@isip.msstate.edu

ABSTRACT

Technological applications for which speech databases are needed can be roughly divided into four major classes: speech synthesis, speech recognition, spoken language systems, and speaker recognition/verification. Depending on the specific application, the speech corpora which are needed are very diverse. In the following sections the four domains of speech research for technological applications as well as corresponding databases will be discussed.

1. SPEECH SYNTHESIS

The seemingly most natural way to synthesize speech is to model human speech production directly by simulating lung pressure, vocal fold vibration, articulatory gestures, etc. However, the human system is not completely understood. This is probably the reason why it turns out to be extremely difficult to determine and control the details of the model parameters in computer simulations. This is the reason that articulatory synthesizers have only been moderately successful in generating perceptually important acoustic features. Yet, modern measurement techniques have allowed the collection of substantial amounts of measurement data. Most of these data are now being made available to the research community.

A relatively simple way to build a speech synthesizer is through concatenation of stored human speech components[1]. In order to achieve natural coarticulation in the synthesized speech, it is necessary to include transition regions in the building blocks. Often-used transition units are diphones, which represent the transition from one phone to another. Since diphone inventories are derived directly from human utterances, diphone synthesis might be expected to be inherently natural sounding. However, this is not completely true, because the diphones have to be concatenated and in practice there will be many diphone junctions that do not fit properly together. In order to be able to smooth these discontinuities the waveform segments have to be converted to a convenient format, such as some form of LPC parameters, often with some inherent loss of auditory quality. Until recently it was believed that a parametric representation was mandatory to be able to change the pitch and timing of utterances without disturbing the spectral envelope pattern. Since the invention of PSOLA-like techniques, high quality pitch and time changes can be effected directly in the time domain. For limited applications, such as train information systems, whole words and even phrases may be stored. Lately, this method of speech synthesis is being applied more and more, because of the possibility of cheap mass storage. The quality of concatenated-word sentences is often acceptable, especially in the light of the still not optimal quality of the other types of synthesis.

Another important method of generating computerized speech is through synthesis by rule[2]. The usual approach is to input a string of allophones to some form of formant synthesizer. Target formant values for each allophone are derived from human utterances and these values are stored in large tables. With an additional set of rules these target values can be adapted to account for all kinds of phonological and phonetic phenomena and to generate proper prosody

.For all types of speech synthesis systems corpora are needed to determine the model parameters. If the user wants many different types of voice, the speech corpus should contain various speakers for the extraction of speaker-specific model parameters. In particular, the user might want to be able to generate both male and female speech. Transformations to convert rule systems between male and female speech have had limited success, so it seems more

convenient to include both sexes in the speech corpus. Application specific corpora are needed to investigate issues related to prosody.

2. SPEECH RECOGNITION

There are several types of speech recognition systems, which may differ in three important ways:

- •the recognition strategies they use,
- •the speakers they have to recognize,
- •the speech they have to recognize.

These different aspects will be discussed below.

2.1. Knowledge-based vs. stochastic systems

With respect to the strategies they use, speech recognition systems can be roughly divided in two classes: knowledge-based systems and stochastic systems. All state-of-the-art systems belong to the second category. In the knowledge-based approach an attempt was made to specify explicit acoustic-phonetic rules that are robust enough to allow recognition of linguistically meaningful units and that ignore irrelevant variation in these units. Stochastic systems, such as Hidden Markov Models (HMMs)[3] or neural networks, do not use explicit rules for speech recognition. On the contrary, they rely on stochastic models which are estimated or trained with (very) large amounts of speech, using some statistical optimization procedure (e.g. the Estimate-Maximize or the Baum-Welch algorithm). Higher level linguistic knowledge can be used to constrain the recognition hypotheses generated at the acoustic-phonetic level. Higher level knowledge can be represented by knowledge-based explicit rules, for example syntactic constraints on word order. More often it is represented by stochastic language models, for example bigrams or trigrams that reflect the likelihood of a sequence of two or three words, respectively. Recently, promising work on enhancing HMMs with morphological and phonological structure has been conducted, pointing to the possibility of convergence between knowledge-based and stochastic approaches.

2.2. Speaker independent vs. dependent systems

Speech recognition systems can be either speaker-dependent or speaker-independent. In the former case the recognition system is designed to recognize the speech of just a single person, and in the latter case the recognition system should be able to recognize the speech of a variety of speakers. All other things being equal, the performance of speaker-independent systems is likely to be worse than in speaker-dependent systems, because speaker-independent systems have to deal with a considerable amount of inter-speaker variability. It is often sensible to train separate recognition models for specific subgroups of speakers, such as men and women, or speakers with different dialects[4].

Some systems can to some extent adapt to new speakers by adjusting the parameters of their models. This can be done in a separate training session with a set of predetermined utterances of the new speaker, or it can be done on-line as the recognition of the new speaker's utterances gradually proceeds. Most recognition systems are very sensitive to the recording environment. In the past, speakers employed to train and develop a system were often recorded under "laboratory" conditions, for instance in an anechoic room. It appears that the performance of speech recognizers which are trained with such high quality recordings severely degrades if they are tested with some form of "noisy" speech[5]. Also the use of different microphones during training sessions and test sessions has a considerable impact on recognition performance.

2.3. Isolated words vs. continuous speech

The third main distinction between speech recognition systems is based on the type of speech they have to recognize. The system can be designed for isolated word recognition or for continuous speech recognition. In the latter case word boundaries have to be established, which can be extremely difficult. Nevertheless, continuous speech recognition systems are nowadays reasonably successful, although their performance of course strongly depends on the size of their vocabulary. Word spotting can be regarded as a special form of isolated word recognition: the recognizer is "listening" for a limited number of words.

These words may come embedded in background noise, possibly consisting of speech of competing speakers, or may come from the target speaker who is producing the word embedded in extraneous speech.

2.3.1. Corpora for speech recognition research

In general, two similar speech corpora are needed for the development of speech recognition systems: one for the training phase and one for the testing phase. The training material is used to set the model parameters of the recognition system. The testing material is used to determine the performance of the trained system. It is necessary to use different speech data for training and testing in order to get a fair evaluation of the system performance.

For speaker-dependent systems, obviously the same speaker is used for the training and testing phase. For speaker-independent systems, the corpora for training and testing could contain the same speakers (but different speech data), or they could contain different speakers to determine the system's robustness for new speakers.

When a system is designed for isolated word recognition, it should be trained and tested with isolated words. And similarly, when a system is designed for telephone speech, it should be trained and tested with telephone speech. The design of corpora for speech recognition research thus strongly depends on the type of recognition system that one wants to develop. Several large corpora for isolated words (e.g. TIDIGITS) and continuous speech recognition (e.g. Switchboard, TIMIT and Wall Street Journal) have been collected and made available[6].

3. SPOKEN LANGUAGE SYSTEMS

Speech synthesis and speech recognition systems can be combined with natural language processing and Dialogue Management systems to form a Spoken Language System (SLS) that allows an interactive communication between man and machine. A spoken language system should be able to recognize a person's speech, interpret the sequence of words to obtain a meaning in terms of the application, and provide an appropriate response to the user.

Apart from speech corpora needed to design the speech synthesis and the speech recognition part of the spoken language system,

speech corpora are also needed to model relevant features of spontaneous speech (pauses, hesitations, turn-taking behavior, etc.) and to model dialogue structures for a proper man-machine interaction.

An excellent overview of spoken language systems and their problems is given in [7].

4. SPEAKER RECOGNITION / VERIFICATION

The task of automatic speaker recognition is to determine the identity of a speaker by machine. Speaker recognition (usually called speaker identification) can be divided into two categories: closed-set and open-set problems. The closed-set problem is to identify a speaker from a group of known speakers, whereas the open-set problem is to decide whether a speaker belongs to a group of known speakers. Speaker verification is a special case of the open-set problem and refers to the task of deciding whether a speaker is who he claims to be.

Speaker recognition can be text-dependent or it can be text-independent. In the former case the text in both the training phase and the testing phase is known, i.e. the system employs a sort of password procedure. One popular example of password-like phrases are the so-called "combination lock" phrases, consisting of sequences of numbers (mostly between 0 and 99) or digits (between 0 and 9). LDC provides a corpus for training and testing speaker verification systems based on combination lock phrases consisting of three numbers between 11 and 99 (e.g. 26-81-57) [8].

There are various application areas for speaker recognition, for instance helping to identify suspects in forensic cases, or controlling access to buildings or bank accounts. As with speech recognition, the corpora needed for speaker recognition or speaker verification are dependent on the specific application. If, for instance, the technology is based on combination lock phrases, a training database should obviously contain a large number of connected number or digit expressions. For the development of text-independent speaker technology

there are no strict requirements as to what the training speakers say.

Corpora for the development and testing of speaker recognition systems differ in a crucial aspect from corpora collected to support speech recognition. For speaker recognition research it is absolutely essential that the corpus contains multiple recordings of the same speaker, made under different conditions. There is a range of conditions that should ideally be sampled, in order to be able to build a model of the natural variation in a person's speech due to realistic variations in the conditions under which the speech is produced. Conditions to be sampled and to be represented in a corpus can be divided into two broad groups, viz. channel conditions, and physiological and psychological conditions of the speaker[8].

4.1. Channel conditions

The details of the acoustic speech patterns depend heavily on the acoustic background in which the speech was produced and on the response of the transmission network. A corpus for speaker recognition research should at least include multiple recordings of the speakers made with different microphones or telephone handsets. Especially the transmission differences between carbon button and electret microphones in telephone handsets are known to affect the performance of speaker recognition systems. In this context, attention should also be paid to the different transmission characteristics of the fixed, landing telephone network and the rapidly growing cellular networks.

4.2. Psychological and physiological conditions

The type of speaker variation addressed under this heading is also very difficult to sample. Given the practical limitations of a corpus collection project it is hardly feasible to require that each speaker be recorded in perfect health conditions, as well as when having a cold, the flu, or any other mild disease.

One simple approximation to sampling within speaker variation that is feasible from a practical point of view is to record speakers at different times of the day (early morning, noon, late night), and on different days of the week. In any case, the period over which the recordings are extended should span at least a couple of months.

Developing and testing speaker recognition systems with a database containing only a single recording session for the speakers should be avoided, because such databases cannot possibly account for even the slightest degree of within-speaker variation.

5. REFERENCES

[1] D. Klatt, "Review of text-to-speech conversion in English", in *Journal of the Acoustical Society of America*, vol. 82, pp. 737-793, 1987.

[2] J. Holmes, "Speech synthesis and recognition", Van Nostrand Reinhold, Wokingham, UK, 1985.

[3] M. Hochberg, et. al., "Large vocabulary continuos speech recognition Using a Hybrid Connectionist/HMM system", *Proceedings of the ICSLP*, Yokohama, Japan, pp. 1499-1502, September 1994.

[4] D. Van Compernolle et al., "Speaker clustering for dialectic robustness in speaker independent recognition", *Proceedings of Eurospeech*, Geneva, Switzerland, pp. 2723-2726, April 1991.

[5] Y. Gong, "Speech recognition in noisy environments: A survey", in *Speech Communication*, vol. 16, pp. 261-291, 1995.

[6] http://www.ldc.upenn.edu

[7] Cole, "The challenge of spoken language systems: Research directions for the nineties", in *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 1-20, March 1995.

[8] J. Campbell, "Testing with the YOHO CD-RON voice verification corpus", *Proceedings of the IEEE Conference on Acoustics, Speech and SIgnal Processing*, Detroit, MI, pp. 341-344, March 1995.

[9] H. Gish & M. Schmidt, "Text-independent speaker identification", *IEEE Signal Processing*, vol. 11, pp. 18-32, June 1994.