CTIMIT: A SPEECH CORPUS FOR THE CELLULAR ENVIRONMENT WITH APPLICATIONS TO AUTOMATIC SPEECH RECOGNITION

Kathy L. Brown, E. Bryan George

Signal Processing Center of Technology
Lockheed Sanders, Inc.
Nashua, NH 03061

ABSTRACT

This paper will report on techniques used in the generation of a continuous speech, multi-speaker, cellular bandwidth database and describe its application to automatic speech recognition in the cellular environment. CTIMIT (cellular TIMIT) has been generated by transmitting the TIMIT speech database over the cellular network. The CTIMIT database can have widespread applicability in the design and development of speech processing and speech recognition products for the cellular market. We will describe the preliminary collection of the CTIMIT database and report on several studies designed to test the utility of the database in a phoneme recognition task. Two HMM-based phoneme recognizers were trained using utterances drawn from the TIMIT database and the CTIMIT database, respectively. Each recognizer was then tested using the test utterances from CTIMIT. Phoneme recognition accuracy for the TIMIT-trained recognizer dropped 58% from its baseline performance on TIMIT test utterances. By comparison, phoneme recognition accuracy of the CTIMIT-trained recognizer increased 82% compared to that of the TIMIT-trained recognizer.

1. INTRODUCTION

Due to the increasing popularity of mobile cellular communications, there is a great deal of interest in the development of speech processing and speech recognition products that perform robustly and operate effectively in the cellular environment. Of particular interest are voice dialing, speech enhancement, and speech coding applications, the performance of which can be significantly improved by training in the target cellular environment.

Standard speech recognition systems trained in low background noise will not function effectively in the severe acoustic conditions that exist in cellular communications [1]. Furthermore, conventional preprocessing techniques based on signal enhancement (such as adaptive noise cancellation) are not sufficiently effective in mobile cellular communications due to the complicated nature of the signal/noise environment [2]. As a result, the performance of practical cellular speech recognition systems is critically dependent on the availability of a large amount of training data matched to the acoustic and linguistic requirements of the cellular environment.

In addition to matching acoustic characteristics, the training database should accurately reflect the linguistic domain of interest. In general, training phoneme-based recognizers requires a large, phonetically-labeled database, such as the commercially available TIMIT database [3], to adequately capture the variation of continuous speech. Attractive features of the TIMIT database include multiple speakers, continuous speech, good coverage of North American standard dialects, and carefully designed breadth and depth of phonetic coverage.

The collection of a similarly diverse acoustic and linguistic database for the cellular environment is both a time-consuming and resource-intensive task. In order to begin design efforts for speech processing systems to operate in the cellular environment without requiring this investment, we have adopted an alternative strategy that utilizes existing database resources.

We have chosen to transmit the TIMIT speech database, which was originally recorded under clean channel conditions, over the cellular network. The resulting *cellular TIMIT* (CTIMIT) database maintains the linguistic richness of the TIMIT database coupled with the acoustic effects introduced by cellular communications environments and transmission characteristics. The strategy used to generate CTIMIT is similar in form to that used to generate the NTIMIT telephone bandwidth database [4], with some differences in implementation to be described later.

Of course, the utility of the CTIMIT database ul-

E. Bryan George is now with the Systems and Information Science Laboratory, Texas Instruments, Inc., Dallas TX 75265

0-7803-2431-5/95 $4.00

timately depends on how well recognizers trained with CTIMIT perform on speech in actual cellular environments. We have therefore performed a series of preliminary experiments to compare the phoneme recognition accuracy of a hidden Markov model (HMM) recognizer trained with CTIMIT with that of a baseline recognizer trained using the TIMIT database. The following sections describe the technical details of the methods used to generate the CTIMIT database and the recognition experiments performed, and provide detailed analysis of test results.

2. TIMIT DATABASE

The TIMIT acoustic/phonetic database consists of 630 speakers, each saying 10 sentences including

- 2 "sa" sentences, which are the same across all speakers.
- 5 "sx" sentences, which were read from a list of 450 phonetically balanced sentences selected by MIT.
- 3 "si" sentences, which were randomly selected by TI.

70% of the speakers are male. Most speakers are adult Caucasians. A complete description of the TIMIT database can be found in [3].

3. CTIMIT DATABASE GENERATION

Figure 1 shows a block diagram of the experimental setup used to generate the CTIMIT database. Clean speech from the training and testing portions of the TIMIT database was randomly ordered, then recorded onto digital audio (DAT) tapes in 24 sessions lasting approximately fifteen minutes, using a bandlimited chirp signal as a marker/separator between successive sentences. The chirp signal was chosen due to its excellent time-frequency localization [5], predictable behavior in the presence of bandlimiting, and distinctiveness compared to both speech and typical VHF interference signals.

A DAT player, along with an equalizer and audio amplifier, was then placed in a van equipped with a DC-AC power converter. As seen in Figure 1, the output of the amplifier was acoustically coupled to a cellular phone by placing a reference speaker in close proximity to the cellular phone. We chose to forego the use of an "artificial mouth" as used to generate the NTIMIT database, on the observations that many cellular speech

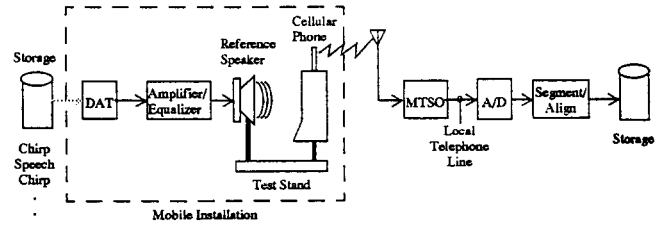

Figure 1: CTIMIT data collection apparatus.

recognition requirements are for cellular phones operated in "hands-free" mode and that modeling the coupling between mouth and handset is therefore less important than it was for NTIMIT.

Having established a cellular link through the mobile telephone switching office with a laboratory telephone, the recorded data was transmitted while the van was in motion and digitized at a rate of 8 kHz from the lab phone line. After digitizing, the data was segmented into utterances by "matched filtering" the digitized speech with the chirp signal marker and organized into a directory structure corresponding to that of the TIMIT database.

In order to control the database collection and improve diversity, the following measures were in place during the experiment:

- The speaker and cellular phone were held in place by clamps on an acoustically-isolated test stand. The cellular transceiver microphone was held one inch from the speaker at various angles, and the mean sound pressure level was calibrated to 85 dBA at this distance.
- The amplifier/speaker chain was further calibrated by feeding a sweep tone through (in an "acoustically dead" room) and displaying the output of a condenser microphone placed next to the speaker on an audio spectrum analyzer. The equalizer was then set such that the output was flat to within ± 1 dBm over the band of interest (300–3000 Hz).
- A separate phone call was placed for each of the 24 sessions. A different phone (three total, two transportable and one hand-held) was used for each successive session. In addition, different driving environmental conditions were set for each session, including varying speeds, rural/urban driving, closed- versus open-cabin, etc., and the test

stand was moved for each session. A number of cell sites in the southern New Hampshire/northern Massachusetts area were involved in the experiment, and no attempt was made to avoid cell switching during sessions.

4. SPEECH RECOGNITION USING CTIMIT

Several experiments have been carried out using a speech recognition application to test the effectiveness and utility of the CTIMIT database. We have designed phoneme recognizers using the popular hidden Markov model development toolkit HTK*. The recognizer makes use of continuous density HMMs (CDHMM) with Gaussian mixture densities for acoustic modeling.

In the front end, input speech is preemphasized using a first backward difference filter with $\alpha = 0.97$. A 25.6 msec Hamming window is then applied at a frame increment of 10 msec. For each frame, a 39 component feature vector is computed that consists of 12 mel-frequency cepstral coefficients with their first and second order derivatives, and a normalized power measurement with its first and second order derivatives.

A set of context dependent triphone models was trained from a set of 48 context independent phone models. Each triphone is modeled by a 3 state, left-to-right CDHMM with five Gaussian mixture densities per state. Model output distributions across different triphone HMMs are shared with each other when they exhibit acoustic similarity. In training, we ran 4 iterations of context dependent, forward-backward training.

Two HMM-based phoneme recognizers were trained using different sets of training utterances: 1) the training set of TIMIT downsampled to 8 kHz, and 2) the identical training utterances from CTIMIT. The TIMIT recognizer was tested using both matched and mismatched training/test conditions. In the matched conditions, the TIMIT recognizer was tested using the test utterances from the TIMIT database downsampled to 8 kHz; in the mismatched conditions, the TIMIT recognizer was tested using CTIMIT test utterances. No language model was employed in these experiments. The number of correct, substituted, inserted, and deleted phones are computed by a dynamic programming match between the correct phone string and the recognized phone string. These results are summarized in Table 1.

As illustrated in Table 1, the mismatched condition in which TIMIT-trained models were used to recognize cellular test speech resulted in a significant decrease in phone accuracy. The TIMIT-trained recognizer experiences a 58% decrease in phonetic recognition accuracy when tested with the CTIMIT test corpus. By contrast, the phone recognition accuracy of the CTIMIT-trained recognizer increased 82% compared to that of the TIMIT-trained recognizer. Recognition rates for four broad phonetic classes (sonorant, stop, fricative, and closure) are reported in Table 2.

Performance Measure	Train/Test Conditions		
	T/T	T/CT	CT/CT
% Correct	70.43	29.5	53.4
% Substitutions	24.1	43.7	33.7
% Deletions	5.2	26.7	13.0
% Insertions	9.9	7.8	7.5

Table 1: Phonetic recognition results for the TIMIT-trained (T) and CTIMIT-trained (CT) recognizers on CTIMIT test utterances.

Class	Train/Test Conditions		
	T/T	T/CT	CT/CT
Sonorant	69.2	46.5	56.2
Stop	74.6	14.1	48.8
Fricative	63.5	24.5	34.1
Closure	84.7	30.0	62.5

Table 2: Phonetic recognition results by broad phonetic class.

In the TIMIT/CTIMIT test, the stop, fricative and closure classes experience a significant deterioration in phone accuracy, which can be attributed largely to the increased levels of acoustic noise and bandlimiting in the cellular environment. Due to the decreased consonant accuracy, it is expected that word accuracy will similarly be adversely affected. Confusion matrices describing the phone accuracies for the combined stop, fricative, and closure classes are given in Tables 3 and 4 for the TIMIT/CTIMIT and the CTIMIT/CTIMIT tests, respectively. In the TIMIT/CTIMIT test, a large number of the consonants are misrecognized as the phoneme /hh/. As illustrated in Table 4, better consonant accuracy is obtained by using models trained from speech matched to the acoustic characteristics of cellular communications.

*HTK is a trademark of Entropic Research Laboratory, Inc.

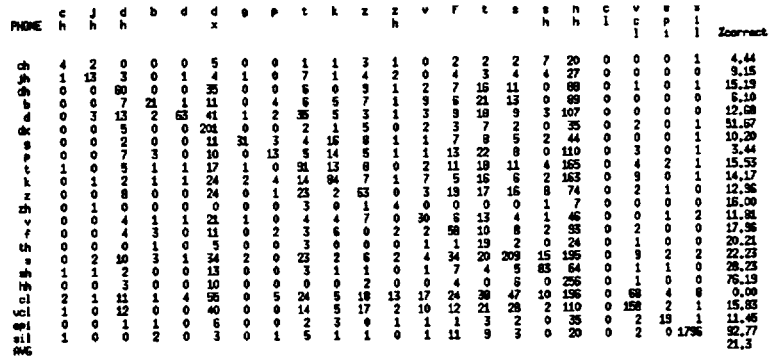

Table 3: Confusion matrix for the TIMIT/CTIMIT test.

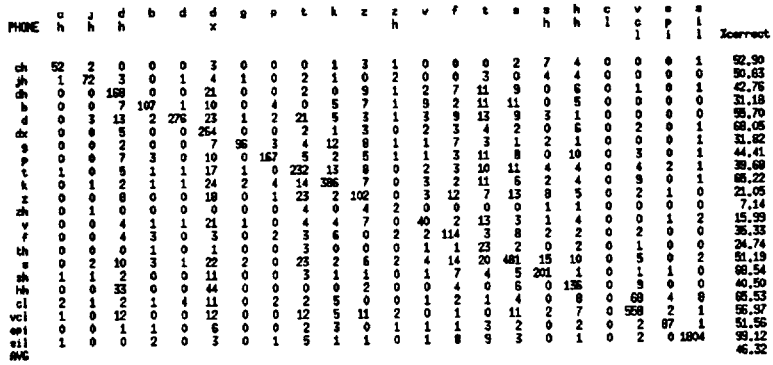

Table 4: Confusion matrix for the CTIMIT/CTIMIT test.

5. CONCLUSION

The CTIMIT database provides a phonetically labeled, multi-speaker speech corpus for acoustic characterization of the cellular communication network. We have been able to demonstrate positive preliminary results based on collection of the CTIMIT database and the use of CTIMIT to train a speech recognition system. These results illustrate the advantages of training a recognizer intended for cellular applications with a database that captures the variability of speech over the cellular network.

However, in order to maximize the performance advantage of CTIMIT when applied to speech from more diverse cellular environments, it will be necessary to address several shortcomings of the described collection experiment. For instance, the need for a power converter in the transmitting vehicle should be eliminated, as this may be a source of RF interference not present in realistic environments. Also, the number of vehicles used in CTIMIT collection, as well as the number of cellular phones used, should be increased to provide greater database diversity. Despite these challenges, our results to date suggest that CTIMIT could be an invaluable tool in the design and development of speech processing and speech recognition products for the cellular market, providing the advantages of a large, fully labeled speech corpus without the need for an intensive collection effort.

6. REFERENCES

[1] I. Lecomte et al. Car noise processing for speech input. In *Proc. ICASSP-89*, pages 512–515, Glasgow, UK, May 1989.

[2] N.Dal Degan and C. Prati. Acoustic noise analysis and speech enhancement techniques for mobile radio applications. *Signal Proc.*, 15:43–56, 1988.

[3] R.G. Leonard. A database for speaker independent digit recognition. In *Proc. ICASSP-84*, pages 42.11.1–4, Mar. 1984.

[4] C. Jankowski et al. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proc. ICASSP-90*, pages 109–112, Apr. 1990.

[5] M.I. Skolnik. *Introduction to Radar Systems*, pages 422–427. New York, New York, second edition, 1980.