# THE CTIMIT CELLULAR BANDWIDTH SPEECH CORPUS

*E. Bryan George, Kathy L. Brown*

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

## ABSTRACT

This paper reports on techniques used in the generation of a continuous speech, multi-speaker, cellular bandwidth database. 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.

## 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.

In order to match acoustic characteristics of the cellular environment for effective system design, 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 popular TIMIT database [1], 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.

E. Bryan George is now with the DSP Research and Development Center, Texas Instruments Incorporated, Dallas TX 75265

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 [2] 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 [3], with some differences in implementation to be described later.

## 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 [1].

## 3. CTIMIT DATABASE GENERATION

A block diagram of the experimental setup used to generate the CTIMIT database is found on page 4 of the accompanying PostScript file "poster.ps," and in [2]. 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 [4], 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. 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 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 $\pm 1$ 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. 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 [2]. 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.

## 5. REFERENCES


[1] W. M. Fisher et al. The DARPA speech recognition research database: specifications and status. In *Proc. DARPA Workshop on Speech Recognition*, pages 93–99, February 1986.

[2] K. L. Brown and E. B. George. CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition. In *Proc. IEEE Int'l Conf. on Acoust., Speech and Signal Processing*, pages 105–109, May 1995.

[3] C. Jankowski et al. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proc. IEEE Int'l Conf. on Acoust., Speech and Signal Processing*, pages 109–112, April 1990.

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