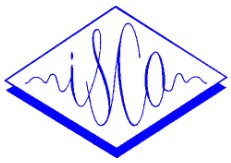

Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

SOME PERSPECTIVES ON SPEECH DATABASE DEVELOPMENT

Lori F. Lamel

LIMSI-CNRS
BP 133
91403 ORSAY Cedex
FRANCE

ABSTRACT

The article, *Speech Database Development: Design and Analysis of the Acoustic Phonetic Corpus* was published in the proceedings of the DARPA Speech Recognition Workshop, held in Palo Alto, February 1986. This article describes some of the issues encountered in the design of the TIMIT database. Below are a few comments related to the design of speech databases, based on the development and subsequent use of TIMIT.

COMMENTS

The following article, *Speech Database Development: Design and Analysis of the Acoustic Phonetic Corpus* was published in the proceedings of the DARPA Speech Recognition Workshop, held in Palo Alto, February 1986. While the paper is several years old, it has never been widely available. It describes some of the issues involved in the design of the TIMIT database, a speech corpus only recently available outside of the United States. The issues presented are still important in the design of speech databases.

The acoustic-phonetic portion of TIMIT was designed to have comprehensive phonemic coverage in a relatively compact set of sentences. As such, the sentences for a portion of the database were hand-selected and the phonemic coverage of the sentences was evaluated using the lexical search program *Alexis*. In addition, an attempt was made to provide adequate frequency of events thought to be difficult to recognize by machine. It was realized that hand-selected sentences are likely to be similar in style – lacking syntactic and semantic variability – so a subset of the database was chosen automatically from a large corpus to provide stylistic variety and complementary phonemic coverage. The hand-crafted sentences were created based on an analysis of phoneme pairs, and the automatically selected ones were chosen by using an objective function to measure the allophonic information in phone triples. However, for both subsets little attention was paid to the word sequences, the sentential position of the phonemes, or the sentence type.

One complaint about the acoustic-phonetic portion of TIMIT is that it does not provide sufficient data to train speaker-dependent recognition systems. It was not designed to; rather, the aim was to provide contexts suitable for observing a large variety of speech

phenomena. Requiring multiple speakers for many of the sentences was necessary to observe the effects of different speaking styles, word pronunciations, and the frequency of optional phenomena.

In developing TIMIT we were faced with a trade-off between the time to create the database and the coverage it would provide. Selecting the sentences to be read was a small portion of the total investment. Database development, even when semi-automated, is extremely labor-intensive. Careful selection of the text material is important to ensure maximum utility of the database and optimum use of the time invested. Besides the selection of the material and the recordings, speech databases should ideally have some form of verification and transcription.

In the attached paper, a summary of some of the phonemic coverage of the database is given. These characteristics, and perhaps others, should be available for all databases. In addition, several environments in which phonological variations were anticipated have been tabulated. It would be very useful to be able to predict the occurrence of different pronunciations and phonological variations from the text corpus. Some correspondences between the predicted environments and realizations for TIMIT will be given.

However, databases obtained solely from read text are insufficient for the development of integrated speech recognition systems. People talk differently when they are reading compared to when they are performing an interactive task. Considering this, some effort must be directed to gathering more realistic data, perhaps using task simulations, for at least a portion of the database. Still, a foundation of read text is useful as it provides a controlled environment for phonetic variability and thus can offer many examples of events occurring rarely in natural speech.