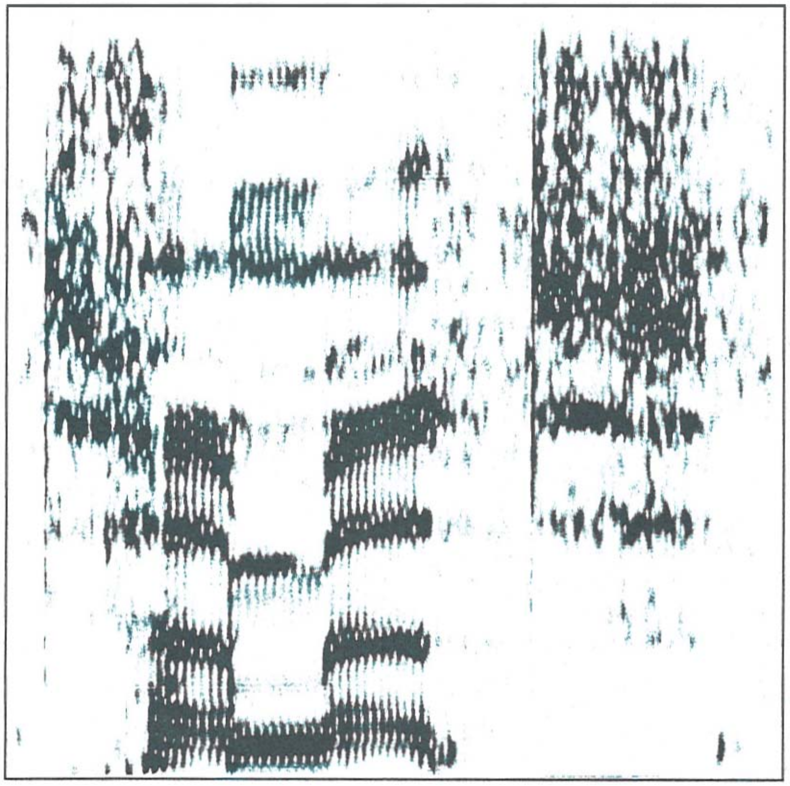
NISTIR 4930

# DARPA TIMIT

## Acoustic-Phonetic Continuous Speech Corpus CD-ROM

NIST Speech Disc 1-1.1



**John S. Garofolo**
**Lori F. Lamel**
**William M. Fisher**
**Jonathan G. Fiscus**
**David S. Pallett**
**Nancy L. Dahlgren**

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

CD-ROM Released October, 1990
Documentation Published February, 1993

NIST

NISTIR 4930

# DARPA TIMIT

## Acoustic-Phonetic Continuous Speech Corpus CD-ROM

NIST Speech Disc 1-1.1

John S. Garofolo
Lori F. Lamel
William M. Fisher
Jonathan G. Fiscus
David S. Pallett
Nancy L. Dahlgren

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

CD-ROM Released October, 1990
Documentation Published February, 1993

**U.S. DEPARTMENT OF COMMERCE**
**Ronald H. Brown, Secretary**

**NATIONAL INSTITUTE OF STANDARDS**
**AND TECHNOLGY**
**John W. Lyons, Director**

# Abstract

The Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains speech from 630 speakers representing 8 major dialect divisions of American English, each speaking 10 phonetically-rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions, as well as speech waveform data for each spoken sentence.

This release of TIMIT contains several improvements over the Prototype CD-ROM released in December, 1988: (1) full 630-speaker corpus, (2) checked and corrected transcriptions, (3) word-alignment transcriptions, (4) NIST SPHERE-headered waveform files and header manipulation software, (5) phonemic dictionary, (6) new test and training subsets balanced for dialectal and phonetic coverage, and (7) more extensive documentation.

The TIMIT CD-ROM has resulted from the joint efforts of several sites under sponsorship from the Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO) [now the Software and Intelligent Systems Technology Office (SISTO)]. Text corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI), and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT, and the data has been verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST).

Certain commercial products are identified in this document in order to adequately specify procedures described. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, the U.S. Department of Commerce, or the United States Federal Government, nor does it imply that the material identified is necessarily the best for the purpose.

# Table of Contents

# List of Tables

# 1 Introduction

The NIST Speech Disc CD1-1.1 contains the complete Texas Instruments/Massachusetts Institute of Technology (TIMIT) acoustic-phonetic corpus of read speech. TIMIT was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT has resulted from the joint efforts of several sites under sponsorship from the Defense Advanced Research Projects Agency -- Information Science and Technology Office (DARPA-ISTO), and Defense Science Office (DARPA-DSO). Text corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI), and Texas Instruments (TI). The speech was recorded at TI, transcribed at MIT, and has been maintained, verified, and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). This publication and the disc were prepared by NIST with assistance by Lori Lamel.

TIMIT contains a total of 6300 utterances, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. 70% of the speakers are male and 30% are female. More information on the selection and distribution of speakers is given in Section 3.1. The recording conditions are described in Section 3.2.

The text material in the TIMIT prompts consists of 2 dialect "shibboleth" sentences designed at SRI, 450 phonemically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. Each speaker read the 2 dialect sentences, 5 of the phonemically-compact sentences, and 3 of the phonetically-diverse sentences. See Section 3.3 for more information on the corpus text material and Section 6 for reprints of publications on the design of TIMIT.

The speech material in TIMIT has been subdivided into dialect-balanced portions for training and testing with complete phonemic coverage. The criteria for the subdivision are described in Section 3.4. A "core" test set contains speech data from 24 speakers, 2 male and 1 female from each dialect region, and a "complete" test set contains 1344 utterances spoken by 168 speakers, accounting for about 27% of the total speech material in the corpus.

Each sentence has an associated orthographic transcription, time-aligned word boundary transcription (provided by NIST, see Section 5.3), and time-aligned phonetic transcription (provided by MIT, see Sections 5.1 and 5.2).

The CD-ROM contains a hierarchical tree-structured directory system which allows the disc to be easily perused. The TIMIT speech and transcription material is located in the "/timit/train" and "/timit/test" directories and online documentation pertaining to the corpus is located in the "/timit/doc" directory. Version 1.5 of the NIST SPeech HEader REsources (SPHERE) software is in the "/sphere" directory and ESPRIT SAM software (CONVERT) to convert TIMIT speech files into a SAM compatible format can be found in the "/convert" directory. Each of these directories contains a "readme.doc" file which may be consulted for further information.

The remainder of this document is structured as follows:

- Section 2 contains a description of the CD-ROM structure and format, including information on how to mount and read the CD-ROM, and a brief description of the SPHERE and CONVERT software.

- Section 3 describes the TIMIT corpus in more detail and the criteria used to divide the speech data into training and test subsets.

- Section 4 provides a description of the accompanying phonemic lexicon and the phonemic and phonetic symbols used in the lexicon and the phonetic transcriptions.

- Section 5 includes a reprint of the article by Seneff and Zue, "Transcription and Alignment of the TIMIT Database", and notes on checking the phonetic transcriptions and on the time-aligned word boundaries.

- Section 6 contains reprints of articles on the design of TIMIT.

# 2 CD-ROM Contents and File Structure

The CD-ROM, NIST Speech Disc CD1-1.1, contains the complete TIMIT acoustic-phonetic speech corpus. Also included on the disc are a new version (1.5) of the NIST SPeech HEader REsources (SPHERE) software and ESPRIT SAM software (CONVERT) for converting TIMIT speech files into a SAM compatible format.

## 2.1 Reading the CD-ROM

The TIMIT CD-ROM (and all NIST speech discs) are formatted according to the ISO-9660 international standard for CD-ROM volume and file structure (ISO, 1988). The ISO-9660 format allows the CD-ROM to be read on any computer platform which supports the standard. To date, ISO-9660 drivers have been implemented for a wide variety of computer systems from personal computers to mainframes. These drivers permit an ISO-9660 disc to emulate a read-only Winchester disk, allowing virtually seamless integration of the CD-ROM. The TIMIT CD-ROM was designed to be usable on any system which supports ISO-9660. The disc contains only data files and source code (with the exception of the PC-executable software in "/convert") which can be easily imported into any speech research environment.

The TIMIT CD-ROM has been designed to be easily browsed or searched programmatically. The TIMIT corpus and documentation (in "/timit") is structured into a directory hierarchy which reflects the organization of the corpus. Several computer searchable text files (in "/timit/doc/*.txt") contain tabular corpus-related information. In addition to TIMIT, a set of software tools "SPeech HEader REsources (SPHERE)" (in "/sphere") is included to ease importation of speech waveform files. The remainder of Section 2 contains more information on the CD-ROM directory and file structure.

## 2.2 CD-ROM Contents

The following files and subdirectories are located in the top-level directory of the CD-ROM. Each of the subdirectories contains a "readme.doc" file which may be consulted for more detailed information.

convert/ - directory containing version 1.2 of the ESPRIT SAM software (CONVERT) for converting TIMIT speech files into a SAM compatible format

/readme.doc - general information file

sphere/ - directory containing version 1.5 of the NIST SPeech HEader REsources (SPHERE) software; SPHERE is a set of "C" library routines and programs for manipulating the NIST header structure prepended to the TIMIT waveform files.

timit/ - directory containing the TIMIT corpus as well as TIMIT-related documentation.

## 2.3 TIMIT Directory and File Structure

This section describes the organization of the files in the "/timit" directory. Descriptions of the file types and a summary of the on-line documentation are given in Sections 2.3.2 and 2.3.3.

### 2.3.1 Organization

On-line documentation and computer-searchable tabular text files are located in the directory "/timit/doc". A brief description of each file in this directory can be found at the end of this section. The speech and associated data are organized on the CD-ROM in the "/timit" directory according to the following hierarchy:

/<CORPUS>/<USAGE>/<DIALECT>/<SEX><SPEAKER_ID>/<SENTENCE_ID>.<FILE_TYPE>

where,

CORPUS :== timit
USAGE :== train | test
DIALECT :== dr1 | dr2 | dr3 | dr4 | dr5 | dr6 | dr7 | dr8
(See Table 3.1 for a description of the dialect codes.)
SEX :== m | f
SPEAKER_ID :== <INITIALS><DIGIT>

where,

INITIALS :== speaker initials, 3 letters
DIGIT :== number 0-9 to differentiate speakers with identical initials

SENTENCE_ID :== <TEXT_TYPE><SENTENCE_NUMBER>

where,

TEXT_TYPE :== sa | si | sx
(See Section 3.2 for the description of sentence text types.)
SENTENCE_NUMBER :== 1 ... 2342

FILE_TYPE :== wav | txt | wrd | phn
(See Table 2.1 for a description of the file types.)

Examples:

/timit/train/dr1/fcjf0/sa1.wav

(TIMIT corpus, training set, dialect region 1, female speaker, speaker-ID "cjf0", sentence text "sa1", speech waveform file)

/timit/test/dr5/mbpm0/sx407.phn

(TIMIT corpus, test set, dialect region 5, male speaker, speaker-ID "bpm0", sentence text "sx407", phonetic transcription file)

### 2.3.2 File Types

The TIMIT corpus includes several files associated with each utterance. In addition to a speech waveform file (.wav), there are three associated transcription files (.txt, .wrd, .phn) for each utterance. These associated files have the form:

```
<BEGIN_SAMPLE> <END_SAMPLE> <TEXT><new-line>
      .
      .
      .
<BEGIN_SAMPLE> <END_SAMPLE> <TEXT><new-line>
```

where,

BEGIN_SAMPLE :== The beginning integer sample number for the segment (Note: the first BEGIN_SAMPLE of each .txt and .phn file is always 0)
END_SAMPLE :== The ending integer sample number for the segment (Note: the last END_SAMPLE in each transcription file may be less than the actual last sample in the corresponding .wav file)

TEXT :== <ORTHOGRAPHY> | <WORD_LABEL> | <PHONETIC_LABEL>

where,

ORTHOGRAPHY :== Complete orthographic text transcription
WORD_LABEL :== Single word from the orthography
PHONETIC_LABEL :== Single phonetic transcription code
(See Section 4.3 for a description of the phone codes.)

Table 2.1: Utterance-associated file types

| File Type | Description |
|---|---|
| .wav | SPHERE-headered speech waveform file. (See Section 2.4 for a description of the speech file manipulation utilities.) |
| .txt | Associated orthographic transcription of the words the person said. (This is usually the same as the prompt, but in a few cases the orthography and prompt disagree.) |
| .wrd | Time-aligned word transcription. The word boundaries were aligned with the phonetic segments using a dynamic programming string alignment program. (See Section 5.3 for information on the alignment procedure.) |
| .phn | Time-aligned phonetic transcription. (See Sections 5.1 and 5.2 for more details on the phonetic transcription protocols.) |

Example transcriptions from the utterance in "/timit/test/dr5/fnlp0/sa1.wav"

Orthography (.txt):

```
0 61748 She had your dark suit in greasy wash water all year.
```

Word label (.wrd):

```
7470 11362 she
11362 16000 had
15420 17503 your
17503 23360 dark
23360 28360 suit
28360 30960 in
30960 36971 greasy
36971 42290 wash
43120 47480 water
49021 52184 all
52184 58840 year
```

Phonetic label (.phn):
(Note: beginning and ending silence regions are marked with h#)

```
0 7470 h#
7470 9840 sh
9840 11362 iy
11362 12908 hv
12908 14760 ae
14760 15420 dcl
15420 16000 jh
16000 17503 axr
17503 18540 dcl
18540 18950 d
18950 21053 aa
21053 22200 r
22200 22740 kcl
22740 23360 k
23360 25315 s
25315 27643 ux
27643 28360 tcl
28360 29272 q
29272 29932 ih
29932 30960 n
30960 31870 gcl
31870 32550 g
32550 33253 r
33253 34660 iy
34660 35890 z
35890 36971 iy
36971 38391 w
38391 40690 ao
40690 42290 sh
42290 43120 epi
43120 43906 w
43906 45480 ao
45480 46040 dx
46040 47480 axr
47480 49021 q
49021 51348 ao
51348 52184 l
52184 54147 y
54147 56654 ih
56654 58840 axr
58840 61680 h#
```

### 2.3.3 On-line Documentation

Compact on-line documentation is located in the "/timit/doc" directory. Files in this directory with a ".doc" extension contain freeform descriptive text, and files with a ".txt" extension contain tables of formatted text which can be searched programmatically. Lines in the ".txt" files beginning with a semicolon are comments and should be ignored on searches. The following is a brief description of each file:

| | |
|---|---|
| phoncode.doc | - List of phone symbols used in the phonemic dictionary and the phonetic transcriptions |
| prompts.txt | - Table of sentence prompts and corresponding sentence-ID numbers |
| spkrinfo.txt | - Table of speaker attributes |
| spkrsent.txt | - Table of sentence-ID numbers for each speaker |
| testset.doc | - Description of the suggested train/test subdivision |
| timitdic.doc | - Description of the phonemic lexicion |
| timitdic.txt | - Phonemic dictionary of all the orthographic words in the prompts |

## 2.4 SPHERE Software (version 1.5)

The NIST SPHERE header format was designed to facilitate the exchange of speech signal data on various media, particularly on CD-ROM. The NIST header is an object-oriented, 1024 byte-blocked structure prepended to the waveform data. See the file "/sphere/headers.doc" for a description of the header format.

NIST SPeech HEader REsources (SPHERE) is a software package for manipulating the NIST-headered speech waveform (.wav) files. The software consists of a library of C-language functions and a set of C-language system-level utilities which can be used to create or modify speech file headers in memory and to read/write the headers from/to disk. See the file, "/sphere/readme.doc", for more information on the SPHERE software, including usage, installation on UNIX systems, and some sample programs.

Please note: The SPHERE library and utilities are modified periodically. The most up-to-date version of the software is available via anonymous ftp from "ssi.ncsl.nist.gov" under the "pub" directory in the compressed tar-formatted file, "sphere-<RELEASE-NUMBER>.tar.Z. Users are encouraged to acquire the most recent source code and documentation.

The SPHERE C-language library contains the following functions:

struct header_t *sp_create_header()

Returns a pointer to an empty header structure.

struct header_t *sp_open_header(fp,parse_flag,error)

Reads an existing header in from file pointer "fp". The file pointer is assumed to be positioned at the beginning of a speech file with a header in NIST SPHERE format. On success, "fp" is positioned at the end of the header (ready to read samples) and a pointer to a header structure is returned. On failure, argument "error" will point to a string describing the problem. If "parse_flag" is false (zero), the fields in the header will not be parsed and inserted into the header structure; the structure will contain zero fields. This is useful for operations on files when the contents of the header are not important, for example when stripping the header.

int sp_clear_fields(h)

Deletes all fields from the header pointed to by "h". Returns a negative value upon failure.

int sp_close_header(h)

Unlinks the header pointed to by "h" and releases the
space allocated for the header. First reclaims all space allocated for the header's fields, if any exist. Returns a negative value upon failure.

int sp_get_nfields(h)

Returns the number of fields stored in the specified header "h". Returns a negative value upon failure.

int sp_get_fieldnames(h,n,v)

Fills in an array "v" of character pointers with addresses of the fields in the specified header "h". No more than "n" pointers in the array will be set. Returns the number of pointers set.

int sp_get_field(h,name,type,size)

Returns the "type" and "size" (in bytes) of the specified header field "name" in the specified header "h". Types are T_INTEGER, T_REAL, T_STRING (defined in "header.h").

The size of a T_INTEGER field is sizeof(long).
The size of a T_REAL field is sizeof(double).
The size of a string is variable and does not include a null-terminator byte (null bytes are allowed in a string).

Returns a negative value upon failure.

int sp_get_type(h,name)

Returns the type of the specified header field "name" of the specified header "h". Types are T_INTEGER, T_REAL, T_STRING (defined in "header.h"). Returns a negative value upon failure.

int sp_get_size(h,name)

Returns the size (in bytes) of the specified header field "name" of the specified header "h".

The size of a T_INTEGER field is sizeof(long).
The size of a T_REAL field is sizeof(double).
The size of a string is variable and does not include a null-terminator byte (null bytes are allowed in a string).

Returns a negative value upon failure.

int sp_get_data(h,name,buf,len)

Returns the value of the specifed header field "name" in the specified header "h" in "buf". No more than "len" bytes are copied; "len" must be positive. It really doesn't make much sense to ask for part of a long or double, but it's not illegal. Remember that strings are not null-terminated. Returns a negative value upon failure.

int sp_add_field(h,name,type,p)

Adds the field "name" to header specified by "h". Argument "type" is T_INTEGER, T_REAL, or T_STRING. Argument "p" is a pointer to a character pointer, or a long integer or a double cast to a character pointer. The specified field must not already exist in the header. Returns a negative value upon failure.

int sp_delete_field(h,name)

Deletes field "name" from header specified by "h". The field must exist in the header. Returns a negative value upon failure.

int sp_change_field(h,name,type,p)

Changes an existing field "name" in header "h" to a new "type" and/or value "p". The field must already exist in the header. Returns a negative value upon failure.

int sp_is_std(name)

Returns TRUE if the specified field "name" is a "standard" field, FALSE otherwise. Standard fields are listed in stdfield.c. The notion of "standard" fields is now archaic.

sp_set_dealloc(n)

Turns on (n<>0) or off (n=0) memory deallocation. The default is on.

int sp_get_dealloc()

Returns the state of memory deallocation.

int sp_write_header(fp,h,hbytes,databytes)

Prints the specified header "h" to stream "fp" in the standard SPHERE header format. The number of bytes in the header block (a multiple of 1024) is returned in "hbytes" and the number of actual header data bytes used is returned in "databytes". Returns a negative value upon failure.

int sp_print_lines(h,fp)

Prints the specified header "h" to stream "fp" in a human-readable format. Returns a negative value upon failure.

int sp_fpcopy(fp,outfp)

Copies stream "fp" to stream "outfp" until end-of-file. Returns a negative value upon failure.

The SPHERE system-level utilities are:

h_read [options] [file ...]
: reads headers from the files listed on the command line; by default, output is lines of tuples consisting of all fieldnames and values; many options modify the program's behavior; see the manual page "h_read.1";

h_add inputfile outputfile
: adds an empty header to the "raw" unheadered speech samples in inputfile and stores the result in outputfile;

h_strip inputfile outputfile
: strips the SPHERE header from inputfile, stores the remaining data in outputfile; if outputfile is "-", writes the sample data to "stdout";

h_edit [-uf] [-D dir] -opchar fieldname=value ... file ...
h_edit [-uf] [-o outfile] -opchar fieldname=value ... file
: edit specified header fields in the specified file(s). In the first form, it either modifies the file(s) in place or copies them to the specified directory "dir". In the second form, it either modifies the file in place or copies it to the specified file "outfile".

The "-u" option causes the original files to be unlinked (deleted) after modification. The "-f" option forces the program to continue after reporting any errors.

The "opchar" must be either "S","I", or "R" to denote string, integer, or real field types respectively.

h_delete [-uf] [-D dir] -F fieldname ... file ...
h_delete [-uf] [-o outfile] -F fieldname ... file
: delete specified header fields in the specified file(s). In the first form, it either modifies the file(s) in place or copies them to the specified directory "dir".

In the second form, it either modifies the file in place or copies it to the specified file "outfile".

The "-u" option causes the original files to be unlinked (deleted) after modification. The "-f" option forces the program to continue after reporting any errors.

Example TIMIT SPHERE-formatted speech waveform header from the waveform file, "timit/train/dr1/fcjf0/sa1.wav":

```
NIST_1A
   1024
database_id -s5 TIMIT
database_version -s3 1.0
utterance_id -s8 cjf0_sa1
channel_count -i 1
sample_count -i 46797
sample_rate -i 16000
sample_min -i -2191
sample_max -i 2790
sample_n_bytes -i 2
sample_byte_format -s2 01
sample_sig_bits -i 16
end_head
```

(The speech data follows the header block.)

## 2.5 Convert Software

The directory "/convert" contains European Strategic PRoject on Information Technology (ESPRIT) Speech input/output Assessment Methodology and Standardization (SAM) Project software (version 1.2) for converting TIMIT speech files to a SAM-compatible format. The software was developed at the Institut de la Communication Parlée, Grenoble, France, in a cooperation with NIST. Some minor modifications to the software were made at NIST to enable the software to run with the TIMIT CD-ROM file structure.

SAM file naming conventions differ from those used in TIMIT. A mapping file "/convert/spkr_map.sam" has been included by NIST on the CD-ROM to be used for automatic filename conversion when the CD-ROM is on-line. The Convert software removes the SPHERE header from the file since SAM speech files contain no header information, and produces 2 SAM files for each TIMIT utterance. The first file is the signal file, and the second contains the orthographic transcription and speaker information. More details about Convert and examples of how to use the package are given in the file "/convert/readme.doc".

# 3 The TIMIT Corpus

The TIMIT corpus of read speech has been designed to provide the speech research community with a standardized corpus for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. The creation of any reasonably-sized speech corpus is very labor intensive. With this in mind, TIMIT was designed so as to balance utility and manageability, containing small amounts of speech from a relatively diverse speaker population and a range of phonetic environments. This section provides more detailed information on the contents of TIMIT and on the division of the TIMIT speech material into subsets for training and testing purposes.

## 3.1 Corpus Speaker Selection and Distribution

TIMIT contains a total of 6300 utterances, 10 sentences spoken by each of 630 speakers from 8 major dialect divisions of the United States. The 10 sentences represent roughly 30 seconds of speech material per speaker. In total, the corpus contains approximately 5 hours of speech. All speakers are native speakers of American English and were judged by a professional speech pathologist to have no clinical speech pathologies. Some speech or hearing abnormalities of subjects are noted in the speaker information file "/timit/doc/spkrinfo.txt" which lists speaker-specific information. In addition to these 630 speakers, a small number of speakers with foreign accents or other extreme speech and/or hearing abnormalities were recorded as "auxiliary" subjects, but they are not included on the CD-ROM.

The speakers were primarily TI personnel, many of whom were new to TI and the Dallas area. They were selected to be representative of different geographical dialect regions of the U.S.[2] A speaker's dialect region was defined as the geographical area of the U.S. where he or she lived during their childhood years (age 2 to 10). The geographical areas correspond with recognized dialect regions of the U.S. (Language Files, Ohio State University Linguistics Dept., 1982), with the exception of the Western dialect region (dr7) in which dialect boundaries are not known with any confidence and "dialect region" 8 where the speakers moved around a lot during their childhood. The dialect regions are illustrated by the lines on the map shown in Figure 3.1. The locale of each speaker's childhood is indicated by a color-coded marker on the map.

TI attempted to recruit speakers who equally represented the 8 dialect regions, but this was found to be impractical given the constraints of time and recording location. As a result, the regions dr1, dr6, and dr8 are less well-represented than the others. Table 3.1 shows the

[2] For more information on American English dialectology see, for example, Atwood, 1980; Bailey and Robinson, 1973; Bronstein, 1960; Davis, 1983; Kurath, 1949; and Williamson and Burke, 1971.

total number of speakers, as well as the number of male and female speakers, for each of the 8 dialect regions. The percentages are given in parentheses.

Table 3.1: Dialect distribution of speakers

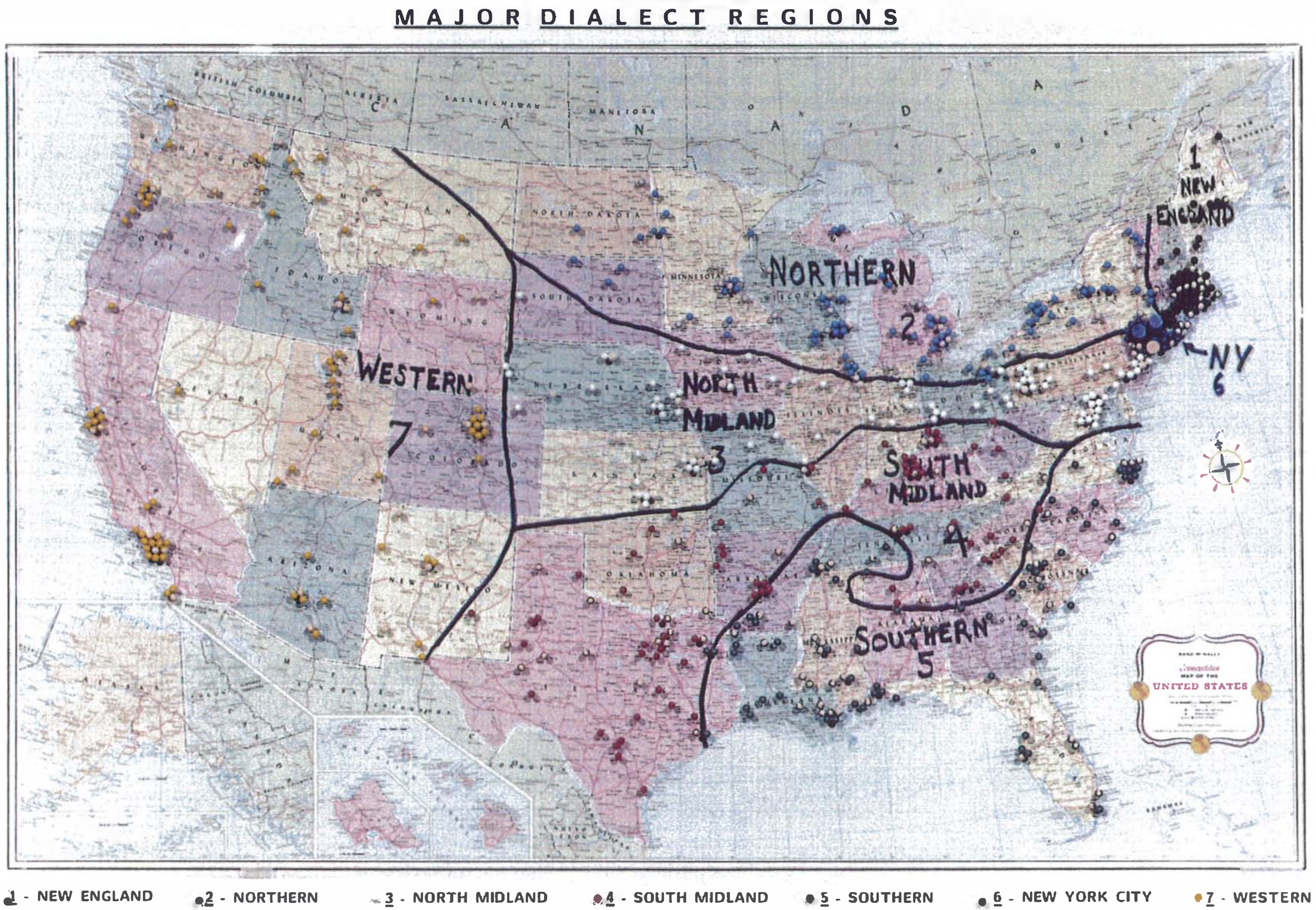
| Dialect Region | | # Male Speakers | # Female Speakers | Total # Speakers |
|---|---|---|---|---|
| Name | Code (dr) | | | |
| New England | 1 | 31 (63%) | 18 (27%) | 49 (8%) |
| Northern | 2 | 71 (70%) | 31 (30%) | 102 (16%) |
| North Midland | 3 | 79 (67%) | 23 (23%) | 102 (16%) |
| South Midland | 4 | 69 (69%) | 31 (31%) | 100 (16%) |
| Southern | 5 | 62 (63%) | 36 (37%) | 98 (16%) |
| New York City | 6 | 30 (65%) | 16 (35%) | 46 (7%) |
| Western | 7 | 74 (74%) | 26 (26%) | 100 (16%) |
| Army Brat (moved around) | 8 | 22 (67%) | 11 (33%) | 33 (5%) |
| Total # Speakers: | | 438 (70%) | 192 (30%) | 630 (100%) |

coretest
test

Table 2: Dialect Distribution of Speakers
in Complete Test Set

| Dialect | #Male | #Female | Total | |
|---|---|---|---|---|
| 1 | 7 | 4 | 11 | (6.5) |
| 2 | 18 | 8 | 26 | (15.5) |
| 3 | 23 | 3 | 26 | (15.5) |
| 4 | 16 | 16 | 32 | (19.0) |
| 5 | 17 | 11 | 28 | (16.7) |
| 6 | 8 | 3 | 11 | (6.5) |
| 7 | 15 | 8 | 23 | (13.7) |
| 8 | 8 | 3 | 11 | (6.5) |
| Total | 112 | 56 | 168 | |

Figure 3.1: Map of TIMIT Dialect Regions



Courtesy of Texas Instruments, Inc.

The on-line file "timit/doc/spkrinfo.txt" contains a table of speaker attributes. For each speaker the information includes the ID (speaker's initials), Sex (male or female), DR (dialect region), Use (train or test), RecDate (recording date), BirthDate, Ht (height), Race, Edu (education level) and optional comments listing interesting speaker attributes or abnormalities.

## 3.2 Recording Conditions and Procedures

Recordings were made in a noise-isolated recording booth at TI, using a semi-automatic computer system (STEROIDS) to control the presentation of prompts to the speaker and the recording. Two-channel recordings were made using a Sennheiser HMD 414 headset-mounted microphone and a Breul & Kjaer 1/2" far-field pressure microphone (#4165). Only the speech data recorded with the Sennheiser microphone is included on this CD-ROM.

The speech was directly digitized at a sample rate of 20 kHz using a Digital Sound Corporation DSC 200 with the anti-aliasing filter at 10 kHz. The speech was then digitally filtered, debiased, and downsampled to 16 kHz. (For more information on the recording conditions and the post-processing of the speech signals see the article by Fisher et al. in Section 6.)

Subjects were seated in the recording booth and prompts were presented on a monitor. The subjects wore earphones through which a low-level (approximately 53 dB SPL) of background noise was played to eliminate the unusual voice quality produced by the "dead room" effect. TI attempted to keep both the recording gain and the level of noise in the subject's earphones constant during the collection. At the beginning of each recording day, a standard calibration tone was recorded from each microphone and the voltage at the subject's earphones was checked and adjusted as necessary.

The speakers were given minimal instructions and asked to read the prompts in a "natural" voice. The recordings were monitored, and any suspected mispronunciations were flagged for verification. Verification consisted of listening to the utterance by both the monitor and the speaker. When a pronunciation error was detected, the sentence was re-recorded. Variant pronunciations were not counted as mistakes.

## 3.3 Corpus Text Material

The text material in the TIMIT prompts, found in the file, "/timit/prompts.doc", consists of 2 dialect "shibboleth" sentences designed at SRI, 450 phonetically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. Table 3.2 summarizes the speech material in TIMIT. The on-line file "/timit/doc/spkrsent.txt" lists the sentence texts read by each speaker.

The dialect sentences (the SA sentences) were meant to expose dialectal variants of the speakers and were read by all 630 speakers. The two dialect sentences are "She had your dark suit in greasy wash water all year." and "Don't ask me to carry an oily rag like that." Some expected variations occur in the pronunciation of the words "greasy" (with an /s/ or /z/) and the vowel color in the word "water". (For a study of such dialectal phenomena see the article by Cohen et al. in Section 6.)

The phonetically-compact sentences (the SX sentences) were hand-designed to be comprehensive as well as compact. The objective was to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. (See the article by Lamel et al. in Section 6 for more information on the design of these sentences.) Each speaker read 5 of these sentences and each text was spoken by 7 different speakers.

The phonetically-diverse sentences (the SI sentences) were selected from existing text sources - the Brown Corpus (Kuchera and Francis, 1967) and a collection of dialogs from recent stage plays (Hultzen et al., 1964) - so as to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts. (See the article by Fisher et al. in Section 6 for more information on the selection of these sentences.) Each speaker read 3 of these sentences, with each sentence being read by only a single speaker.

Table 3.2: TIMIT speech material

| Sentence Type | #Sentences | #Speakers/ Sentence | Total | #Sentences/ Speaker |
|---|---|---|---|---|
| Dialect (SA) | 2 | 630 | 1260 | 2 |
| Compact (SX) | 450 | 7 | 3150 | 5 |
| Diverse (SI) | 1890 | 1 | 1890 | 3 |
| Total: | 2342 | ▨ | 6300 | 10 |

## 3.4 Suggested Training/Test Subdivision

The texts and speakers in TIMIT have been subdivided into suggested training and test sets using the following criteria:

1 - Roughly 20 to 30% of the corpus should be used for testing purposes, leaving the remaining 70 to 80% for training.

2 - No speaker should appear in both the training and testing portions.

3 - All the dialect regions should be represented in both subsets, with at least 1 male and 1 female speaker from each dialect.

4 - The amount of overlap of text material in the two subsets should be minimized; if possible the training set and test set should have no sentence texts in common.

5 - All the phonemes should be covered in the test material; preferably each phoneme should occur multiple times in different contexts.

The next three subsections provide more details on the training and test partitions of TIMIT. In order to ensure adequate coverage in the test subset, the test material was selected from the entire corpus according to the above criteria. Two test sets were selected. The "core" test set, containing a minimal balanced set of test data is described in Section 3.4.1. A description of the larger test set, the "complete" test set is given in Section 3.4.2. After exclusion of the selected test material, the remainder of the corpus was designated as the training set. Some properties of the training partition are specified in Section 3.4.3.

> NOTE: *This subdivision has no correspondence with the original "training" material distributed on the prototype CD-ROM. The original division of training and test material was based ONLY on dialect and sex distribution without other considerations. In contrast, the training and test division on this CD-ROM is based on more factors and is better balanced. Therefore, only the designated training material on CD-ROM "1-1.1" should be used for training purposes.*

### 3.4.1 Core Test Set

Using the above criteria, 2 male speakers and 1 female speaker from each dialect were selected, providing a "core" test set of 24 speakers. Each speaker read a completely different set of 5 SX sentence texts. Since each SI sentence was read by only one speaker, these texts did not impose constraints in selecting the texts or speakers.

The selected texts were checked to ensure that the set included at least one occurrence of each phoneme. The phonemic analysis was based on concatenated phonemic transcriptions of the words in the sentence, not the actual, realized phonetic transcription. Thus, the phonetic allophones found in the test data may be expected to differ from the underlying phonemic forms in accordance with typical phonological variations.

The core test set contains 192 different texts ((5 SX + 3 SI sentences) x 24 speakers). To avoid overlap with the training material the 2 SA sentences have been excluded from the core and complete test sets.

NOTE: *The SA sentences for the test speakers are included on the CD-ROM for completeness. However, they should not be used for training or test purposes if the suggested training and test subsets are used, since they exist for both training and test speakers.*

Table 3.3 lists the speakers in the core test set for each dialect region. This set is the minimum recommended set for test purposes.

Table 3.3: Speakers in the core test set

| Dialect | Male | Female | #Texts/Speaker | Total Texts |
|---|---|---|---|---|
| 1 | DAB0, WBT0 | ELC0 | 8 | 24 |
| 2 | TAS1, WEW0 | PAS0 | 8 | 24 |
| 3 | JMP0, LNT0 | PKT0 | 8 | 24 |
| 4 | LLL0, TLS0 | JLM0 | 8 | 24 |
| 5 | BPM0, KLT0 | NLP0 | 8 | 24 |
| 6 | CMJ0, JDH0 | MGD0 | 8 | 24 |
| 7 | GRT0, NJM0 | DHC0 | 8 | 24 |
| 8 | JLN0, PAM0 | MLD0 | 8 | 24 |
| Total: | 16 | 8 | ▨ | 192 |

### 3.4.2 Complete Test Set

The`"complete" test set was formed by including all 7 repetitions of the SX texts in the core test set. Thus, the utterances from 144 (6x24) additional speakers were added, including the 3 unique SI sentences spoken by each speaker. This insured that no sentence text appeared in both the training and test material. The 168 speakers in the complete test set represent 27% of the total number of speakers in the corpus. The resulting dialect distribution of the complete speaker test set is given in Table 3.4. As in the entire TIMIT corpus, dialects 1, 6, and 8 are less represented than the other dialects.

Table 3.4: Dialect distribution of speakers in complete test set

| Dialect | #Male | #Female | Total |
|---|---|---|---|
| 1 | 7 | 4 | 11 |
| 2 | 18 | 8 | 26 |
| 3 | 23 | 3 | 26 |
| 4 | 16 | 16 | 32 |
| 5 | 17 | 11 | 28 |
| 6 | 8 | 3 | 11 |
| 7 | 15 | 8 | 23 |
| 8 | 8 | 3 | 11 |
| Total: | 112 | 56 | 168 |

The complete test set contains a total of 1344 sentences, 8 sentences from each of the 168 speakers. In this set there are 120 distinct SX texts and 504 different SI texts. Thus, roughly 27% (624) of the texts have been reserved for the test material.

The minimum recommended test material is the core test set, consisting of 2 male speakers and 1 female speaker from each dialect region and 192 unique texts. Those wishing to perform more extensive testing should use the complete test set.

### 3.4.3 Training Set

The training material consists of all the speech data NOT included in either the "core" or "complete" test sets. There are 462 speakers in the training set, comprising 73% of the speakers. The training material contains a total of 4620 utterances, with 10 utterances/speaker. The dialect distribution of the training speakers is given in Table 3.5.

The training material contains 1718 unique texts: the 2 SA texts, 330 different SX texts, and 1386 distinct SI texts. The 2 SA texts were spoken by all the speakers in the corpus. Each of the SX sentence texts were read by 7 speakers, and each SI text was spoken by a single speaker. With the exception of the 2 SA sentences, there is no overlap between the texts read by the test speakers and those read by the training speakers.

*Note: The SA sentences should not be used for training or test purposes if the suggested training and test subsets are used, since they exist for both training and test speakers. Even if they are only used in training, the SA sentences might skew training models since the words contained in them would be over-represented. Their suggested use is for comparative dialectal research.*

Table 3.5: Dialect distribution of speakers in the training set

| Dialect | #Male | #Female | Total |
|---|---|---|---|
| 1 | 24 | 14 | 38 |
| 2 | 53 | 23 | 76 |
| 3 | 56 | 20 | 76 |
| 4 | 53 | 15 | 68 |
| 5 | 45 | 25 | 70 |
| 6 | 22 | 13 | 35 |
| 7 | 59 | 18 | 77 |
| 8 | 14 | 8 | 22 |
| Total: | 326 | 136 | 462 |

### 3.4.4 Distributional Properties of the Training and Test Subsets

Table 3.6 shows some of the distributional properties of the training and test subsets. All of the 45 phonemes are found in the three text subsets, as determined by lookup of each word in the lexicon supplied on the CD-ROM. (See Section 4 for more information on the lexicon.) The total number of distinct words in the TIMIT scripts is 6099. In the core test set 912 distinct words occur, 403 of which also occur in the training texts. The complete test set contains 624 different texts and 2371 distinct words - 1108 of these words also occur in the training texts. Approximately 45% of the words in the texts of the test material also occur in the texts of the training material. The remaining words in the test material are "new". This is due in part to the design of the corpus itself. TIMIT was designed to provide a corpus of acoustic-phonetic speech data for the evaluation of recognition systems at the phonemic level. Because the primary focus in the design of the corpus was the coverage of phonemic elements, emphasis was placed on providing multiple contextual environments for the phonemes during text selection. In order to provide contextual and lexical variation, new words were preferentially chosen over old words during the generation of the phonemically-compact SX sentences. The phonetically-diverse SI sentences were selected so as to maximize allophonic contexts, and thus also favored selection of texts containing new words or word sequences.

Table 3.6: Distributional properties of training and test subsets

| | Entire Corpus | Train | Test | |
|---|---|---|---|---|
| | | | Core | Complete |
| Sentences | 6300 | 4620 | 192 | 1344 |
| Distinct Texts | 2342 | 1718 | 192 | 624 |
| Distinct Words | 6099 | 4891 | 912 | 2371 |
| Distinct Phonemes | 45 | 45 | 45 | 45 |

## 3.5 Transcriptions

The TIMIT corpus includes several transcription files associated with each utterance. These files contain an orthographic transcription, a time-aligned word transcription, and a time-aligned phonetic transcription. Details on the file formats are given in Section 2.3.

The orthographic transcription contains the text of the sentence the speaker said. The orthographic transcription is usually the same as the prompt, but in a few cases they disagree. Word boundaries were assigned using a dynamic programming string alignment program (see Section 5.3) which aligned the word pronunciations found in the lexicon (see Section 4) with the phonetic segments. Information on the phonetic transcription conventions can be found in the article by Seneff and Zue in Section 5.1 and in the notes on checking the phonetic transcriptions in Section 5.2.

# 4 TIMIT Lexicon

The lexicon found in the file "/timit/doc/timitdic.txt" contains entries for all of the words in the TIMIT prompts. There are a total of 6229 entries[3] in the dictionary.[4] The lexicon was derived in part from the MIT adapted version of the Merriam-Webster Pocket Dictionary of 1964 ("pocket") and a preliminary version of a general English dictionary under development at CMU. The pronunciations in the MIT pocket lexicon have been verified and modified over the years. However, many of the words in the TIMIT scripts did not appear in the pocket lexicon, and needed to be added. These include other forms of words found in "pocket" and words not found in any form. Rules were used to generate pronunciations in the former case and the derived pronunciations were hand-checked. In the latter case, consisting mainly of proper names and abbreviated forms (such as "takin'" instead of "taking" or "'em" for "them"), the pronunciations were added by hand.

The symbols in the lexical representation are abstract, quasi-phonemic marks representing the underlying sounds and typically correspond to a variety of different sounds in the actual recordings. The term *quasi-phonemic* is used because some differences represented in the lexicon are not phonemically distinctive in English, such as the /er/~ /axr/ in which /er/ co-occurs with stress.

The term *quasi-phonemic* is used because some differences represented in the lexicon are not phonemically distinctive in English, such as the contrast between /er/ and /axr/. (Since the former always occurs with stress and the latter never occurs with it, as in "burner" /b er1 n axr/, the two are in complementary distribution and could be considered different allophones of the same phoneme.)

## 4.1 Format of the Lexicon

All entries have been converted to lower case. Stress is represented as a "1" for primary stress and a "2" for secondary stress, appended to the end of the vowel symbol. Hyphenated words such as "head-in-the-clouds" can be found both as a single entry and as the individual words "head", "in", "the", and "clouds", which result when the hyphens are replaced by spaces. This was done to allow more flexibility in the parsing of sentences into their constituent lexical items. If these parts of hyphenated words occur only as bound forms and never as free words, the hyphen is left in their entry, as in "knick-" and "-knack" from "knick-knack." Due to vagaries of English orthography, this procedure sometimes results in lexical entries that are neither words nor proper constituents of words, such as

[3] This number is greater than the number of distinct words listed in Table 3.6 because the dictionary includes entries for compound words and their components.

[4] The terms "lexicon" and "dictionary" are used interchangeably throughout this publication.

"-upmanship" from "one-upmanship".

One pronunciation is provided per entry except in the case where the same orthography corresponds to different parts of speech with different pronunciations, ***and*** both forms exist in the TIMIT prompts. To differentiate these words, multiple entries are given, with the syntactic class following the symbol ~. The classes found in the lexicon are:

| | |
|---|---|
| ~n | noun |
| ~v | verb |
| ~adj | adjective |
| ~pres | present tense |
| ~past | past tense. |

An example is the word "live", with the entries:

live~v /l ih1 v/
live~adj /l ay1 v/

## 4.2 Pronunciation Conventions

The pronunciation is specified using the the "CMU" symbol set (see Section 4.3 for a description of the symbols). While we realize that representing only one pronunciation is often not sufficient to cover commonly observed pronunciations, many of the alternate pronunciations may be predicted by use of phonological rules and may be highly dialect dependent. Using only one pronunciation per word forced the somewhat vexing decision of which one to use. We did not put extensive study into such issues, and do not make any claims of the theoretical correctness of our decisions on particular words. Our tendency has been to use the more marked alternate because we think it is harder to predict. We tried to make the pronunciations as consistent as possible. In a number of cases we referred to the authorities Kenyon and Knott (1953) and Webster's Third New International Dictionary (1966).

### 4.2.1 Vowel Variability

Many of the pronunciation differences for vowels occur in semi-vowel environments and in unstressed syllables.

- The vowel in words like "for", "pour", and "more" are often represented using either the vowel /ow/ or the vowel /ao/. This lexicon uses /ao/.

- The vowel in words like "air" and "care" has been represented using /ae/, to differentiate this vowel from the /eh/ in "berry". Some speakers actually make a three-way distinction

("Mary", "merry", "marry"), with the vowel in Mary being somewhat in between an /eh/, /ae/, and /ey/. These speakers may use the same vowel in words like "care".

- The vowel /ih/ (as opposed to /iy/) has been systematically used in the representation of words like "fear" and "year".

- unstressed schwa alternation /ix/~/ax/: /ix/ is usually used for schwas between 2 alveolars ("roses" /r ow1 z ix z/), otherwise /ax/ is used ("ahead" /ax hh eh1 d/).

- /r/ following the diphthongs /aw/ ("hour") and /ay/ ("fire") has been represented as /axr/, except where the /r/ is syllable-initial as in words like "irate" and "virus".

- vowel reduction: In some cases the pronunciation of a word may alternate between a full vowel and a highly reduced one. In these cases, preference was given to the pronunciation with the more marked vowel instead of the schwa. For example, the pronunciation of "accept" is given as /ae k s eh1 p t/, not /ax k s eh1 p t/.

### 4.2.2 Stress Differences

- /er/~/axr/ alternation -- /axr/ is used in unstressed syllables and /er/ in stressed syllables.

- /ih/~/ix/, /ah/~/ax/ -- once again the distinction is based on stress. The forms /ix/ and /ax/ are used in unstressed syllables.

- /y uw/~/y uh/ -- the tendency is to use /y uw/ in stressed positions as in "attribution" /ae t r ih b y uw1 sh ix n/, and /y uh/ in unstressed position as in "attribute~v" /ax t r ih2 b y uh t/.

### 4.2.3 Syllabics

The syllabics /em/, /en/, and /el/ are used frequently in the phonemic representations even though they may be pronounced as a sequence of a schwa followed by /m/, /n/, or /l/. For example, words ending in "-ism" are represented as /ih z em/ even though a short schwa often appears in the transition from the /z/ to the /em/.

- /en/ must follow a coronal, except in rare occurrences such as "cap'n" /k ae1 p en/ and "haven't" /h ae1 v en t/.

- in general, the syllabic /el/ is used instead of /ax l/ except before a stressed vowel. Some exceptions are found in words ending in the "-ly" suffix. For example, "angrily" is represented /ae1 ng g r ax l iy/, not /ae1 ng g r el iy/. The only occurrences of /el l/ are found in compound words such as "jungle-like" and "liberal-led".

## 4.3 Phonetic and Phonemic Symbol Codes

This following table contains a list of all the phonemic and phonetic symbols used in the TIMIT lexicon and in the phonetic transcriptions. These include the stress markers {1,2} found only in the lexicon and the following symbols which occur only in the transcriptions:

1) the closure intervals of stops which are distinguished from the stop release. The closure symbols for the stops /b,d,g,p,t,k/ are /bcl,dcl,gcl,pcl,tck,kcl/, respectively. The closure portions of /jh/ and /ch/ are /dcl/ and /tcl/.

2) allophones that do not occur in the lexicon. The use of a given allophone may be dependent on the speaker, dialect, speaking rate, and phonemic context, among other factors. Since the use of these allophones is difficult to predict, they have not been used in the phonemic transcriptions in the lexicon.

- flap /dx/, as in words "muddy" or "dirty"
- nasal flap /nx/, as in "winner"
- glottal stop /q/, which may be an allophone of /t/, or may mark an initial vowel or a vowel-vowel boundary
- voiced-h /hv/, a voiced allophone of /h/, typically found intervocalically
- fronted-u /ux/, an allophone of /uw/, typically found in an alveolar context
- devoiced-schwa /ax-h/, a very short, devoiced vowel, typically seen when reduced vowels are surrounded by voiceless consonants

3) other symbols include two types of silence: "pau", marking a pause; "epi", denoting the epenthetic silence often found between a fricative and a semivowel or nasal, as in "slow"; and "h#", used to mark the silence and/or non-speech events found at the beginning and end of the signal.

|  | Symbol | Example Word | Possible Phonetic Transcription | Comment |
|---|---|---|---|---|
| Stops: | b | bee | BCL B iy |  |
|  | d | day | DCL D ey |  |
|  | g | gay | GCL G ey |  |
|  | p | pea | PCL P iy |  |
|  | t | tea | TCL T iy |  |
|  | k | key | KCL K iy |  |
|  | dx | muddy, dirty | m ah DX iy, dcl d er DX iy | flap |
|  | q | bat | bcl b ae Q | glottal stop |
| Affricates: | jh | joke | DCL JH ow kcl k |  |
|  | ch | choke | TCL CH ow kcl k |  |
| Fricatives: | s | sea | S iy |  |
|  | sh | she | SH iy |  |
|  | z | zone | Z ow n |  |
|  | zh | azure | ae ZH er |  |
|  | f | fin | F ih n |  |
|  | th | thin | TH ih n |  |
|  | v | van | V ae n |  |
|  | dh | then | DH eh n |  |
|  | m | mom | M aa M |  |
|  | n | noon | N uw N |  |
|  | ng | sing | s ih NG |  |
|  | em | bottom | b aa dx EM |  |
|  | en | button | b ah q EN |  |
|  | eng | washington | w aa sh ENG tcl t ax n |  |
|  | nx | winner | w ih NX axr | nasal flap |
| Semivowels and Glides: | l | lay | L ey |  |
|  | r | ray | R ey |  |
|  | w | way | W ey |  |
|  | y | yacht | Y aa tcl t |  |
|  | hh | hay | HH ey |  |
|  | hv | ahead | ax HV eh dcl d |  |
|  | el | bottle | bcl b aa dx EL |  |

| | | | |
|---|---|---|---|
| Vowels: | iy | beet | bcl b IY tcl t |
| | ih | bit | bcl b IH tcl t |
| | eh | bet | bcl b EH tcl t |
| | ey | bait | bcl b EY tcl t |
| | ae | bat | bcl b AE tcl t |
| | aa | bott | bcl b AA tcl t |
| | aw | bout | bcl b AW tcl t |
| | ay | bite | bcl b AY tcl t |
| | ah | but | bcl b AH tcl t |
| | ao | bought | bcl b AO tcl t |
| | oy | boy | bcl b OY |
| | ow | boat | bcl b OW tcl t |
| | uh | book | bcl b UH kcl k |
| | uw | boot | bcl b UW tcl t |
| | ux | toot | tcl t UX tcl t |
| | er | bird | bcl b ER dcl d |
| | ax | about | AX bcl b aw tcl t |
| | ix | debit | dcl d eh bcl b IX tcl t |
| | axr | butter | bcl b ah dx AXR |
| | ax-h | suspect | s AX-H s pcl p eh kcl k tcl t |

| | Symbol | Description |
|---|---|---|
| Others: | pau | pause |
| | epi | epenthetic silence |
| | h# | begin/end marker (non-speech events) |
| | 1 | primary stress |
| | 2 | secondary stress |

## 4.4 Errata

A few errors were found in the phonemic lexicon file, "/timit/doc/timitdic.txt", after the CD-ROM was pressed. The corrections are as follows:

1. delete
   "-knacks /n ae1 k s/"
   "-upmanship /ah1 p m ax n sh ih p/"
   "-ups /ah p s/"
   "-zagged /z ae1 g d/"
   "bodied /b aa1 d iy d/"
   (These aren't words or combining forms.)

2. change "castorbeans /k ae1 s axr b iy1 n z/"
   to "castorbeans /k ae1 s t axr b iy1 n z/"

3. change "fast-closing /f ae1 s t ao l ow1 z ix ng/"
   to "fast-closing /f ae1 s t k l ow1 z ix ng/

4. change "cloverleaf /ao l ow1 v axr l iy2 f/"
   to "cloverleaf /k l ow1 v axr l iy2 f/"

5. change "constantly /ao aa1 n s t ix n t l iy/"
   to "constantly /k aa1 n s t ix n t l iy/"

6. change "countryside /ao ah1 n t r iy s ay2 d/"
   to "countryside /k ah1 n t r iy s ay2 d/"

7. change "nancy's /n ae1 n ao iy z/"
   to "nancy's /n ae1 n s iy z/"

8. change "singer's /s ih1 ng g axr z/"
   to "singer's /s ih1 ng axr z/"

9. change "uncomfortable /ah n ao ah1 m f axr t ax b el/"
   to "uncomfortable /ah n k ah1 m f axr t ax b el/"

10. change "backward /b ae1 k w er d z/"
    to "backward /b ae1 k w er d/"

11. change "cleaners /al l iy1 n axr z/"
    to "cleaners /k l iy n axr z/"

12. change "cruelty /k r uw1 l iy/"
    to "cruelty /k r uw1 l t iy/"

13. change "detectable /d ih t eh1 k ax b el/"
    to "detectable /d ih t eh1 k t ax b el/"

14. change "distinct /d ih s t ih1 ng t/"
    to "distinct /d ih s t ih1 ng k t/"

15. change "ellipsoids /ax l ih1 p s oy d/"
    to "ellipsoids /ax l ih1 p s oy d z/"

16. change "entity /eh1 n ix t iy/"
    to "entity /eh1 n t ix t iy/"

17. change "halloween /hh ae2 l ow iy1 n/"
    to "halloween /hh ae2 l ow w iy1 n/"

18. change "headquarters /hh eh1 d k w ao2 t axr z/"
    to "headquarters /hh eh1 d k w ao2 r t axr z/"

19. change "identified /ay d eh1 n t ix f ay2/"
    to "identified /ay d eh1 n t ix f ay2 d/"

20. change "instinct /ih1 n s t ih2 ng t/"
    to "instinct /ih1 n s t ih2 ng k t/"

21. change "musical /m uw1 z ih k el/"
    to "musical /m y uw1 z ih k el/"

22. change "presented /p r ax z eh1 t ix d/"
    to "presented /p r ax z eh1 n t ix d/"

23. change "unwaveringly /ah n w ey1 v axr ix ng/"
    to "unwaveringly /ah n w ey1 v axr ix ng l iy/"

These following are less surely errors, but probably should be fixed:

1. change "photochemical /f ow2 t ax k eh1 m ix k el/"
   to "photochemical /f ow2 t ow k eh1 m ix k el/"

2. change "photographs /f ow1 t ow g r ae2 f s/"
   to "photographs /f ow1 t ax g r ae2 f s/"

3. change "reorganization  /r iy2 ao r g ix n ay z ey1 sh ix n/"
   to "reorganization  /r iy ao2 r g ix n ay z ey1 sh ix n/"

4. change "tyranny  /t ih1 r ae n iy/"
   to "tyranny  /t ih1 r ax n iy/"

# 5 Transcription Protocols

This section includes information pertaining to the protocols used in obtaining the transcription files associated with each utterance. Section 5.1 reprints an article on the phonetic transcription methodology. Additional details are given in the notes in Section 5.2. The word boundary alignment procedure is described in Section 5.3.

## 5.1 Reprint of a Publication Describing TIMIT Transcription Conventions

This section contains a reprint of the article "Transcription and Alignment of the TIMIT Database," by Stephanie Seneff and Victor W. Zue. The paper was presented at *The Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, Hawaii, November 20-22, 1988.

# Transcription And Alignment of the Timit Database

Victor W. Zue and Stephanie Seneff

Spoken Language Systems Group,
Laboratory for Computer Science,
Massachusetts Institute of Technology,
Cambridge, MA 01890, USA.

## ABSTRACT

The TIMIT acoustic-phonetic database was designed jointly by researchers at MIT, TI and SRI. It was intended to provide a rich collection of acoustic phonetic and phonological data, to be used for basic research as well as the development and evaluation of speech recognition systems. The database consists of a total of 6,300 sentences from 630 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. This paper describes the transcription and alignment of the TIMIT database, which was performed at MIT.

## 1. BACKGROUND

When the DARPA Strategic Computing speech program was first formulated in 1984, the consensus of the research community was that the amount of speech data available is woefully inadequate. As a result, a significant effort on database development was mounted in order to provide the research community with a large body of acoustic data for research, system development, and performance evaluation. One such database is the so-called TIMIT acoustic-phonetic database. The TIMIT database was designed jointly by researchers at MIT, TI, and SRI. It consists of a total of 6,300 sentences from 630 speakers, representing over 5 hours of speech material, and was recorded by researchers at TI. This paper describes the transcription and alignment of the TIMIT database, which was performed by researchers at MIT.

Each speaker in the TIMIT database recorded 10 sentences drawn from three different corpora as follows. Each speaker read two sentences, designated as S1 and S2, which were designed by Jared Bernstein of SRI in order to compare dialectal and phonological variations across speakers. Five sentences, designated as SX sentences, were drawn from a small set of sentences designed at MIT. The remaining three sentences for each speaker, designated as SI sentences, were selected from the Brown corpus by Bill Fisher of TI [1].

There are a total of 450 "MIT" sentences used in the TIMIT database. These were generated by hand in an iterative fashion, with the goal that they should be phonetically rich. Care was taken to have as complete a coverage of left- and right- context for each phone as possible. Some of the more problematic sequences, such as vowel-vowel and stop-stop, were particularly emphasized. An attempt was also made to ensure that many of the frequently-occurring low-level phonological rules were adequately represented. To aid in the sentence generation process, we made use of an on-line, Webster's Pocket Dictionary containing nearly 20,000 words. Words or word-sequences containing particular phone pairs could be accessed from this dictionary automatically, which greatly facilitated the database design process. We performed a detailed analysis of the resulting sentence set, as well as the SI sentences that make up the remainder of the database . The interested reader should consult Lamel et al.[3] for further information about the corpora.

## 2. THE ACOUSTIC PHONETIC LABEL SET

All of the recorded sentences were provided with a time- aligned sequence of acoustic-phonetic labels. The label set is intended to represent a level somewhat intermediate between phonemic and acoustic. Our motivation was that clear acoustic boundaries in the waveform should all be marked, and that the criteria for positioning the boundaries between units should in part be based on our ability to mark them consistently. Table I lists all of the acoustic-phonetic labels that were used. Most of these labels are phonemic, although several symbols have been included for labelling acoustically distinct allophones as well as other special acoustic events.

### 2.1 Stops

Stops are characterized by a sequence of two events: a closure and a release. This departure from phonemic form is, we believe, important in order to preserve a boundary marking the onset of the release. There are six closure symbols for the stops. The closure region for affricates is identical with that of the corresponding alveolar stop. (e.g., the /č/ in "char" is represented as [tᵒč]).

There are two major allophones for the stops. The glottal stop, [ ʔ ], is often inserted preceding a word-initial vowel. Sometimes a /t/ can also be realized as a glottal stop, as in "cotton". The symbol [ ɾ ] is used to label a flap, which can either be an underlying /t/ or /d/. We make a separate flapping decision for every phonemic /t/ and /d/, based on listening and the spectrographic evidence. We allow flapping to occur in environments for which theory is violated, if in fact we believe that flap is what was heard/seen.

### 2.2 Nasal and Semivowels

We recognize four allophones for the nasals, three of them are the syllabics, [ m̩, n̩, ŋ̍ ]. If there is any evidence of a preceding schwa, the non-syllabic form is preferred. The alveolar nasal, /n/ can be realized as a nasal flap, denoted by the symbol [ ɾ̃ ]. Sometimes an underlying /nt/ sequence is realized as a nasal flap, as in "*ent*ertain".

The liquid, /l/, has a syllabic allophone, denoted as [l̩]. Again, a non-syllabic form is preferred whenever a preceding schwa is observed.

### 2.3 Vowels

Two vowels, /i o/, are represented by symbols that included their corresponding off-glides. This is because they are usually realized as diphthongs in American English. The four diphthongs, /ɑʸ/, /ɑʷ/, /ɔʸ/, and /eʸ/, are each represented as a single label, with no separate region defined for the off-glide portion. The retroflexed vowel /ɝ / is also represented as a single unit. This represents a departure from the International Phonetic Alphabet, which would represent this steady-state vowel as the sequence /ʌr/.

Reduced vowels are represented by four separate allophones: back schwa ([ ə ]), front schwa ([ ɨ ]), retroflexed schwa ([ ɚ ]), and voiceless schwa ([ ə̥ ]). The decision for [ ə ] vs [ ɨ ] is based on whether the second formant is closer to the first or to the third. A low third formant leads to / ɚ /. Schwas can often be devoiced in words such as "secure".

English does not distinguish phonemically between the fronted vowel /ʉ/ and the standard back /u/; however the difference in $F_2$ for the two forms can be as much as 800 Hz. We felt it was unsatisfactory to group two forms with such diverse formant frequencies into the same vowel category. The decision is made as for schwa: if $F_2$ is closer to $F_1$, it's considered a back /u/. Similar trends of fronting are also observed for /o/ and /ʊ/ in certain environments; however, the effect is most dramatic for /u/.

At present, we make no attempt to provide further sub-phonemic characterizations for vowels other than this front/back distinction for /u/ and the four schwas. For instance, many vowels are nasalized when they are followed by a nasal, or lateralized when followed by an /l/. Such information would surely be useful, but the decision-making process is prone to judgement error, and would require a significant increase in time and effort.

## 2.4 Others

We make a distinction between two types of /h/: voiced ([ ɦ ]) and unvoiced ([h]). The decision is based mainly on an examination of the waveform for clear low-frequency periodicity, and spectrogram for voicing striations. The voiced form is most common between two vowels.

Our label set includes a category "epenthetic silence," ], which we use to mark acoustically distinct regions of weak energy separating sounds that involve a change in voicing. These short gaps are typically due to articularory timing errors. The most common occurrences of such gaps are between an /s/ and a semivowel or nasal, as in "small", "swift", or "prince". Two other non-phonetic symbols are included: # is used to mark regions preceding and following a sentence, and □ is used to mark pauses within a sentence.

## 3. CRITERIA FOR BOUNDARY ASSIGNMENTS

The acoustic-phonetic transcription for the TIMIT sentences is time aligned with the speech waveform. The alignment is useful in that specific acoustic events can be accessed conveniently based on the transcription. We must stress, however, that the aligned transcription is intended to establish a *correspondence* between the transcription and important acoustic landmarks. One should not directly associate a region between two time markers as a distinct phonetic unit, since the encoding of phonetic information in the speech signal is extremely complicated.

In most cases, the boundaries between two acoustic-phonetic events are clear and well-defined, such as that between a stop closure and its release. However, there are a number of cases where the exact placement of a boundary is problematic (as is the case between a semivowel and a vowel), or cases where it's not clear whether a region should be represented as one or two acoustic-phonetic units (as is the case for diphthongs). In these cases, we tried to define a set of criteria that would be systematic and least subject to human error, in order to produce boundary positionings that were as consistent as possible.

As mentioned previously, we decided that the boundary between the closure interval and the release of a stop is an important one that should be assigned. It is certainly a very distinct landmark in the waveform. Anyone interested in studying the burst characteristics of a stop would then be able to focus on just that region that includes only the released portion. In a strictly phonemic representation, the closure and release would be represented as a single unit, and therefore that critical boundary would remain unmarked.

A problematic boundary is one that separates a prevocalic stop from a following semivowel, as in "*t*ruck." Typically, part of the /r/ is devoiced, and therefore is absorbed into the aspiration portion of the stop. If listening were the only criterion, then the left boundary of the /r/ would occur somewhere in the aspiration, and the right boundary would occur somewhere after voicing onset. A clear acoustic boundary at the point of voice onset would remain unmarked. It would also be difficult to decide where to mark the boundary between the stop burst and the aspirated /r/ portion. Since voice-onset time (VOT) is a parameter that has been a focus of many research efforts, it seems unsatisfactory not to include a reliable mechanism for measuring VOT based on the labelled boundaries. Therefore, we adopted the policy of always absorbing into the stop release all of the unvoiced portion of a following vowel or semivowel.

The boundary between many semivowels and their adjacent vowels is rather ill-defined in the waveform and spectrogram, because transitions are slow and continuous. It is not possible to define a single point in time that separates the vowel from the semivowel. In such cases, we decided to adopt a simple heuristic rule, in which one-third of the vocalic region is assigned to the semivowel, thus giving the vowel twice the duration of the adjacent semivowel. Previous investigators have also made use of such consistent rules for defining acoustically ambiguous boundaries [4].

One obscure condition is a /ts/ or /dz/ sequence, where typically there is little or no spectral change to characterize a boundary between the homorganic stop and fricative, yet the onset of acoustic energy of the unit is sufficiently abrupt such that a /t/ is heard. Our convention here is that, if a clear /t/ is heard, the early portion of the /s/ is marked as a /t/ release.

When gemination occurs, we do not attempt to mark a boundary between the two units. This situation occurs exclusively at word boundaries, as in "so*me m*oney." Furthermore, in the case of a stop-stop sequence where the first stop is unreleased, the closure interval is assigned to the first stop and the release to the second one.

## 4. PROCEDURES FOR TRANSCRIPTION AND ALIGNMENT

The transcription and alignment process involves three stages:

1. An acoustic-phonetic sequence is entered manually by a transcriber as a string.
2. The speech waveform is aligned automatically with the acoustic-phonetic sequence, using an alignment program developed at MIT.
3. The boundaries generated automatically are then hand corrected by experienced acoustic phoneticians.

### 4.1 Transcription

In both stages 1 and 3, the labeller makes her/his acoustic-phonetic decision based on careful listening of portions of the speech waveform, as well as visual examination using displays such as the spectrogram and the original waveform. The process takes place within the SPIRE software facility for speech analysis, a powerful interactive tool that is well-matched to this task [2]. Stage 1 requires less intensive use of SPIRE than stage 3, because it is only necessary to record what was heard, without identifying the time locations of the events. Furthermore, minor errors of judgement made at this stage can be readily corrected in stage 3. The labels can be entered either by typing or by mousing a displayed set. Figure 1 shows the SPIRE layout used for entering the transcription. The completed transcription is shown in the top window of this display.

In general, we try to label what we hear/see, rather than what we expect. Thus, if a person says "imput" for "input," the nasal will be marked as an /m/. However, in conditions of ambiguity, the underlying phonemic form is selected preferentially.

4.2 Automatic Alignment

The alignment of a phonetic transcription with the corresponding speech waveform is essential for making use of the database in speech research, since time-aligned phonetic transcriptions provide direct access to specific phonetic events in the waveform. Traditionally, this alignment is done manually by a trained acoustic-phonetician. This is an extremely time-consuming procedure, requiring the expertise of one or a very small number of people. Therefore, the amount of data that can be labeled is limited. In addition, manual labeling often involves decisions which are highly subjective, and thus the results can vary substantially from one person to another.

Transcription alignment of the TIMIT database makes use of CASPAR, an automatic alignment system developed at MIT. Descriptions of preliminary implementation of the system can be found elsewhere [5,6]. Basically, the alignment is accomplished by the system in three steps. First, each 5 ms frame of the speech data is assigned to one of five broad-class labels: *sonorant, obstruent, voiced-consonant, nasal/voicebar*, and silence, using a nonparametric pattern classifier. The assignment process makes use of a binary decision tree, based on a set of acoustically motivated features. Each sequence of identically-labelled frames is then collapsed into a segment of the same label, thus establishing a broad-class segmentation of the speech. The output of the initial classifier is then aligned with the phonetic transcription using a search strategy with some look-ahead capability, guided by a few acoustic phonetic rules. For those segments which correspond to two or more phonetic events after preliminary alignment, further segmentation is done using specific algorithms based on knowledge of the phonetic context. In some cases heuristic rules are invoked (as between a vowel and a semivowel) to assign consistent, but somewhat arbitrary boundaries.

Over the past two years, two major modifications of CASPAR have taken place. First, the alignment of the broad-class acoustic labels with the phonetic symbols has been cast into a probabilistic framework. By using a large body of training data, a set of robust, context-dependent and durational statistics were obtained. Second, a fourth module has been added to the system to improve the resolution of the boundaries. This module computes appropriate acoustic attributes at a high analysis rate using different window shapes that depend on the specific context. The boundaries are then adjusted on these attributes.

In a formal evaluation, it was found that CASPAR can correctly perform over 95% of the labelling task previously done by human transcribers. The boundary locations produced by the system agree well with those produced by human transcribers. For example, over 75% of the automatically generated boundaries were within 10 msec of a boundary entered by a trained phonetician.

Figure 2 displays the output for the sentence, "She had your dark suit in greasy wash water all year." The transcription and boundaries are overlaid on the spectrogram for ease of examination.For this example, most of the boundaries have been found correctly by CASPAR. Note, however, that boundaries are missing in the [iɦæ] sequence of "She had." The waveform displays the word "dark" and the [s] of "suit." Note that the initial boundary of the first [d] is slightly too far forward in time.

### 4.3 Post-Processing

The final step is to correct by hand any errors in the automatically aligned acoustic-phonetic sequence. Some of the errors are due to the fact that CASPAR is not able to determine certain boundaries, such as some of those between two vowels. In other cases the boundaries may have been misplaced.

Hand correction of the aligned transcription is based on critical listening of portions of the utterance as well as visual examination of the spectrogram and the waveform. The spectrogram covers close to 3 seconds worth of speech at one time, whereas the waveform is displayed on a much more expanded time scale. For example, to accurately mark the onset of the release of a stop, the cursor is first positioned on the spectrogram at the approximate point in time. The waveform display automatically moves to synchronize in time with the cursor, and a fine-tuning of the boundary can be achieved by mousing the exact time point in the waveform.

The mouse can be used with ease to move an existing boundary to a new point in time, to erase a boundary, or to insert a boundary. Furthermore, a specified mouse click on any segment allows the labeller to change the acoustic-phonetic label associated with that segment. This step is sometimes necessary to correct an error of judgement in stage 1.

An example of the screen layout used for the correction process is shown in Figure 3. The boundary for the [d] burst onset has been corrected. Missing boundaries were inserted for the [iɦæ] sequence. In addition, the boundaries associated with the first [w] were extended on both sides, and an epenthetic silence was inserted between the [ ǯ ] and the following [w].

## 5. CONCLUDING REMARKS

Once the acoustic-phonetic transcription has been aligned, it is rather straightforward to propagate the alignment up to the orthographic transcription as well as the intermediate phonemic transcription. A time-aligned orthographic transcription is useful when searching for a specific word, while a time-aligned phonemic transcription can be used to relate the lexical representation of words to their acoustic realizations. For example, the lexical representation of the word sequence "gas shortage" contains a word-final /s/ and a word-initial / ǯ /, whereas its acoustic realization may simply be a long [ǯ]. In this case, the time-aligned phonemic transcription will map the long to [ǯ] both the underlying fricative. Researchers interested in studying the frequency of occurrence of certain low-level phonological rules will thus be able to derive the information from these transcriptions.

We have developed a system that maps a time-aligned acoustic-phonetic transcription to the phonemic and orthographic transcriptions [7]. However, the alignment effort for these transcriptions lags somewhat behind the phonetic alignment. In the interest of expeditiously making as much data available to the interested parties, we have decided to provide these other transcriptions in future releases.

The transcription and alignment of the TIMIT database is a sizable project. At this writing, all of the sentences have been processed and delivered to the National Bureau of Standards. A significant portion of the database is now available to the general public via magnetic tapes, and plans for distributing them by way of compact disc is well under way. Despite our best intention to provide as correct a set of transcriptions as possible, however, errors undoubtedly exists. We urge users of this database to communicate errors to us whenever possible, so that future users can benefit from this effort.

Finally, we would like to thank Dave Pallett, Jim Hieronymus, and their colleagues at NBS for the cooperation, patience, and good humor that they provided. Their help, particularly regarding data transfer, verification, distribution, and fending off eager inquiries, have been indispensable to this project.

The development of the TIMIT database at MIT was supported by the DARPA-ISTO under contract N00039-85-C-0341, as monitored by the Naval Space and Warfare Systems Command. Major participants of the project at MIT include Corine Bickley, Katy Isaacs, Rob Kassel, Lori Lamel, Hong Leung, Stephanie Seneff, Lydia Volaitis, and Victor Zue.

## REFERENCES

[1] Fisher, W.M. and G.R. Doddington, "The DARPA Speech Recognition Research Database: Specification and Status," Proceedings of the DARPA Speech Recognition Workshop, Palo Alto, CA, February 19-20, 1986, pp. 93-99.

[2] Zue, V.W., D.S. Cyphers, R.H. Kassel, D.H. Kaufman, H.C. Leung, M.A. Randolph, S. Seneff, J.E. Unverferth, III, and T. Wilson, "The Development of the MIT LISP-Machine Based Speech Research Workstation," Proceedings of ICASSP-86, Tokyo, Japan, Apr. 8-11, 1986.

[3] Lamel, L.F., R.H. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proceedings of the DARPA Speech Recognition Work- shop, Palo Alto, CA, February 19-20, 1986, pp. 100-109.

[4] Peterson, G.and I. Lehiste, "Duration of Syllable Nuclei in English," J. Acoust. Soc. Am., Vol. 32, 693, 1960.

[5] Leung, H.C.and V.W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," Proc. ICASSP 84, pp. 2.7.1-2.7.4, March 1984.

[6] Leung, H.C., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Insitute of Technology, January 1985.

[7] Kassel, R.H., "Aids for the Design, Acquisition, and Use of Large Speech Databases," S.B. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute

of Technology, May 1986.

**Phonetic Symbol Mapping**

| IPA | Char | Notes | IPA | Char | Notes |
|---|---|---|---|---|---|
| **Stops** | | | | | |
| p | p | | b | b | |
| t | t | | d | d | |
| k | k | | g | g | |
| p° | ϴ | Symbol-÷ | b° | δ | Symbol-shift-D |
| t° | ∞ | Symbol-i | d° | † | Symbol-g |
| k° | ∂ | Symbol-p | g° | ≐ | Symbol-: |
| ɾ | F | | ʔ | ? | |
| **Nasals** | | | | | |
| m | m | | m̩ | M | |
| n | n | | n̩ | N | |
| ŋ | G | | ŋ̍ | π | Symbol-shift-P |
| ɾ̃ | ϵ | Symbol-shift-E | | | |
| **Fricatives** | | | | | |
| s | s | | š | S | |
| z | z | | ž | Z | |
| č | C | | ǰ | J | |
| θ | T | | ð | D | |
| f | f | | v | v | |
| **Liquids, Glides, Silence, and h** | | | | | |
| l | l | | l̩ | L | |
| r | r | | w | w | |
| y | y | | | | |
| □ | λ | Symbol-shift-L | ] | ⊂ | Symbol-t |
| h | h | | ɦ | H | |
| **Vowels** | | | | | |
| ɛ | E | | ɪ | I | |
| ɔ | c | | æ | @ | |
| ɑ | a | | ʌ | ^ | |
| u | u | | ʊ | U | |
| ɝ | R | | ü | : | |
| aʸ | Y | | ɔʸ | O | |
| eʸ | e | | iʸ | i | |
| aʷ | W | | oʷ | o | |
| ə | x | | ɚ | X | |
| ɨ | \| | | ə̥ | γ | Symbol-shift-G |

Table 1: A list of the acoustic phonetic symbols used for the transcription of the TIMIT database.

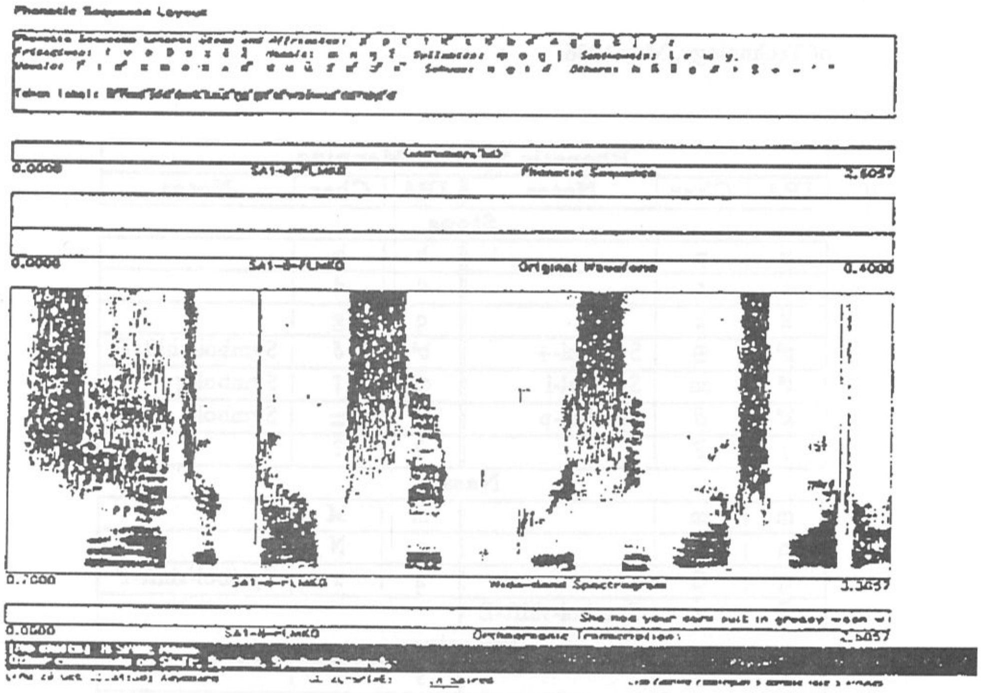


Figure 1: SPIRE layout for entering the acoustic-phonetic transcription.

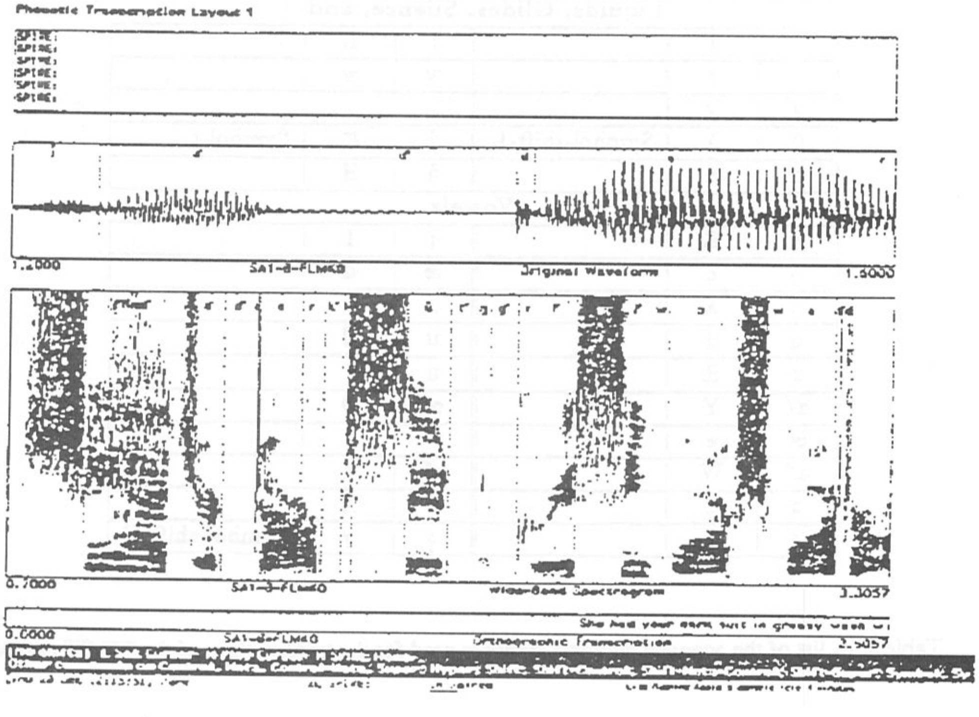


Figure 2: SPIRE layout showing the alignment produced by CASPAR.

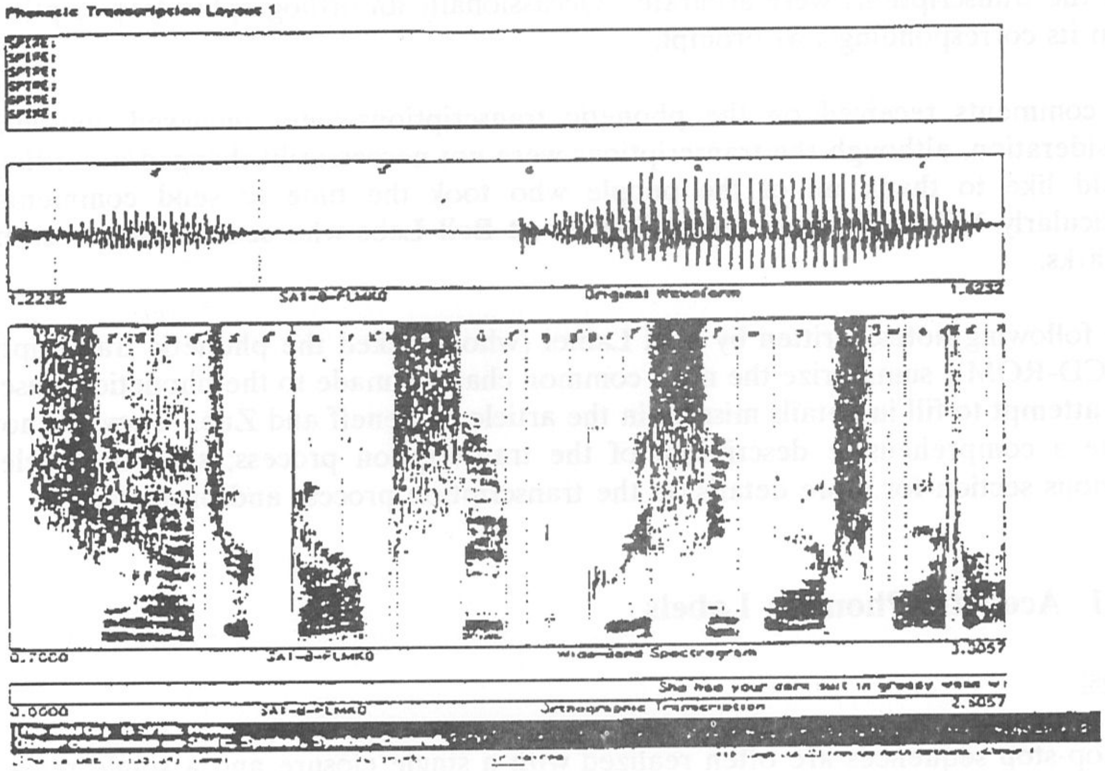


**Figure 3: SPIRE layout showing the aligned transcription following post-processing.**

## 5.2 Notes on Checking the Phonetic Transcriptions

The phonetic and orthographic transcriptions have been re-checked before this release. The aim in checking these transcriptions was to correct any blatant errors, often due to mistyping, and to make the transcriptions a bit more consistent. The phonetic transcriptions were checked at MIT using the SPIRE system (Zue et al., 1986). The phonetic transcription of any utterance is highly subjective, particularly with regard to fine phonetic distinctions, such as the exact vowel color and the partial devoicing of voiced consonants.

The orthographic transcriptions were checked primarily for spelling errors and to ensure that the transcriptions were accurate. Occassionally an orthographic transcription differs from its corresponding text prompt.

All comments received on the phonetic transcriptions were reviewed and taken into consideration, although the transcriptions were not neccessarily changed accordingly. We would like to thank all of the people who took the time to send comments to us, particularly Mike Riley and his colleagues at Bell Labs who sent us the most extensive remarks.

The following notes, written by Lori Lamel (who checked the phonetic transcriptions for the CD-ROM), summarize the most common changes made to the phonetic transcriptions and attempt to fill in details missing in the article by Seneff and Zue. They are not meant to be a comprehensive description of the transcription process; see the article in the previous section for more details on the transcription process and protocols.

### 5.2.1 Acoustic-Phonetic Labels

<u>Stops:</u>

- Stop-stop sequences are often realized with a single closure and a single release. For example, "big boy" is often realized as /bcl b ih gcl b oy/, with the /g/ being unreleased. Generally the closure interval is given to the first stop and the release to the second, unless it is clear that there was no gesture towards the first stop. For example, if there are clear labial transitions at the end of the /ih/, the sequence would be transcribed as /bcl b ih bcl b oy/.

- The glottal stop /q/ is used to denote several different types of acoustic-phonetic events, leading to some apparent confusions. A stop, typically /t/, may be realized as a glottal stop. In this case, the signal is carefully reviewed to ensure that there are no alveolar formant transitions and that one really hears a glottal stop. When a /t/ is transcribed as a glottal stop, no closure interval is marked.

/q/ is also used to mark the glottalization found at the beginning of a word starting with a vowel, or the glottal stop or glottalization that may be used to mark a vowel-vowel boundary. The /q/ is not used to mark non-event specific glottalization, such as may be found at the end of sentences, or as may be characteristic of the speech of some speakers.

- Stop closures are not marked after pauses, except in the few cases where there was a pause, followed by clear prevoicing for a voiced stop.

- Nasal-stop sequences are sometimes transcribed without a closure interval for the stop. Thus, "undo" may be found as /uh n dcl d u/ or as /uh n d u/. The latter case occurs when there is no visible weakening in the nasal murmur prior to the stop release.

- There is a relatively broad use of flaps in the transcriptions. Flaps may be as long as 40-50 ms at times, or even contain a weak, line-like release, if they are heard as a flap.

- Fricative-like allophones of voiced stops are transcribed as having only a closure interval, since there is no visible release. While it might be more realistic to transcribe fricative-like voiceless stops with only a release portion, they are typically transcribed as only a closure for consistency with the voiced stops.

Nasals:

- Sometimes, particularly when followed by a voiceless consonant as in words like "can't" and "dance," there is no segment corresponding to the nasal murmur and the only evidence of the underlying nasal is found in the nasalization of the vowel. Since there is no symbol in the set to mark the nasal in this way, a small nasal segment is marked when a nasal is heard, even if it is almost impossible to locate the nasal in the signal. In this case, the last 1-2 pitch periods of the preceeding vowel are labelled as the nasal.

Liquids and Glides:

- Post-vocalic /r/ is typically transcribed using the syllabic symbols, /er/ or /axr/, depending on the stress. This convention is used since the post-vocalic /r/ is acoustically more similiar to the syllabic than the consonantal form. This does not necessarily mean that there are two syllables present. When a post-vocalic /r/ occurs intervocalically, it is transcribed as /r/ if there are good initial-/r/ transitions into the following vowel.

Vowels:

- Since fine distinctions in vowel color are highly subjective, the vowel color was left unchanged except when the verifier strongly disagreed with the label. Thus, there may still be many differences of opinion in whether a given vowel is an /aa/, /ao/, or /ow/, and in distinguishing between the use of /iy/ and /ih/, etc.

- In general, /r/-color greatly affects the quality of the vowel. It is hard to distinguish between /ow r/ and /ao r/. The vowel preceeding /r/ in words like "ear" and "year" has been systematically transcribed as /ih/. Occassionally, a speaker pronounced an extreme /iy/ in this context, and the vowel is so marked. Under these conditions the word may be pronounced with two syllables (/y iy er/). Similarly the vowel in words like "wear" has been labelled /eh r/.

- Schwas are liberally used in the transcriptions to represent unstressed and reduced vowels. As a reminder, four types of schwas are used. The use of /ax/ or /ix/ is based on the position of the second formant: if it is closer to the third formant /ix/ is used, otherwise /ax/ is used. A devoiced schwa (/ax-h/) is marked when there is no "vocalic" portion or when there are only 1 or 2 pitch periods visible in the waveform. /axr/ marks the observation that the third formant is low, indicating retroflection. Schwa off-glides are marked only when they can be heard or seen as a syllable.

- The fronted-/uw/ (ux) is found in the transcriptions even though it is not phonemically distinct in English. Sometimes the first part of the vowel is /ux/-like and the end part is /uw/-like. In some of these cases, the vowel had been transcribed as a sequence of two vowels, /ux uw/. Since there is really only one vowel present, these transcriptions were changed. If the /ux/-like portion was longer than the /uw/-like portion, or if the second formant never really got close enough to the first formant, the symbol /ux/ was used. In other cases, the vowel was labelled /uw/, despite the fronting at the onset.

Fricatives:

The labels used for the fricatives were not discussed in the Seneff/Zue article. Some comments may help to clarify the transcriptions.

- Voiced fricatives have a tendency to be devoiced in English, with the primary cue to voicing carried in the segment duration. Thus voiced fricatives are labelled as such even though vocal fold vibration is not present throughout the segment if at least one of the following holds:

  (1) there is evidence of vocal fold vibration during part of the segment, typically found at the beginning, or

  (2) the segment duration is short relative to the voiceless fricatives in the sentence, or

(3) the duration of the preceeding vowel is lengthened.

In some cases the voicing characteristic may be very hard to determine and disagreements may arise, particularly when the fricative is part of a cluster or sentence-final.

- Fricative-fricative sequences often show modification. For example, the sequence /s sh/ is most often seen as a long /sh/. When the /s/ is visible, it is labelled. Similarly, a voiced fricative preceding a voiceless one is often devoiced. /z s/ is usually seen as a long /s/. The /z/ is marked only when voicing is evident, and the /s/ begins where the periodicity ends.

- Some speakers tend to produce stop-like allophones of the weak fricatives. These are typically transcribed as the weak fricative, recognizing this as an allophone. Sometimes, however, there is evidence of a very clear stop closure, followed by a stop-release-like fricative. In these cases a stop closure has been marked in the transcription. A /dcl/ is used before /dh/ and a /tcl/ before /th/; for the other fricatives a homorganic stop closure is used.

- There may be a small period of silence between a nasal and a fricative. Homorganic stop closures have been used to mark this interval in the place of epenthetic silence. Sometimes a stop release is also inserted. This process is known as homorganic stop insertion and is transcribed as a stop.

### 5.2.2 Boundaries

Few adjustments were made to segment boundary locations. Severe misalignments (a relatively rare happening) were corrected. The most common boundary change was the adjustment of the start location of a stop release, which occassionally cut off the beginning of the release.

### 5.2.3 Disclaimer

Phonetic transcriptions are inherently extremely subjective; thus, we expect that there will always be disagreements with some of the decisions made in transcribing and checking TIMIT. Our goal was to provide a relatively broad acoustic-phonetic transcription where the most reliable acoustic landmarks have been marked. The re-checking of TIMIT, aimed at correcting relatively blatant errors and not at making finer distinctions, represents about 200 hours of human-interactive time, and as such is a task subject to error. We hope to have minimized the transcription errors in TIMIT and to have made the transcriptions more consistent.

## 5.3 Notes on Automatic Generation of Word Boundaries

This section describes the program used to automatically associate word boundaries with phonetic segments. The program is similar in concept to the work of Kassel (1986).

### 5.3.1 General Methodology

The automatic generation of word boundaries is accomplished using the following algorithm:

> (1) A phonemic transcription of a sentence is generated from an orthography by concatenating the phonemic form of the lexical entry for each word.
>
> (2) The resulting string is then aligned with the phonetic transcription using a dynamic programming-based string alignment program (available from NIST) with weights based on phonetic features.
>
> (3) After alignment, the word boundaries in the phonetic string are inferred from the phonemic string by applying a set of phonological rules.

Automatically-generated word boundaries using the above algorithm agreed with 96% of the available human-checked boundaries on a sample of 4000 sentences.

### 5.3.2 Alignment Procedure

The alignment of phonemes and phones is performed using a dynamic programming string alignment algorithm to determine a mapping from phonemes to phones which minimizes a distance function. The distance function which was used, technically a weighted Levenshtein metric, is a weighted sum of all insertion, deletion, and substitution operations necessary to edit the phoneme string into the phone string. The weight of each elementary operation is the sum of the phonetic features that are different between the mapped phoneme and phone. By convention, deleted phonemes and inserted phones are mapped to "null", a symbol defined as having no phonetic features, so that their contribution to the distance is the number of phonetic features defining them. The alignment code using this concept of phonological distance was reported on at ICASSP90 (Pallett et al., 1990) and is available from NIST.

### 5.3.3 Phonological Rule Post-Processing

The following rules are applied to the aligned phonemic and phonetic strings. If the rule's preconditions are met, the rule is activated (or "fired") and the word boundary is modified. This modification can add, delete, orphan, or share phones at the boundary. The rules are listed below in the order of precedence in which they are applied. Only one rule is active at a time. The rule format is:

[precondition] : [phone sequence] -> [phoneme sequence]

where "->" means "is mapped to".

Disclaimer: While this rule set is likely to be incomplete, we feel it provides adequate agreement with human-checked boundaries.

Rule 1: Orphanization of silence periods

The phones in the set {h#, pau, epi} were orphaned unless the alignment routine matched them with a word final phoneme.

[any context] : (pau) -> ()

Rule 2: Orphanization of glottal stop insertions

If a glottal stop was inserted between a word final vowel and the following word initial vowel, it was left as an orphan phone.

[(vow1 q vow2)] : (q) -> ()

Rule 3: Stop Closure and Release Merges

The phonetic transcription and phonemic representations differ with respect to the representation of stops. In the phonemic transcription the stop is a single token, while in the phonetic transcription stop closures are marked separately from stop releases. This rule searches for a "closure release", e.g. tcl-t, bcl-b, dcl-d, that spans the inferred word boundary. If this condition occurs, the boundary is shifted to include both phones in the proper word.

[any context] : (tcl t) -> (t)

### Rule 4: Sharing of geminate phones

MIT's phonetic transcription convention for geminate phones, i.e. where the word final and word initial phones were identical, was to mark them as a single segment. This rule adjusts the word boundaries to allow this single phone to be shared.

[geminate phoneme (m m)] : (m) -> (m m)

### Rule 5: Sharing word final and initial vowels

Word final vowels followed by word initial vowels were occasionally transcribed as a single vowel segment. Typically at least one of the 2 vowels was unstressed. This rule searches for a missing vowel at the inferred word boundary, then forces the remaining vowel to be shared.

[phoneme (vow1 vow2)] : (vow) -> (vow1 vow2)

### Rule 6: y-Palatization sharing

A fairly common phonological transformation is y-palatalization of stops and fricatives across word boundaries. In this case the underlying phonemic sequence of (d y) may be manifest phonetically as (dcl jh).

The following set of rules account for these phonomena.

```
a.  [phoneme (d y)] : (dcl jh) -> (d)
                      (jh) -> (y)

b.  [phoneme (t y)] : (tcl ch) -> (t)
                      (ch) -> (y)

c.  [phoneme (s y)] : (sh) -> (s)
                      (sh) -> (y)

d.  [phoneme (z y)] : (zh) -> (z)
                      (zh) -> (y)
```

In cases (a) and (b), sometimes the closure interval is missing, and the d, t are aligned only with jh, ch.

# 6 Reprints of Selected Articles

This section includes reprints of the following three TIMIT-related articles that appeared in the proceedings of DARPA Speech Recognition Workshops.

Fisher, William M., Doddington, George R., and Goudie-Marshall, Kathleen M. (1986), "The DARPA Speech Recognition Research Database: Specifications and Status," *Proceedings of the DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, February 1986, Palo Alto.

Lamel, Lori F., Kassel, Robert H., and Seneff, Stephanie (1986), "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proceedings of the DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, February 1986, Palo Alto.

Cohen, Michael, Baldwin, Gay, Bernstein, Jared, Murveit, Hy, and Weintraub, Mitchel (1987), "Studies for an Adaptive Recognition Lexicon," *Proceedings of the DARPA Speech Recognition Workshop*, Report No. SAIC-87/1644, March 1987, San Diego.

# THE DARPA SPEECH RECOGNITION RESEARCH DATABASE: SPECIFICATIONS AND STATUS

William M. Fisher
George R. Doddington
Kathleen M. Goudie-Marshall

Texas Instruments Inc.
Computer Sciences Center
P.O. Box 226015, MS 238
Dallas, Texas 75266, USA
Tel. (214) 995-0394

## ABSTRACT

This paper describes general specifications and current status of the speech databases that Texas Instruments (TI) is collecting to support the Darpa speech recognition research effort. Emphasis is placed on the portion of the database development work that TI is specially responsible for. We give specifications in general, our recording procedures, theoretical and practical aspects of sentence selection, selected characteristics of selected sentences, and our progress in recording.

## 1. INTRODUCTION

This paper is a report on the specification and current status of the work done by Texas Instruments, Inc. (TI) on Darpa-funded Acoustic Phonetic Database development as of the early part of February, 1986. It is meant to be complementary to similar reports from other groups included in this volume.

## 2. GENERAL SPECIFICATIONS

Originally three data bases were planned: "stress," "acoustic-phonetic," and "task-specific." The stress data base was to investigate variations of speech with stress, and would be done primarily by AFAMRL. The acoustic-phonetic database, to be done by TI in collaboration with MIT and SRI, was intended to uncover general acoustic-phonetic facts about all major dialects of continental U.S. English. And the task-specific data base, providing data for the study of the effect on speech recognition of limiting domain of discourse, would be defined later. At our last meeting, there was a consensus that the task-specific data base should be folded into the acoustic-phonetic data base, becoming one of the later phases.

The acoustic-phonetic data base is phased so that a small amount of speech is initially recorded from a large number of subjects, followed by successively larger durations of speech from fewer subjects, culminating in two hours recorded from each of two subjects. MIT and SRI have helped us design the material to be read by subjects. Figure 1 below shows the current general specifications for this data base.

## 3. RECORDING PROCEDURES

### 3.1 STEROIDS

This large scale database collection would be difficult or impossible to collect without the VAX Fortran automated speech data collection system developed here at TI, called the STEReO automatic Interactive Data collection System, or STEROIDS. Use of STEROIDS requires a stereo DSC 200 sound system directly connected to 2 DSC 240 audio control boxes, one for each of the 2 channels of stereo input. (No multiplexor is used.)

STEROIDS uses a file called the Vocabulary Master Library (VML). The VML file is a formatted direct access file which contains records holding data for each utterance in a recording session: the text of the prompt, a speech file name, and variables holding the number of recorded versions and which one is best. A prompt may be any text string less than 133 characters long.

When STEROIDS is executed, it first reads in values for several parameters that effect its decisions about when each utterance begins and ends, and a name for the subject. It then, under the control of the director, displays prompts to the subject and records his responses in speech files. The director may listen to recorded versions, decide which version is best, and re-prompt.

Recording conditions:
- o Low noise (acceptable to NBS)
- o 2 channel recording: 1 noise-cancelling (Sennheiser) mike, 1 far-field pressure (Bruel and Kjaer) mike.
- o Subjects exposed to 75 dB SPL noise through earphones

Style:
- o Read from prompts

Material:

| Phase | Speech/Subject | # Subjects | Contents, etc. |
|---|---|---|---|
| 1 | 30 sec. | 630 | Broad Phonetic Coverage |
| 2 | 2 min. | 160 | |
| 3 | 8 min. | 40 | W/Standard Paragraph |
| 4 | 30 min. | 10 | W/Explicit Variations |
| 5 | 2 hrs. | 2 | Interview Format |

Figure 1. General Specifications of Acoustic-Phonetic Database.

### 3.2 GENERAL PROCEDURE

We created and ran a program which read sentences and sentence assignments and made 630 VML files. Our recording procedure then takes five steps: 1. At the beginning of each day, calibration tones are recorded from both channels; 2. For each subject, one of the 630 VML files is copied to his named sub-directory and STEROIDS is used to collect his data; 3. At the end of the day, a REDUCE procedure is run on all data collected that day, which produces the files that we send out, by splitting the initial stereo file into two mono files, de-biasing each, high-pass filtering the BK file at 70 Hz., and down-sampling each to 16,000 samples per second; 4. A backup procedure is then run, which makes three tape copies of the VML files, the calibration tone files, and all the speech files recorded on that day; and 5. The disk is cleaned up for re-use by deleting the files that were put onto tape. One copy of the back-up tape is then sent to NBS.

Data on each subject recorded in each session is added to an ASCII text file for documentation.

### 3.3 NOISE

After the sound booth was moved to the third floor of the North Building, a very large noise signal was observed coming from the combination BK power supply and preamplifier. At first this noise was thought to be the result of a defect in the amplifier, but the BK service center could find no problem. It was then that we realized that the noise was actually an acoustical signal being picked up by the microphone. Figure 2 shows the spectrum of the noise signal below 500 Hz for a 5 second segment of "silence". The spectrum is flat from 300 Hz up to 10 kHz. (The spectrum of the signal from the Sennheiser noise-cancelling microphone is flat from DC to 10 kHz, which indicates that the noise-cancellation property and the low-frequency roll-off of the Sennheiser is adequate to render the acoustic rumble of no consequence for this microphone.)

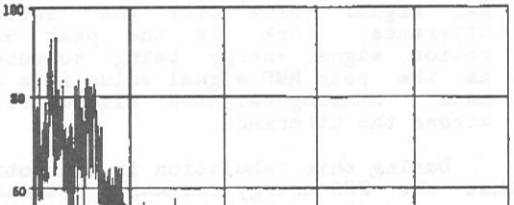


Figure 2. Amplitude Spectrum of Acoustic Rumble. Recorded in the TI sound booth over a 5 second period of "silence."

With consultation from an acoustical engineering consultant it was judged that the acoustical noise in our double-walled sound booth is being introduced by mechanical vibrations transmitted through the floor. Opinion varies as to the amount of reduction that may be achieved by better isolation from the floor, from less than 3 dB to more than 20 dB. Current plans are to install an air suspension vibration isolation mount system under the sound booth to reduce the rumble as much as possible.

As an interim solution, a 1581-point FIR filter has been designed to provide a high-pass filter function, with a cut-off at 70 Hz and an in-band ripple of less than 0.1 dB above 100 Hz. Using this filter, reasonably acceptable S/N ratios have been achieved during data collection. The following S/N ratios have been measured, using seventeen subjects' (nine men and eight women) utterances of sentence SA1.

| Condition | ENrms | SNavg | SNpk |
|---|---|---|---|
| No HP | 421 | 8 dB | 16 dB |
| 70 Hz HP | 95 | 21 dB | 29 dB |
| 200 Hz HP | 4 | 48 dB | 56 dB |

Table 1. Raw S/N Ratios.

Explanatory notes for Table 1: ENrms is the RMS energy of the noise; SNavg is the average S/N ratio, signal energy being computed as the average RMS signal value over the entire utterance; SNpk is the peak S/N ration, signal energy being computed as the peak RMS signal value in a 30 msec. Hamming-weighted window slid across the utterance.

During this tabulation it was noticed that the RMS energy for men's utterances averaged 4 dB greater than that for women. There are variations of signal level with speaker and with utterance, of course, and the weakest of the seventeen utterances used for this tabulation showed an average S/N ratio of minus 1 dB for the original signal.

The effective S/N ratios for speech processing and listening or perceptual purposes will be somewhat higher for the no high-pass and 70 Hz high-pass results listed above, because typically a preemphasis is performed on the speech signal before further processing, and because the human ear is progressively less sensitive to sound at frequencies below 200 Hz. For a preemphasis constant of 1.0 (at a sampling frequency of 16 kHz), the S/N ratios were measured as follows:

| Condition | ENrms | SNavg | SNpk |
|---|---|---|---|
| No HP | 6 | 35 dB | 46 dB |
| 70 Hz HP | 4 | 39 dB | 50 dB |
| 200 Hz HP | 3 | 41 dB | 52 dB |

Table 2. Pre-emphasized S/N Ratios. Symbols are same as in Table 1.

## 4. ACOUSTIC-PHONETIC DATA BASE PHASE 1

### 4.1 GENERAL

The sentences constituting the phase 1 material will have a mean value of expected reading time of 3 seconds, so that each of the 630 subjects reading ten sentences will give us the specified 30 seconds per subject of speech data.

Altogether 630x10=6300 sentence tokens will be collected. The sentence types are divided into three sorts: 1. Two "dialect" or "calibration" sentences; 2. 450 "MIT" sentences; and 3. 1890 "TI" sentences. Each subject reads both the dialect sentences, a selection of five of the MIT sentences, and a selection of three TI sentences. Each MIT sentence will be read by seven speakers and each TI sentence by one. This variation in the number of subjects reading different sentences is a compromise between the desiderata of breadth and depth of phonetic coverage across subjects.

The dialect sentences were devised by SRI and the MIT sentences by MIT, who will report separately on their design.

### 4.2 THE TI NATURAL PHONETIC SENTENCES

Our strategy in selecting our 1890 sentences was almost identical to one we reported on earlier [1]: use a computer procedure to select from a large or infinite set of sentences a subset that meets certain feasibility criteria, trying to optimize an objective function of the selected sentences. The ideal set of sentences to draw from in this case is the set of normal, acceptable American English sentences. Lacking an off-the-shelf grammar of sufficient generality, we approximate this set with the largest set of American English sentences in computer readable form that we know of, the "Brown Corpus [2]." Responding to concerns of some in the DARPA Database SIG that these sentences were "written" English instead of "spoken" English, we augmented our final pool of sentences from this corpus with 136 sentences of playwrights' dialog from the corpus published by Hultzen et al. [3]. (We are not concerned that our sentences are too "written": the alternative, naturally "spoken" sentences, are replete with run-on sentences, self-corrections, and ungrammaticality.)

There may be some slight discrepency between the original written form of these Hultzen sentences and the form in which we use them, since we reconstructed their spellings from the phonemic transcriptions published in the Hultzen book using TI off-the-shelf speech-to-text technology.

A series of programs was executed that produced a file of pointers to the beginnings of sentences in the Brown corpus, then filtered out sentences from this set until about 10,000 were left in the selection pool. Sentences were eliminated if they were over 80 characters long, included any proscribed words, or included characters other than letters and punctuation. This pool was augmented with 136 Hultzen sentences.

The fixed set of sentences -- the two dialect sentences and the revised set of 450 sentences that TI received from MIT in the middle of November -- were transcribed phonemically by TI's best off-the-shelf text-to-phoneme program and, after careful checking by two experts in phonetics and phonology, files of allophonic transcriptions of them were computed as described below. The selection program assumed this set of utterances as a base to build on in the selection of the 1890 TI sentences.

The selection pool of 10,000 sentences was prepared in a similar way, except that it was not feasible to hand-check the transcriptions.

The selection program accesses these allophonic transcription files, in addition to a file of pointers to sentences that have previously been selected and one of pointers to sentences that have been manually zapped (ruled out). It produces a new version of the sentence selection file. Both the sentence selection file and the zapped sentence file are in ASCII text file format so that they can be manipulated with a text editor. One of the program's typed-in parameters tells it how many sentences to select. The program was run in a series of batch jobs, each typically selecting an additional 100 or so sentences. The additional sentences selected in each batch run were examined, and unacceptable ones were stricken from the selected sentence file and added to the zapped sentence file before the next run.

The internal procedure used by the program is this:

1. Build the initial version of the data structure holding phonetic data on the selected sentences by reading in the dialect sentences weighted by 630, the MIT sentences weighted by 7, and the previously selected TI sentences weighted by 1;
2. Repeat this until this run's quota of sentences has been selected:
    a. Scan through a list of prospective sentences from the pool of unselected and unzapped sentences, calculating for each the increase in the phonetic objective function under the hypothesis that the sentence is added to the selected set, remembering the one producing the highest value;
    b. Add the remembered sentence to the selected sentence list.
3. Write out the new version of sentence selections.

The program knows two basic ways of making a list of sentences from the pool for examination: 1. take N (typically 400) random grabs; and 2. look at them all. This option is selectable by the user, and both were used in actual runs selecting sentences. The first is faster and less optimal than the second.

### 4.3 CONTROL OF AVERAGE UTTERANCE DURATION

In order to control the average duration of utterances, a heuristic was used. The expected speech duration of each sentence was calculated using the formula

$$SPDUR = -0.0928 + .06302 * NLETTS$$

where NLETTS is the number of letters in the spelling of the sentence and SPDUR is the speech duration of the sentence in units of seconds. This formula was derived by the least-squared-error fit of a linear function to speech duration data obtained from a previously collected data base of continuous speech: 750 sentences from each of eight subjects, half male and half female. The mean value of speech utterance duration of the current selected sentence set was kept track of, and if it was lower than the target duration (three seconds) minus a tolerance, the next sentence selection was taken from a list of longer-than-average pool sentences; if the mean speech duration was greater than the target plus a tolerance, the next selection was from the subset of short pool sentences; and if within the tolerances, any of the 10,000 pool sentences could be selected. The tolerance used in the final selection was 1%.

### 4.4 OBJECTIVE FUNCTION

The function that is used to measure the aggregate phonetic coverage of the set of selected utterances, called "allophone information", is:

$$Ial= SUM(Ni*LOG2(Ni/Ntot))$$

where Ni is the frequency of phonetic unit i and Ntot is the total number of phonetic units in the utterance set. A user-specified switch determines whether the function is used in its absolute form as given above, or normalized by dividing by the number of letters in the sentences. Most of the later runs were made using the relative form of the function.

Following most authorities on phonetics, we take the relevant set of phonetic units to be phones, allophones or variants of phonemes of American English [4,5], roughly equivalent to Pike's "speech sounds" [6, pp. 42]. The problem of calculating or defining the complete set of allophones is equivalent to defining the set of possible phonological rules. The first-order approximation to this that was used is: an allophone is a variant of a phoneme that is distinguished by the phone on its immediate left, the phone on its immediate right, and, if it is syllabic, by a binary mark of stressed or non-stressed; part of the allophonic representation, also, is whether there are word boundaries on its immediate right or left before the adjacent segments. For the purposes of this specification, left and right environmental phones are the segmental phonemes with vowels marked as stressed or nonstressed and the complex phonemes /ch/, /jh/ written as [t sh] and [d zh]. (This is a correction and generalization of a proposal for psycholinguistic units of speech recognition made by Wickelgren some years ago [7, chap. 6,7].)

It is important to use phones instead of phonemes as possible phonetic conditioning environments for several reasons.

Complex phones condition phonetically according to their separate parts. If you think, as we do, that the vowels of "chew" and "shoe" are phonetically identical, then always counting phones as different if they have different adjacent phonemes won't work: the two vowels have different phonemes on their left -- /ch/ vs. /sh/ -- but the identical phone, [sh]. And /oy/ and /aw/ probably cause rounding assimilation on different ends, /oy/ at the beginning and /aw/ at the end, although there is no principled way to distinguish them with the phonological feature of rounding if they are regarded as holistic segments.

In general, conditioning phones should also be marked redundantly for features that can assimilate over an intervening segment. Only if the /t/ of "stew" is marked for lip rounding will the /s/ be in an environment that will cause it to become rounded, but lip rounding is not phonemic in English consonants. If you think, as we do, that the /s/'s of "stew" and "sty" are phonetically different, then the relevant conditioning environment cannot be just the immediately following phoneme.

Of course, supra-segmental features affect phonetics also. As a first approximation to this, we mark vowels as being stressed or non-stressed and include word and utterance boundaries in conditioning environments. Something like this must be done if you think, as we do, that the /t/'s of "deter" and "veto" are phonetically different, and that the /ay/'s of "Nye trait" and "night rate" are also different.

Because of exigencies of time and resources available, the allophonic codes actually used were 4-byte integers consisting of these bit patterns:

EACH ALLOPHONE CODE:

- o 6 bits for segmental phone code
- o 6 bits for segmental phone on left
- o 6 bits for segmental phone on right
- o 1 bit for word boundary on left
- o 1 bit for word boundary on right

where segmental phone codes are classical phonemes except:

- o Vowels marked stressed/unstressed
- o Complex phonemes are split: /CH/=[T SH]; /JH/=[D ZH]
- o Utterance begin/end mark used: /$/

Figure 3. Allophone Codes.

The simplest way to decode and write one of these allophone codes is as a phone with an environment specified as in linguistic phonological rules. Here is an example from the log of a computer program run testing allophone coding and decoding that shows how this method of counting phonetics handles three well-known phrases that are distinguished by allophones of /T/:

```
ORTH="Nye trait/nitrate/night-rate"
PRON=/- N AY1 - T R EY1 T - - N AY1
 T R EY1 T - - N AY1 T - R EY1 T -/

THESE ARE THE PHONETIC UNITS:

I     ALLO(I)         DECODED:
1     868422        N / [$] __ [AY1]
2     292253        AY1 / [N] __ [# T]
3     643562   *    T / [AY1 #] __ [R]
4     960268        R / [T] __ [EY1]
5      64156        EY1 / [R] __ [T]
6     639957        T / [EY1] __ [# N]
7     878406        N / [T #] __ [AY1]
8     292252        AY1 / [N] __ [T]
```

(continued on next page)

(continued from previous page)

| | | | |
|---|---|---|---|
| 9 | 643560 | * | T / [AY1] ___ [R] |
| 10 | 960268 | | R / [T] ___ [EY1] |
| 11 | 64156 | | EY1 / [R] ___ [T] |
| 12 | 639957 | | T / [EY1] ___ [# N] |
| 13 | 878406 | | N / [T #] ___ [AY1] |
| 14 | 292252 | | AY1 / [N] ___ [T] |
| 15 | 643561 | * | T / [AY1] ___ [# R] |
| 16 | 960270 | | R / [T #] ___ [EY1] |
| 17 | 64156 | | EY1 / [R] ___ [T] |
| 18 | 639745 | | T / [EY1] ___ [$] |

*: 3 DIFFERENT ALLOPHONES OF /T/.

Figure 4. Allophone Encoding/Decoding

### 4.5 RESULTING SENTENCES

The sentences resulting from this selection process were checked for accepability by two experts with Ph.D.'s in Linguistics with major areas of Phonetics and Phonology (WMF and KGM), and one Registered Speech Pathologist (Jane McDaniel), who has been hired as a consultant to help record the data base. The only area in which there was some disagreement was on whether or not to allow utterances consisting of just a well-formed noun phrase. The decision was made to allow such fragments if they were not otherwise unacceptable, because they are perfectly common and normal in speech, according to such authorities as Sledd [8,p. 169]:

> "We often say things, in perfectly normal speech, which do not contain a complete subject and a complete predicate. We might very well say, possibly in answer to a question,
>
> the choir ↓
>
> just as we might say,
>
> the choir → will sing now ↓
>
> Both answers are correct English utterances, and both end in the terminal ."

Figure 5 below shows some characteristics of the sentences finally selected:

| SENTENCES | #Utts. | Ial | # Allophone Types | # Allophone Tokens |
|---|---|---|---|---|
| DIA+MIT | 4410 | 1584 | 7,296 | ~147k |
| DIA+MIT+TI | 6300 | 2562 | 19,853 | ~212k |

Figure 5. Phase I Utterances Summary. Allophone information, Ial, is in kbits.

### 4.6 SENTENCE-TO-SUBJECT ASSIGNMENT

Sentences were assigned to subjects represented as indices; as particular real subjects are chosen they are assigned a subject index arbitrarily. The initial assignment of sentences to subject indices was made by a looping program that assigned consecutive sentences from the selected sentence set to different subjects.

Another program then re-assigned sentences to subjects in order to reduce the range of expected total speech durations assigned to subjects. The program used a simple fast repetitive heuristic of finding the subjects with the longest and the shortest assigned speech durations, then interchanging the two sentences between them that make the greatest reduction in the difference of their speech durations, respecting the constraints of the experimental design (dialect sentences interchange only with dialect sentences, MIT only with MIT sentences, and TI only with TI sentences). Before this program ran, the minimum, average, and maximum speech durations assigned to subjects were 23, 30, and 37 seconds respectively; running the program increased the minimum to 28.5 and reduced the maximum to 32.

### 4.7 RECORDING PROGRESS

Last fall TI made a commitment to send a sample of at least ten speakers' recordings to NBS for evaluation by December 21, which was done. In addition, we made a commitment to record and send out an average of 20 speakers per week beginning January 1. Figure 6 below shows how well we have met that commitment.

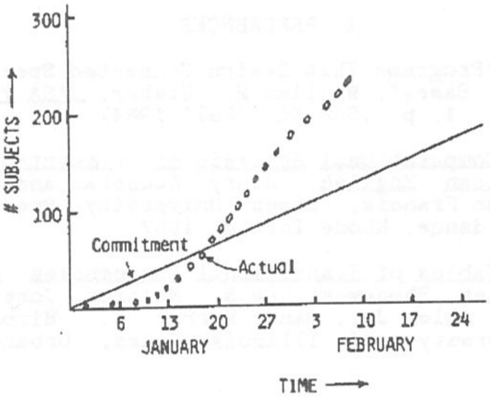


Figure 6. Recording Progress. Phase I.

We expect to finish Phase 1 and start Phase 2 recording during April 1986.

Figure 7 below shows the geographical distribution of speakers we have recorded as of February 16, 1986, along with the definition of the assumed dialectal areas we are using in an attempt to get more even numbers of speakers from all major dialects. Table 3 below gives a numerical summary of this same information.

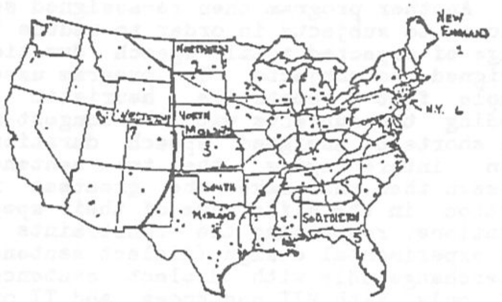


Figure 7. Current Geographical/Dialect Distribution of Speakers

| AREA# | AREA NAME | NMALES | NFEMALES | TOTAL |
|---|---|---|---|---|
| 1 | New England | 6 (60%) | 4 (40%) | 10 ( 4%) |
| 2 | Northern | 25 (64%) | 14 (36%) | 39 (17%) |
| 3 | North Midland | 29 (76%) | 9 (24%) | 38 (17%) |
| 4 | South Midland | 34 (67%) | 17 (33%) | 51 (23%) |
| 5 | Southern | 25 (63%) | 15 (38%) | 40 (18%) |
| 6 | New York City | 4 (57%) | 3 (43%) | 7 ( 3%) |
| 7 | Western | 23 (77%) | 7 (23%) | 30 (13%) |
| 8 | Army Brat | 5 (56%) | 4 (44%) | 9 ( 4%) |
| | | ---- | ---- | ---- |
| | TOTAL: | 151<br>(67%) | 73<br>(33%) | 224 |

Table 3. Breakdown of Subjects (2/18/86)

## 4. REFERENCES

[1] "Programs That Design Connected Speech Data Bases", William M. Fisher, JASA 74, supp. 1, p. S48 (A) (Fall 1984)

[2] Computational Analysis of Present-Day American English, Henry Kuchera and W. Nelson Francis, Brown University Press, Providence, Rhode Island, 1967.

[3] Tables of Transitional Frequencies of English Phonemes, Lee S. Hultzen, Josech H.D. Allen Jr., and Murray S. Miron, University of Illinois Press, Urbana, 1964.

[4] A Course in Phonetics, Peter Ladefoged, Harcourt Brace Jovanovich, Inc., New York, 1975.

[5] General Phonetics, R-M. S. Heffner, The University of Wisconsin Press, Madison, 1964.

[6] Phonetics, Kenneth L. Pike, The University of Michigan Press, Ann Arbor, 1966.

[7] Speech and Cortical Functioning, ed. John H. Gilbert, Academic Press, New York, 1972.

[8] A Short Introduction to English Grammar, James Sledd, Scott, Foresman and Company, Chicago, 1959.

# SPEECH DATABASE DEVELOPMENT: DESIGN AND ANALYSIS OF THE ACOUSTIC-PHONETIC CORPUS*

Lori F. Lamel, Robert H. Kassel, and Stephanie Seneff

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

## ABSTRACT

The need for a comprehensive, standardized speech database is threefold: first, to acquire acoustic-phonetic knowledge for phonetic recognition; second, to provide speech for training recognizers; and third, to provide a common test base for the evaluation of recognizers. There are many factors to consider in corpus design, making it impossible to provide a complete database for all potential users. It is possible, however, to provide an acceptable database that can be extended to meet future needs. After much discussion among several sites, a consensus was reached that the initial acoustic-phonetic corpus should consist of calibration sentences, a set of phonetically compact sentences, and a large number of randomly selected sentences to provide contextual variation. The database design has been a joint effort including MIT, SRI, and TI. This paper describes MIT's role in corpus development and analyzes of the phonetic coverage of the complete database. We also include a description of the phonetic transcription and alignment procedure.

## INTRODUCTION

The development of a common speech database is of primary importance for continuous speech recognition efforts. Such a database is needed in order to acquire acoustic-phonetic knowledge, develop acoustic-phonetic classification algorithms, and train and evaluate speech recognizers. The acoustic realization of phonetic segments results from a multitude of factors, including the canonical characteristics of the phoneme, contextual dependencies, and syntactic and extralinguistic factors. A large database will make it possible to examine in detail many of these factors, with the hope of eventually understanding acoustic variability well enough to design robust speech recognizers. A complete database should include different styles of speech, such as isolated words, sentences and paragraphs read aloud, and conversational speech. The speech samples should be gathered from many speakers (at least several hundred) of varying ages, both male and female, with a good representation of the major regional dialects of American English.

## DESIGN CONSIDERATIONS

There are many factors to consider in designing a large corpus for speech analysis. Unfortunately, some of the goals are limited by practical considerations. Ideally we would like to include multiple samples of all phonemes in all contexts, a goal that is clearly impractical for a manageable database.

At the last DARPA review meeting it was decided that an initial acoustic-phonetic database would be designed to have good phonetic coverage of American English. It was agreed that the initial acoustic-phonetic corpus would include calibration sentences (spoken by every talker), a small set of phonetically compact sentences (each spoken by several talkers) and a large number of sentences (each to be spoken by a single talker). This combination was chosen to balance the conflicting desires for compact phonetic coverage, contextual diversity, and speaker variability. Another requirement of the corpus was that the sentences should be reasonably short and easy to say.

The database design is a joint effort between MIT, SRI, and TI. The speaker *calibration sentences*, provided by SRI, were designed to incorporate phonemes in contexts where significant dialectical differences are anticipated. They will be spoken by all talkers. The second set of sentences, the *phonetically compact* sentences, was hand-designed by MIT with emphasis on as complete a coverage of phonetic pairs as is practical. Each of these sentences will be spoken by several talkers, in order to provide a feeling for speaker variation. Since it is extremely time-consuming and difficult to create sentences that are both phonetically compact and complete, a third set of *randomly selected* sentences, chosen by TI, provides alternate contexts and multiple occurrences of the same phonetic sequence in different word sequences.

A breakdown of the actual sentence corpus is shown in Table 1. This arrangement was chosen to balance the conflicting desires for capturing inter-speaker variability and providing contextual diversity. Since the calibration

*This research was supported by DARPA under contract N00039-85-C-0341, monitored through Naval Electronic Systems Command.

|  | No. Talkers | No. Sentences | Total |
|---|---|---|---|
| Calibration (SRI) | 640 | 2 | 1280 |
| Compact (MIT) | 7 | 450 | 3150 |
| Random (TI) | 1 | 1890 | 1890 |
| Total | — | — | 6320 |

**Table 1:** Breakdown of Frequencies of Occurrence of Sentences in Corpus

sentences are spoken by all of the speakers, they should be useful for defining dialectical differences. For multiple instances of the exact same phonetic environments, but with a much richer acoustic-phonetic content than in the calibration sentences, the MIT set would be appropriate. The TI sentences, to be spoken by one talker per sentence, should provide data for phoneme sequences not covered by the MIT database.

## DESIGN OF THE COMPACT ACOUSTIC-PHONETIC SENTENCES

A set of 450 sentences was hand-designed at MIT, using an iterative procedure, to be both compact and comprehensive. We made no attempt to phonetically balance the sentences. We used *ALexis* and the Merriam-Webster Pocket Dictionary (Pocket) to interactively create sentences and analyze the resulting corpus. We began with the "summer" corpus created for the MIT speech spectrogram reading course to include basic phonetic coverage and interesting phonetic environments. We initially augmented these sentences by looking at pairs of phonemes, trying to have at least one occurrence of each phoneme pair sequence. *ALexis* was used to search the Pocket dictionary for words having sequences that were not represented and for words beginning or ending with a specific phoneme. We then created sentences using the new words and added them to the corpus. Certain difficult sequences were emphasized, such as vowel-vowel and stop-stop sequences. Some phoneme pairs are impossible; others are extremely rare and occur only across word boundaries. For example, /w/ and /y/ never close a syllable, except as an off-glide to a vowel, so many /w/-phoneme pairs are impossible. After filling some of the gaps in coverage, we reanalyzed the sentences with regard to phoneme pair coverage, consonant sequence coverage, and the potential for applying phonological rules both within words and across word boundaries. In a final pass through the sentence set, we modified and enriched sentences where simple substitutions could introduce variety or generate an instance of a rare phoneme pair.

## ANALYSIS OF PHONETIC COVERAGE

This section discusses the phonetic coverage of the compact sentence set developed at MIT and the resulting corpus consisting of the combined MIT and TI sentences. This analysis does not include the calibration sentences as we consider their use to be of a different nature.

|  | POCKET | HL | MIT-450 | APDB |
|---|---|---|---|---|
| # sentences |  | 720 | 450 | 5040 |
| # unique words | 19,837 | 1894 | 1792 | 5107 |
| # words | 19,837 | 5745 | 3403 | 41,161 |
| ave # words/sent |  | 7.9 | 7.6 | 8.2 |
| min # words/sent |  | 5 | 4 | 2 |
| max # words/sent |  | 12 | 13 | 19 |
| ave # syls/word | 1.38* | 1.1 | 1.58 | 1.54 |
| ave # phones/word | 3.34* | 2.97 | 4.0 | 3.89 |

* The ave # syls/word and ave # phones/word have been weighted by Brown Corpus[1] word frequencies.

**Table 2: Description of Databases**

Table 2 compares some of the distributional properties of the Pocket Lexicon (Pocket), the Harvard List (HL)[2], the MIT-selected sentences (MIT-450), and the Acoustic-Phonetic Database selected sentences (APDB). The APDB includes seven copies of each MIT-450 sentence, to account for the number of talkers per sentence, and a single copy of each randomly selected sentence (TI-1890). Since we were given only the orthographies for the TI-1890 sentences, we generated phonemic transcriptions by dictionary lookup, by rule-based expansion of the dictionary entries, and, as a last resort, by a text-to-speech synthesizer. We expect that there are pronunciation variations between the dictionary and the text-to-speech synthesizer, particularly with respect to vowel color. There may also be some pronunciation errors, but we think these will be statistically insignificant.

The proportion of unique words relative to the total number of words is substantially larger in the MIT-450 than the APDB, probably due to the selection procedure. We tried to use new words in sentences and to avoid duplication when at all possible. Roughly 50% of the MIT-450 words are unique, as compared to only 25% of the APDB words. The TI-1890 sentences are, on the average, slightly longer than those in the MIT-450. The 10 most frequently occurring words for all of the corpora are function words or pronouns. In both the MIT-450 and the APDB corpora, the most common word is "the," accounting for roughly 7%of all words.

The average numbers of syllables and phones per word are longer for the MIT-450 and the APDB than for the HL. This is presumably due to the higher percentage of polysyllabic words.

Figure 1 shows the distribution of the number of syllables per word for the two corpora. The distributions are quite similar, with the majority of the words being mono- or bi-syllabic. The MIT-450 corpus has a slightly higher percentage of polysyllabic words than does the combined

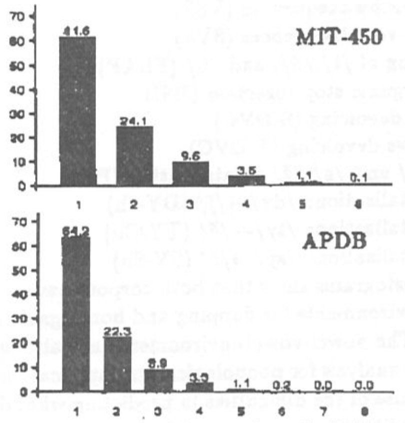


**Figure 1:** Histograms of the number of syllables per word.

corpus. We specifically tried to include polysyllabic words in the sentences, since these are likely to be spoken with greater variability.

Distributions of the number of phonemes per word are shown in Figure 2. The 10 most common phonemes and their frequency of occurrence are given in Figure 3.

Table 3 shows the distribution of within-word consonant sequences for the four databases. The MIT-450 sentence set covers most of the consonant sequences occurring within words. The APDB has more complete coverage, particularly for the word-final and word-medial sequences. We examined a list of all of the word-initial and word-final clusters in the sentence list, and compared these with the occurrences in Pocket. We verified that essentially every initial cluster that occurred more than once in the Pocket lexicon was included at least once in the APDB, and that most of the final clusters were covered. Often, if a word-final cluster did not occur in word-final position in the APDB, the sequence did occur within a word or across a word boundary. Generally, the sequences occurring in Pocket that are not covered by APDB are from borrowed words such "moire" and "svelte."

The APDB includes many word-final consonant sequences that were not present in MIT-450. In fact, there are more word-final consonant sequences in the APDB than actually occur in Pocket. The reason is that the Pocket lexicon does not include suffixes.

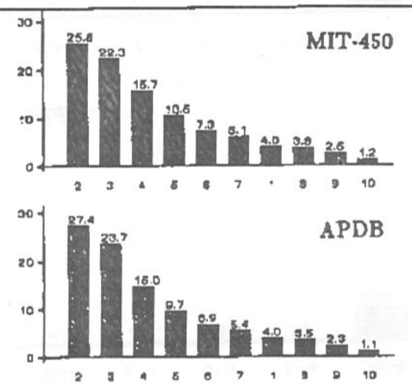


**Figure 2:** Histograms of the number of phonemes per word.

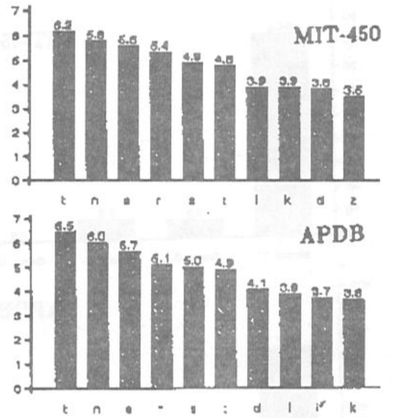


**Figure 3:** Histograms of the 10 most common phonemes.

| | POCKET | HL | MIT-450 | APDB |
|---|---|---|---|---|
| # unique words | 19,837 | 1894 | 1792 | 6103 |
| # WI | 75 | 59 | 64 | 68 |
| # WF | 129 | 105 | 102 | 146 |
| # WM | 608 | 123 | 228 | 388 |
| # boundaries | | 4305 | 2953 | 36,121 |
| # WB | | 976 | 805 | 1639 |

**Table 3: Distribution of Consonant Sequences**

A more detailed phonetic analysis of all *phoneme pairs* is included in Appendix 1 in tabular form. The tables are broken down into phoneme subsets, and data are included for both the MIT-450 and the APDB. Some of the gaps in the MIT-450 table have been filled in by sentences in the TI-1890 corpus (e.g., the syllabic /l/ column of the vowel-sonorant pairs table and the /y/ column of the vowel-sonorant pairs table). Note also that some gaps occur in both tables. Such gaps are expected, since some phoneme sequences are impossible or quite rare. For example, the lax vowels (excluding schwa) are never found in syllable-final position in English. As a consequence, table entries requiring lax vowels as the first member of a pair have many gaps (see for example, the vowel-vowel entries in the pair tables.)

Figure 4 compares histograms of the sentence types for the MIT-450 and the APDB. Simple sentences (Simple S.) and questions (Simple Q.) have no major syntactic markers. Complex sentences (Complex S.) and questions (Complex Q.) are expected to have a major syntactic boundary when read. As can be seen, the APDB has a wider variety of sentence types, with 75% being simple declarative sentences. In the MIT-450, almost 85% of the sentences are of the simple declarative form. Questions form about

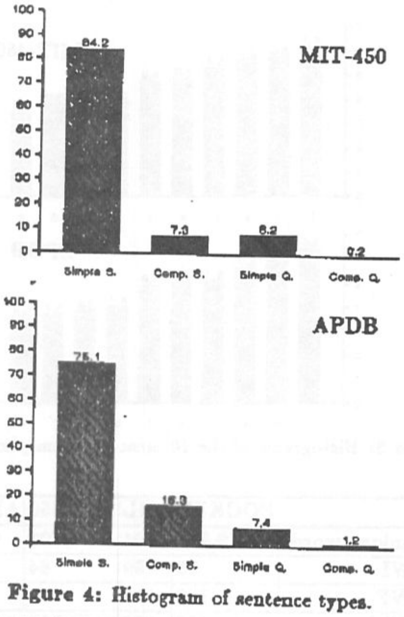


**Figure 4:** Histogram of sentence types.

10% of both corpora.

Figure 5 shows counts of environments where major phonological rules may apply. We chose to gather information on the following possibilities:

- gemination (GEM)
- vowel-vowel sequences (VVS)
- vowel-schwa sequences (VSS)
- schwa-vowel sequences (SVS)
- flapping of /t/,/d/, and /n/ (FLAP)
- homorganic stop insertion (HSI)
- schwa devoicing (S-DVC)
- fricative devoicing (F-DVC)
- /s/-/š/ and /z/-/ž/ palatalization (PAL)
- y-palatalization: /dy/→/ǰ/ (DY-Jh)
- y-palatalization: /ty/→/č/ (TY-Ch)
- y-palatalization: /sy/→/š/ (SY-Sh)

The histograms show that both corpora have many potential environments for flapping and homorganic stop insertion. The vowel-vowel environments are also well covered. The analysis for phonological rule application is difficult, because of the difficulties in predicting what different speakers will say.

## RECORDING, LABELING, AND ALIGNMENT

The recording of the sentences is currently under way at TI. Speech is recorded digitally at 20 kHz, simultaneously on a pressure-sensitive microphone and on a Sennheiser close-talking microphone. Digital tapes are shipped to NBS, where they are filtered and downsampled to 16 kHz. The resampled tapes are then shipped to MIT where the orthographic and phonetic transcriptions are generated.

Transcriptions are generated using the *Spire* facility, in conjunction with the automatic alignment system pro-

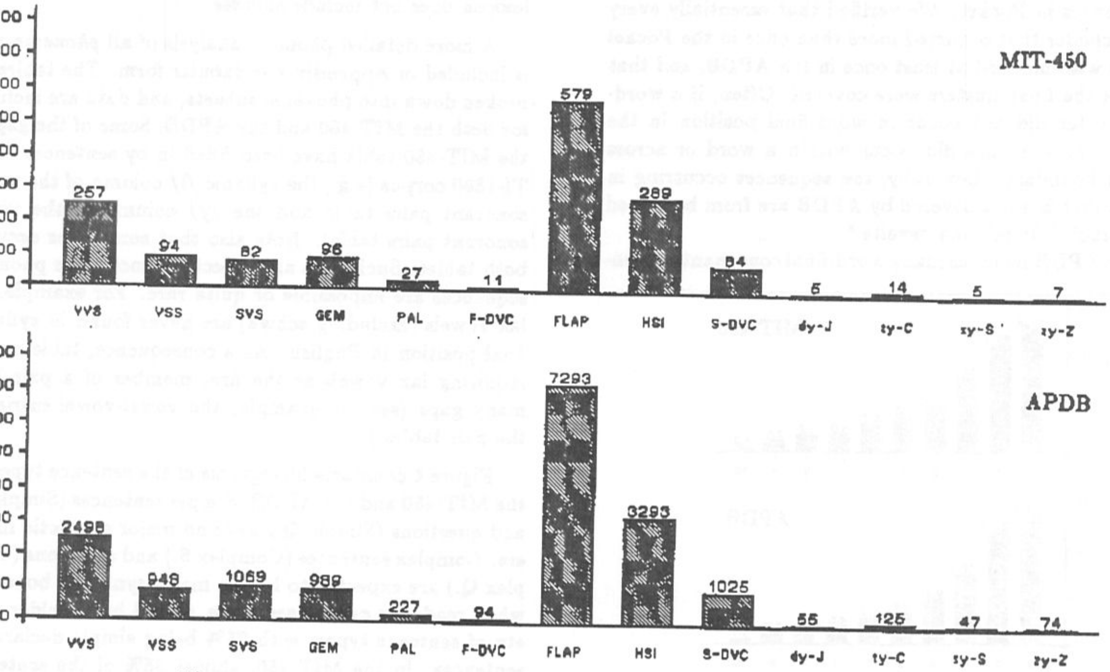


**Figure 5:** Histogram for potential application of phonological rules.

| | |
|---|---|
| Unvoiced Stops: | p t k č |
| Voiced Stops: | b d g ǰ |
| Stop Gaps: | p° t° k° ʔ b° d° g° ɾ |
| Nasals: | n m ŋ ɾ̃ |
| Syllabic Nasals: | n̩ m̩ ŋ̩ l̩ |
| Unvoiced Fricatives: | s š f θ |
| Voiced Fricatives: | z ž v ð |
| Glides: | l r w y |
| Vowels: | iʸ ɪ ɛ eʸ æ a aʷ aʸ |
| | ʌ ɔ ɔʸ oʷ ʊ u ü ɝ |
| Schwa: | ə ə̥ ɨ ɚ |
| H, Silences: | h ɦ ▯ ▫ |

Figure 6: Phones used for labeling.

vided by Leung [3]. The transcription process involves three steps:

1. A "Phonetic Sequence," which consists of a list of the phones of the utterance in correct temporal order but with no boundaries marked in time, is entered.
2. The utterance is run through an automatic system to generate an alignment for the sequence.
3. The automatically generated alignment is hand-corrected.

Only the data recorded through the pressure microphone are transcribed. Transcriptions for the close-talking version are generated by duplicating the results for the pressure microphone.

The phones used in the labeling are shown in Figure 6. In many cases, it is not possible to define a boundary between two phones, such as /ɔr/, because features appropriate for both phones often occur simultaneously in time. When no obvious positioning of the boundary is apparent, arbitrary rules, such as an automatic 2/3:1/3 split, are invoked. There are also some cases in which none of our standard phones are appropriate for a given portion of the speech, primarily because of severe coarticulation effects. In such cases, the segment is labeled as the nearest phone equivalent, according to the transcriber's judgment. There are other difficult cases, such as syllable-initial /pl/, where the /l/ is devoiced at onset. Should the portion before voicing begins be thought of as part of the aspiration of the /p/, or as part of the /l/? We have decided, somewhat arbitrarily, to define the onset time of the phone following an unvoiced stop as coincident with the onset of voicing. These remarks serve simply as examples of some of the difficulties that arise in transcribing continuous speech. We are mainly interested in using consistent methods of transcribing in situations where ambiguity exists. Currently the transcription rate is 100 sentences per week.

## SUMMARY

We have described various components of the preliminary acoustic-phonetic database and discussed some of the issues in its design. Evaluating the phonetic coverage of the database is difficult primarily because no dard for comparison exists. We have chosen to compare the phonetic coverage of the database to two well-known sources, the Merriam-Webster Pocket Dictionary of 1964 and the Harvard List sentences. The dictionary does not reflect spoken English very well, and can only guide us in judging the possible phonemic sequences within words. The Harvard List sentences, while phonemically balanced, consist primarily of very simplistic sentences and monosyllabic words. In addition, they are balanced for phoneme occurrences, whereas we tried to account for occurrences of phoneme pairs.

We believe that we have adequate coverage of most phonemes and phoneme pairs. In cases where the phoneme pairs are scarce, there are often other phoneme pairs that will provide similar information. For example, the class sequence **[alveolar consonant]** **[back vowel]** is more general than /t/ /ɔ/, and has a higher frequency of occurrence.

We hope that the APDB database will provide guidelines for the development of future databases. An analysis of the spoken corpus will enable us to judge our phonetic analysis procedure. In particular, we will be able to evaluate the relationship between our phonological rule predictions and the frequency with which a phonological modification actually occurred.

## REFERENCES

[1] Kucera, H. and W.N. Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, R.I.

[2] Egan,J. (1944) "Articulation testing methods II," OSRD Report No. 3802, U.S. Dept. of Commerce Report PB 22848, November.

[3] Leung, H. C. and V.W. Zue (1984) "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP-84*, 2.7.1-2.7.4.

# Appendix 1

## MIT

| | iʸ | eʸ | ɔ | oʷ | u | ɑʸ | ɔʸ | ɑʷ | ɝ |
|---|---|---|---|---|---|---|---|---|---|
| iʸ | 3 | 7 | 8 | 9 | | 3 | 1 | 3 | 1 |
| eʸ | 1 | 1 | 4 | 4 | | 1 | | 1 | 1 |
| ɔ | 1 | 1 | 1 | 1 | | | | 1 | |
| oʷ | 2 | 2 | 3 | 1 | | 1 | 1 | 3 | |
| u | 1 | 5 | 5 | 3 | 1 | 2 | 1 | 2 | 1 |
| ɑʸ | 4 | 4 | 6 | 2 | 1 | 3 | 1 | | 1 |
| ɔʸ | 1 | 1 | 1 | 1 | | | | | |
| ɑʷ | 1 | 2 | 1 | 1 | | 1 | 1 | | |
| ɝ | 7 | 1 | 1 | 2 | | | | | 1 |
| ɪ | | | | | | | | | |
| ɛ | | | | | | | | | |
| æ | | | | | | | | | |
| ɑ | 1 | | | | | | | | |
| ʌ | | | | | | | | | |
| ʊ | | | | | | | | | |
| ɚ | 4 | 7 | 4 | 1 | | 6 | 1 | 3 | 1 |
| ə | 3 | 1 | | 7 | | 4 | 1 | 1 | 1 |
| ɨ | 1 | | | | | | | | |

| | iʸ | eʸ | ɔ | oʷ | u | ɑʸ | ɔʸ | ɑʷ | ɝ |
|---|---|---|---|---|---|---|---|---|---|
| n | 38 | 5 | 8 | 11 | 10 | 14 | 4 | 3 | |
| m | 17 | 24 | 10 | 18 | 9 | 24 | 2 | 1 | 4 |
| ŋ | 1 | 1 | 3 | | | | | 1 | 1 |
| m̩ | | | 1 | | | | | | |
| n̩ | | 1 | | | | | | | |
| l̩ | 2 | | 2 | | | 1 | | 1 | |
| l | 88 | 23 | 13 | 25 | 9 | 22 | 4 | 12 | 7 |
| r | 81 | 28 | 16 | 28 | 24 | 24 | 3 | 5 | |
| w | 20 | 20 | 20 | 6 | 1 | 17 | 1 | 1 | 10 |
| y | | | 1 | 1 | 88 | | | | 1 |

| | iʸ | eʸ | ɔ | oʷ | u | ɑʸ | ɔʸ | ɑʷ | ɝ |
|---|---|---|---|---|---|---|---|---|---|
| b | 27 | 2 | 9 | 10 | 1 | 28 | 4 | 3 | 8 |
| d | 18 | 22 | 8 | 10 | 14 | 7 | | 9 | 2 |
| g | 1 | 3 | 1 | 13 | 4 | 1 | | 1 | 1 |
| p | 14 | 14 | 2 | 11 | 2 | 9 | 8 | 3 | 12 |
| t | 44 | 19 | 11 | 13 | 86 | 18 | 1 | | 4 |
| k | 9 | 9 | 19 | 15 | 7 | 3 | 4 | 8 | 11 |
| č | 6 | 9 | 1 | 2 | 4 | 3 | 1 | | 3 |
| ǰ | 7 | 2 | 3 | 1 | 6 | 2 | 5 | | 4 |
| s | 22 | 5 | 12 | 5 | 9 | 13 | 2 | 2 | 10 |
| z | 14 | 7 | 10 | 3 | 2 | 2 | | 1 | 4 |
| š | 11 | 2 | 5 | 5 | 3 | 1 | | 1 | 1 |
| ž | | | | | 1 | | | | |
| f | 9 | 5 | 17 | 46 | 2 | 8 | | 3 | 12 |
| v | 10 | 6 | 2 | 5 | | 3 | 2 | 1 | 4 |
| θ | 6 | | 2 | | | | | 3 | 5 |
| ð | 3 | 9 | | 5 | | | | | |
| h | 8 | | 7 | 4 | 5 | 10 | | 12 | 18 |

## APDB

| | iʸ | eʸ | ɔ | oʷ | u | ɑʸ | ɔʸ | ɑʷ | ɝ |
|---|---|---|---|---|---|---|---|---|---|
| iʸ | 28 | 70 | 75 | 88 | | 27 | 7 | 26 | 22 |
| eʸ | 10 | 7 | 38 | 31 | | 10 | | 7 | 9 |
| ɔ | 7 | 7 | 7 | 7 | | | | 7 | |
| oʷ | 14 | 14 | 26 | 10 | | 9 | 7 | 21 | 4 |
| u | 12 | 48 | 43 | 26 | 7 | 19 | 7 | 16 | 10 |
| ɑʸ | 28 | 30 | 47 | 16 | 8 | 21 | 7 | 2 | 8 |
| ɔʸ | 10 | 7 | 7 | 7 | | | | | |
| ɑʷ | 7 | 14 | 9 | 7 | | 7 | 7 | | |
| ɝ | 62 | 10 | 12 | 18 | | 3 | | 2 | 7 |
| ɪ | | | | | | | | | |
| ɛ | | | | 2 | | | | | 3 |
| æ | | | | | | | | | |
| ɑ | 8 | | | | | | | | |
| ʌ | | | | | | | | | |
| ʊ | | | | | | | | | |
| ɚ | 64 | 73 | 37 | 13 | | 52 | 10 | 25 | 7 |
| ə | 22 | 8 | 7 | 60 | | 33 | 7 | 11 | 10 |
| ɨ | 7 | | | | | | | | |

| | iʸ | eʸ | ɔ | oʷ | u | ɑʸ | ɔʸ | ɑʷ | ɝ |
|---|---|---|---|---|---|---|---|---|---|
| n | 422 | 79 | 88 | 252 | 117 | 133 | 32 | 95 | 21 |
| m | 246 | 318 | 100 | 227 | 86 | 248 | 18 | 22 | 48 |
| ŋ | 9 | 7 | 32 | 8 | | 1 | 1 | 10 | 7 |
| m̩ | | | 7 | | | | | | |
| n̩ | | 7 | | | | | | | |
| l̩ | 55 | 5 | 17 | 5 | | 13 | | 7 | 2 |
| l | 1001 | 287 | 168 | 286 | 115 | 308 | 34 | 110 | 60 |
| r | 936 | 312 | 169 | 303 | 260 | 272 | 29 | 71 | 6 |
| w | 294 | 212 | 199 | 67 | 8 | 210 | 7 | 9 | 134 |
| y | 3 | | 9 | 7 | 933 | | | 5 | 10 |

| | iʸ | eʸ | ɔ | oʷ | u | ɑʸ | ɔʸ | ɑʷ | ɝ |
|---|---|---|---|---|---|---|---|---|---|
| b | 364 | 35 | 78 | 118 | 15 | 282 | 48 | 54 | 79 |
| d | 270 | 212 | 95 | 133 | 165 | 80 | 1 | 107 | 44 |
| g | 7 | 51 | 18 | 137 | 29 | 15 | 1 | 7 | 28 |
| p | 149 | 141 | 32 | 142 | 21 | 81 | 75 | 38 | 137 |
| t | 501 | 234 | 157 | 175 | 1060 | 220 | 9 | 21 | 99 |
| k | 99 | 137 | 191 | 146 | 56 | 31 | 32 | 75 | 106 |
| č | 54 | 73 | 10 | 18 | 56 | 28 | 9 | 1 | 26 |
| ǰ | 61 | 16 | 25 | 12 | 55 | 17 | 40 | | 36 |
| s | 275 | 83 | 137 | 176 | 85 | 181 | 20 | 26 | 123 |
| z | 130 | 63 | 109 | 58 | 18 | 33 | | 12 | 37 |
| š | 198 | 33 | 42 | 54 | 35 | 9 | | 11 | 15 |
| ž | | | | | 22 | | | | |
| f | 120 | 72 | 172 | 506 | 25 | 117 | 4 | 32 | 120 |
| v | 92 | 68 | 28 | 44 | | 60 | 28 | 15 | 63 |
| θ | 56 | 2 | 31 | 3 | 3 | 3 | | 37 | 53 |
| ð | 65 | 192 | | 59 | | | | 2 | 3 |
| h | 381 | 18 | 58 | 79 | 54 | 117 | | 137 | 213 |

## MIT

| | ɪ | ɛ | æ | ɑ | ʌ | ʊ | ɚ | ə | ɨ |
|---|---|---|---|---|---|---|---|---|---|
| i$^y$ | 11 | 4 | 16 | 11 | | | 5 | 28 | 1 |
| e$^y$ | 4 | 3 | 3 | 1 | 1 | | 1 | 3 | |
| ɔ | 1 | 1 | 1 | 2 | | | | | 1 |
| o$^w$ | 4 | 1 | 2 | 3 | | | 1 | | 3 |
| u | 6 | 7 | 4 | 6 | 1 | | 1 | 4 | |
| ɑ$^y$ | 5 | 2 | 7 | 3 | 2 | | 4 | 11 | 6 |
| ɔ$^y$ | 3 | 1 | 1 | | | | 2 | 1 | 2 |
| ɑ$^w$ | | 1 | 1 | | 1 | | 3 | 3 | |
| ɝ | 2 | | 1 | | 1 | | | 1 | 1 |
| ɪ | | | | | | | | | |
| ɛ | | | | | | | | | |
| æ | | | | | | | | | |
| ɑ | | | | | | | | | |
| ʌ | | | | | | | | | |
| ʊ | | | | | | | | | |
| ɚ | 6 | 4 | 7 | 6 | 2 | | 1 | 7 | 4 |
| ə | 5 | 6 | 8 | 2 | 2 | | | 3 | |
| ɨ | | | | | | | | | |

| | ɪ | ɛ | æ | ɑ | ʌ | ʊ | ɚ | ə | ɨ |
|---|---|---|---|---|---|---|---|---|---|
| n | 38 | 16 | 13 | 19 | 10 | | 16 | 29 | 14 |
| m | 17 | 15 | 13 | 7 | 23 | | 9 | 44 | 8 |
| ŋ | 12 | | 3 | 1 | | | | 2 | 2 |
| m̩ | | | | | | | | | |
| n̩ | 2 | 3 | 1 | 1 | | | | 1 | 1 |
| l̩ | 10 | 2 | 2 | 1 | | | | 7 | 3 |
| l | 37 | 22 | 29 | 20 | 11 | 5 | 12 | 33 | 24 |
| r | 69 | 39 | 33 | 26 | 38 | 1 | 2 | 49 | 16 |
| w | 60 | 25 | 3 | 7 | 7 | 12 | 6 | 30 | |
| y | 4 | 4 | | 3 | 6 | 38 | 2 | 1 | 1 |

| | ɪ | ɛ | æ | ɑ | ʌ | ʊ | ɚ | ə | ɨ |
|---|---|---|---|---|---|---|---|---|---|
| b | 25 | 12 | 16 | 15 | 11 | 3 | 5 | 12 | 5 |
| d | 56 | 9 | 17 | 6 | 7 | 1 | 16 | 56 | 22 |
| g | 9 | 10 | 9 | 13 | 7 | 4 | 9 | 3 | 5 |
| p | 12 | 13 | 14 | 20 | 2 | 5 | 23 | 16 | 3 |
| t | 63 | 21 | 30 | 16 | 8 | 2 | 44 | 47 | 45 |
| k | 10 | 11 | 42 | 25 | 23 | 6 | 6 | 33 | 14 |
| č | 6 | 4 | 2 | 7 | 1 | | 10 | 3 | 6 |
| ǰ | 12 | 12 | 4 | 4 | 6 | | 5 | 2 | 25 |
| s | 44 | 15 | 11 | 8 | 19 | 1 | 7 | 39 | 20 |
| z | 37 | 7 | 17 | 13 | 4 | | 4 | 36 | 9 |
| š | 5 | 5 | 3 | 2 | | 9 | 1 | 3 | 42 |
| ž | | | | | | 1 | 4 | | 3 |
| f | 23 | 5 | 6 | 5 | 5 | 2 | 2 | 15 | 2 |
| v | 12 | 12 | 10 | 3 | | | 33 | 16 | 3 |
| θ | 14 | 1 | 1 | 1 | | | 4 | 3 | |
| ð | 15 | 14 | 21 | | 1 | | 12 | 219 | |
| h | 24 | 7 | 31 | 15 | 3 | 3 | | | |

## APDB

| | ɪ | ɛ | æ | ɑ | ʌ | ʊ | ɚ | ə | ɨ |
|---|---|---|---|---|---|---|---|---|---|
| i$^y$ | 174 | 47 | 182 | 102 | 7 | | 46 | 399 | 27 |
| e$^y$ | 38 | 22 | 26 | 21 | 9 | | 8 | 38 | 7 |
| ɔ | 9 | 7 | 9 | 14 | 2 | | | 3 | 8 |
| o$^w$ | 54 | 14 | 26 | 25 | 3 | | 28 | 28 | 34 |
| u | 75 | 58 | 48 | 54 | 11 | | 8 | 79 | 13 |
| ɑ$^y$ | 51 | 18 | 61 | 23 | 14 | | 30 | 107 | 61 |
| ɔ$^y$ | 21 | 8 | 8 | | | | 14 | 7 | 15 |
| ɑ$^w$ | 5 | 17 | 9 | 1 | 8 | | 32 | 30 | |
| ɝ | 27 | 8 | 13 | 1 | 9 | | | 30 | 16 |
| ɪ | | | | | | | | | |
| ɛ | | | | | | | | 1 | |
| æ | | | | | | | | | |
| ɑ | | | | | | | | | |
| ʌ | | | | | | | | | |
| ʊ | | | | | | | | | |
| ɚ | 89 | 38 | 73 | 48 | 21 | | 8 | 106 | 52 |
| ə | 57 | 52 | 82 | 19 | 23 | | | 42 | 2 |
| ɨ | | | | | | | | | |

| | ɪ | ɛ | æ | ɑ | ʌ | ʊ | ɚ | ə | ɨ |
|---|---|---|---|---|---|---|---|---|---|
| n | 434 | 185 | 185 | 260 | 148 | 3 | 144 | 388 | 214 |
| m | 231 | 223 | 219 | 93 | 240 | | 91 | 516 | 87 |
| ŋ | 125 | 6 | 51 | 10 | 11 | | | 43 | 14 |
| m̩ | 1 | | 3 | | | | | 1 | |
| n̩ | 18 | 24 | 11 | 9 | | | 2 | 11 | 10 |
| l̩ | 93 | 29 | 37 | 10 | 2 | | | 96 | 43 |
| l | 397 | 281 | 330 | 191 | 127 | 67 | 119 | 342 | 279 |
| r | 808 | 415 | 364 | 271 | 393 | 10 | 17 | 562 | 211 |
| w | 659 | 293 | 31 | 125 | 150 | 145 | 144 | 376 | 5 |
| y | 44 | 55 | 3 | 29 | 46 | 353 | 44 | 15 | 9 |

| | ɪ | ɛ | æ | ɑ | ʌ | ʊ | ɚ | ə | ɨ |
|---|---|---|---|---|---|---|---|---|---|
| b | 271 | 123 | 185 | 145 | 204 | 31 | 69 | 185 | 41 |
| d | 646 | 130 | 209 | 82 | 121 | 10 | 174 | 595 | 245 |
| g | 99 | 132 | 97 | 118 | 67 | 57 | 87 | 30 | 54 |
| p | 169 | 150 | 166 | 208 | 41 | 65 | 222 | 223 | 44 |
| t | 764 | 288 | 326 | 160 | 105 | 35 | 456 | 690 | 481 |
| k | 127 | 97 | 415 | 266 | 251 | 81 | 51 | 404 | 178 |
| č | 61 | 41 | 29 | 56 | 14 | 4 | 101 | 47 | 70 |
| ǰ | 110 | 120 | 35 | 39 | 68 | 1 | 50 | 29 | 230 |
| s | 523 | 349 | 166 | 88 | 268 | 8 | 86 | 488 | 270 |
| z | 386 | 69 | 228 | 157 | 52 | | 36 | 437 | 112 |
| š | 56 | 48 | 40 | 32 | 14 | 112 | 13 | 84 | 472 |
| ž | | | 1 | | | 7 | 39 | 3 | 34 |
| f | 268 | 87 | 96 | 74 | 46 | 26 | 39 | 172 | 31 |
| v | 152 | 145 | 124 | 40 | 11 | | 345 | 230 | 52 |
| θ | 171 | 16 | 18 | 7 | 3 | | 34 | 50 | 19 |
| ð | 238 | 196 | 323 | | 14 | | 187 | 2230 | |
| h | 455 | 118 | 452 | 137 | 46 | 27 | 9 | 2 | 4 |

## MIT

| | b | d | g | p | t | k | č | ǰ |
|---|---|---|---|---|---|---|---|---|
| $i^y$ | 7 | 27 | 9 | 27 | 32 | 25 | 17 | 1 |
| $e^y$ | 8 | 13 | 2 | 9 | 47 | 17 | 2 | 7 |
| ɔ | 2 | 5 | 5 | 1 | 13 | 4 | 3 | 1 |
| $o^w$ | 5 | 2 | 2 | 9 | 15 | 12 | 3 | |
| u | 10 | 14 | 5 | 15 | 17 | 10 | 4 | 7 |
| $a^y$ | 7 | 17 | 2 | 2 | 24 | 14 | 1 | 1 |
| ɔ$^y$ | 2 | 2 | 2 1 | 2 | | | | |
| $a^w$ | | 2 | 1 | 1 | 22 | 1 | | |
| ɝ | 8 | 5 | 1 | 5 | 10 | 10 | 6 | 5 |
| ɪ | 5 | 13 | 28 | 16 | 21 | 87 | 8 | 12 |
| ɛ | 2 | 16 | 8 | 5 | 20 | 37 | 1 | 8 |
| æ | 10 | 10 | 10 | 9 | 37 | 27 | 4 | 2 |
| ɑ | 13 | 7 | 3 | 19 | 31 | 19 | 2 | 10 |
| ʌ | 10 | 3 | 5 | 15 | 12 | 4 | 8 | 3 |
| ʊ | 1 | 24 | 1 | 1 | 4 | 13 | 1 | |
| ɚ | 6 | 14 | 2 | 3 | 10 | 8 | 4 | |
| ə | 48 | 46 | 28 | 45 | 43 | 46 | 2 | 14 |
| ɨ | 8 | 13 | 2 | 3 | 27 | 9 | | 8 |

## APDB

| | b | d | g | p | t | k | č | ǰ |
|---|---|---|---|---|---|---|---|---|
| $i^y$ | 116 | 357 | 102 | 292 | 387 | 270 | 155 | 13 |
| $e^y$ | 113 | 188 | 20 | 84 | 503 | 228 | 18 | 62 |
| ɔ | 16 | 41 | 48 | 8 | 130 | 52 | 23 | 7 |
| $o^w$ | 68 | 61 | 30 | 104 | 157 | 124 | 27 | |
| u | 121 | 173 | 58 | 152 | 196 | 120 | 28 | 57 |
| $a^y$ | 60 | 249 | 19 | 31 | 292 | 166 | 7 | 8 |
| ɔ$^y$ | 15 | 22 | 7 | 14 | 1 | 3 | | |
| $a^w$ | 3 | 42 | 9 | 8 | 262 | 8 | | 3 |
| ɝ | 69 | 75 | 11 | 53 | 121 | 112 | 52 | 46 |
| ɪ | 48 | 186 | 266 | 186 | 501 | 873 | 100 | 145 |
| ɛ | 16 | 224 | 76 | 83 | 269 | 438 | 8 | 71 |
| æ | 95 | 192 | 99 | 118 | 519 | 357 | 43 | 26 |
| ɑ | 133 | 91 | 33 | 195 | 408 | 183 | 16 | 87 |
| ʌ | 101 | 54 | 60 | 171 | 202 | 64 | 104 | 23 |
| ʊ | 8 | 314 | 10 | 8 | 48 | 149 | 7 | |
| ɚ | 63 | 203 | 23 | 43 | 101 | 83 | 34 | 5 |
| ə | 537 | 616 | 307 | 477 | 517 | 514 | 31 | 116 |
| ɨ | 79 | 168 | 28 | 37 | 387 | 137 | | 75 |

| | b | d | g | p | t | k | č | ǰ |
|---|---|---|---|---|---|---|---|---|
| n | 13 | 119 | 7 | 11 | 120 | 17 | 5 | 16 |
| m | 15 | 3 | 3 | 33 | 6 | 6 | 1 | 1 |
| ŋ | 3 | 4 | 9 | 3 | 4 | 10 | 1 | 1 |
| m̩ | | | | | | | | |
| n̩ | 1 | 3 | | | 11 | | 1 | |
| l̩ | | 5 | 4 | 5 | 8 | 4 | 1 | |
| l | 7 | 29 | 1 | 6 | 7 | 8 | 1 | 1 |
| r | 16 | 36 | 9 | 5 | 22 | 15 | 5 | 15 |
| w | | | | | | | | |
| y | | | | | | | | |

| | b | d | g | p | t | k | č | ǰ |
|---|---|---|---|---|---|---|---|---|
| n | 144 | 1482 | 72 | 118 | 1343 | 204 | 76 | 155 |
| m | 181 | 63 | 26 | 369 | 80 | 55 | 9 | 9 |
| ŋ | 39 | 37 | 125 | 41 | 77 | 146 | 7 | 10 |
| m̩ | | 1 | | | | | | |
| n̩ | 8 | 28 | | 1 | 141 | 2 | 7 | |
| l̩ | 24 | 75 | 40 | 55 | 84 | 47 | 10 | 1 |
| l | 71 | 348 | 15 | 74 | 129 | 79 | 14 | 15 |
| r | 156 | 359 | 93 | 86 | 277 | 173 | 41 | 121 |
| w | | | | | | | | |
| y | | | | | | | | |

| | b | d | g | p | t | k | č | ǰ |
|---|---|---|---|---|---|---|---|---|
| b | 3 | 5 | 1 | 2 | 4 | 2 | 1 | 3 |
| d | 25 | 5 | 3 | 5 | 14 | 8 | 2 | 3 |
| g | 4 | 3 | 3 | 2 | 3 | 1 | 1 | 1 |
| p | 4 | 1 | 2 | 4 | 11 | 2 | 2 | 1 |
| t | 18 | 14 | 6 | 11 | 18 | 13 | 1 | 3 |
| k | 2 | 6 | 3 | 7 | 45 | 6 | 1 | 1 |
| č | 2 | 2 | 3 | 1 | 4 | 3 | 2 | 1 |
| ǰ | 1 | 8 | 1 | 3 | 4 | 1 | 2 | 2 |
| s | 9 | 7 | 4 | 46 | 158 | 56 | 7 | 1 |
| z | 15 | 21 | 9 | 11 | 16 | 17 | 2 | 4 |
| š | 1 | 1 | 1 | 1 | 4 | 1 | 1 | |
| ž | | 1 | | | | | | |
| f | 1 | 1 | 1 | 2 | 12 | 3 | 1 | |
| v | 1 | 3 | 1 | 7 | 3 | 5 | 1 | 1 |
| θ | | 5 | | 2 | 1 | 2 | | 1 |
| ð | 2 | 1 | | 1 | | 1 | 1 | |
| h | | | | | | | | |

| | b | d | g | p | t | k | č | ǰ |
|---|---|---|---|---|---|---|---|---|
| b | 24 | 41 | 8 | 14 | 30 | 14 | 7 | 27 |
| d | 283 | 82 | 51 | 95 | 196 | 105 | 16 | 30 |
| g | 28 | 30 | 21 | 17 | 21 | 8 | 7 | 7 |
| p | 33 | 11 | 16 | 32 | 138 | 18 | 15 | 7 |
| t | 210 | 161 | 75 | 119 | 214 | 154 | 15 | 24 |
| k | 26 | 48 | 31 | 58 | 513 | 59 | 26 | 7 |
| č | 16 | 17 | 22 | 9 | 46 | 26 | 14 | 8 |
| ǰ | 12 | 67 | 8 | 25 | 31 | 9 | 14 | 14 |
| s | 98 | 77 | 39 | 503 | 1778 | 523 | 69 | 12 |
| z | 174 | 221 | 85 | 124 | 189 | 175 | 19 | 40 |
| š | 8 | 8 | 7 | 8 | 44 | 7 | 7 | |
| ž | | 7 | | | | | | |
| f | 13 | 17 | 8 | 15 | 155 | 31 | 7 | |
| v | 36 | 77 | 21 | 67 | 50 | 61 | 8 | 9 |
| θ | 7 | 38 | 3 | 16 | 14 | 24 | | 7 |
| ð | 14 | 8 | | 7 | | 7 | 7 | |
| h | | | | | | | | |

## MIT

| | n | m | ŋ | m̩ | n̩ | l̩ | l | r | w | y |
|---|---|---|---|---|---|---|---|---|---|---|
| iʸ | 23 | 15 | | | | 1 | 15 | 11 | 15 | 2 |
| eʸ | 29 | 11 | | | | | 8 | 1 | 2 | 1 |
| ɔ | 29 | 1 | 10 | | | | 39 | 61 | | |
| oʷ | 24 | 6 | | | | | 20 | 77 | 3 | |
| u | 15 | 23 | | | | 1 | 9 | 5 | 7 | 5 |
| ɑʸ | 16 | 15 | | | | 1 | 15 | 12 | 2 | 2 |
| ɔʸ | 9 | 1 | | | | | 4 | | | |
| ɑʷ | 20 | 2 | | | | | 2 | 10 | 2 | |
| ɝ | 8 | 7 | | | | | 4 | 1 | 5 | |
| ɪ | 112 | 32 | 46 | | | | 49 | 24 | 2 | |
| ɛ | 56 | 17 | 3 | | | | 48 | 26 | | |
| æ | 125 | 24 | 8 | | | | 23 | 31 | | |
| ɑ | 17 | 15 | 1 | | | | 16 | 100 | | |
| ʌ | 46 | 34 | 11 | | | | 11 | | | |
| ʊ | | 2 | | | | | 16 | 30 | | |
| ɚ | 7 | 5 | | | | 5 | 6 | 2 | 13 | 2 |
| ə | 104 | 62 | | | | | 50 | 10 | 10 | 4 |
| ɨ | 94 | 1 | 60 | | | | | | | |

| | n | m | ŋ | m̩ | n̩ | l̩ | l | r | w | y |
|---|---|---|---|---|---|---|---|---|---|---|
| n | 5 | 15 | | | | 8 | 9 | 3 | 8 | 8 |
| m | 3 | 10 | | | | 2 | 5 | 2 | 3 | 10 |
| ŋ | 2 | 1 | | | | | 2 | 5 | 5 | 4 |
| m̩ | | | | | | | 1 | | | |
| n̩ | | | | | | 2 | | | 1 | |
| l̩ | 2 | 1 | | | | | 4 | 3 | 2 | |
| l | 2 | 8 | | | | | 3 | 3 | 12 | 8 |
| r | 16 | 23 | | | 1 | 2 | 18 | 5 | 13 | 6 |
| w | | | | | | 1 | | | | |
| y | | | | | 1 | | | | | |

| | n | m | ŋ | m̩ | n̩ | l̩ | l | r | w | y |
|---|---|---|---|---|---|---|---|---|---|---|
| b | 1 | 1 | | | | 21 | 21 | 27 | 2 | 9 |
| d | 8 | 5 | | | 11 | 4 | 7 | 32 | 10 | 5 |
| g | 3 | 2 | | | | 4 | 8 | 36 | 6 | 5 |
| p | 4 | 3 | | | | 10 | 41 | 57 | 4 | 10 |
| t | 6 | 10 | | 5 | 5 | 7 | 18 | 63 | 16 | 14 |
| k | 2 | 3 | | | 1 | 25 | 36 | 30 | 26 | 18 |
| č | 1 | 1 | | | 1 | 1 | 2 | 1 | 1 | 2 |
| ǰ | 1 | 2 | | | | | 1 | 1 | 1 | 1 |
| s | 5 | 18 | | | 6 | 2 | 16 | 4 | 18 | 5 |
| z | 9 | 14 | | 2 | 7 | | 4 | 9 | 14 | 7 |
| š | 1 | 3 | | | 7 | 5 | 2 | 2 | | 1 |
| ž | | 1 | | | | | | | | |
| f | 1 | 1 | | | | 6 | 12 | 35 | 1 | 6 |
| v | 8 | 5 | | | 1 | 6 | 3 | 12 | 3 | 5 |
| θ | 1 | 1 | | | | | 1 | 12 | 1 | |
| ð | | | | 1 | | | | | | 1 |
| h | | | | | | | | | 18 | 5 |

## APDB

| | n | m | ŋ | m̩ | n̩ | l̩ | l | r | w | y |
|---|---|---|---|---|---|---|---|---|---|---|
| iʸ | 276 | 224 | | | | 41 | 213 | 143 | 219 | 31 |
| eʸ | 293 | 157 | | | | 6 | 109 | 32 | 43 | 10 |
| ɔ | 335 | 10 | 127 | | | | 480 | 574 | | 2 |
| oʷ | 293 | 92 | | | | 1 | 258 | 987 | 38 | 3 |
| u | 161 | 247 | | | | 60 | 115 | 72 | 78 | 57 |
| ɑʸ | 220 | 171 | | | | 14 | 161 | 151 | 22 | 17 |
| ɔʸ | 83 | 7 | | | | 5 | 55 | | 1 | 2 |
| ɑʷ | 241 | 22 | | | | 2 | 21 | 122 | 20 | 2 |
| ɝ | 104 | 79 | | | | 6 | 63 | 19 | 43 | 6 |
| ɪ | 1372 | 423 | 500 | | | | 509 | 330 | 16 | |
| ɛ | 795 | 213 | 27 | | | | 540 | 340 | | |
| æ | 1544 | 233 | 90 | | | 1 | 231 | 350 | | |
| ɑ | 201 | 160 | 8 | | | | 183 | 1022 | | |
| ʌ | 522 | 437 | 108 | | | | 121 | | | |
| ʊ | | 28 | | | | | 163 | 299 | | |
| ɚ | 74 | 58 | | | | 63 | 62 | 35 | 110 | 20 |
| ə | 1324 | 680 | 6 | | | | 537 | 162 | 152 | 37 |
| ɨ | 1013 | 37 | 758 | | | | 5 | | | |

| | n | m | ŋ | m̩ | n̩ | l̩ | l | r | w | y |
|---|---|---|---|---|---|---|---|---|---|---|
| n | 59 | 169 | | | | 88 | 129 | 50 | 113 | 108 |
| m | 33 | 88 | | | | 33 | 49 | 24 | 53 | 90 |
| ŋ | 23 | 28 | | | | | 33 | 48 | 50 | 31 |
| m̩ | 1 | | | | | | 7 | | 2 | |
| n̩ | 1 | 1 | | | | 20 | 7 | 1 | 9 | |
| l̩ | 24 | 20 | | | | | 85 | 31 | 24 | 2 |
| l | 27 | 85 | | 2 | | | 29 | 41 | 115 | 81 |
| r | 160 | 262 | | | 9 | 34 | 167 | 55 | 142 | 58 |
| w | | | | | | 8 | | 1 | | |
| y | | | | | 7 | 1 | | | | |

| | n | m | ŋ | m̩ | n̩ | l̩ | l | r | w | y |
|---|---|---|---|---|---|---|---|---|---|---|
| b | 8 | 9 | | | | 236 | 228 | 274 | 17 | 82 |
| d | 114 | 106 | | | 116 | 43 | 123 | 330 | 148 | 55 |
| g | 46 | 19 | | | | 43 | 98 | 357 | 47 | 56 |
| p | 30 | 33 | | | | 102 | 425 | 633 | 35 | 85 |
| t | 99 | 152 | | 39 | 60 | 102 | 222 | 728 | 250 | 125 |
| k | 36 | 37 | | | 7 | 251 | 376 | 324 | 287 | 185 |
| č | 9 | 12 | | | 7 | 15 | 20 | 10 | 13 | 14 |
| ǰ | 8 | 16 | | | | 4 | 10 | 7 | 9 | 11 |
| s | 66 | 184 | | | 70 | 31 | 197 | 54 | 213 | 47 |
| z | 154 | 163 | | 29 | 100 | 8 | 71 | 98 | 175 | 74 |
| š | 10 | 24 | | | 49 | 55 | 17 | 22 | 1 | 7 |
| ž | | 7 | | | | | | | 1 | |
| f | 10 | 11 | | | | 98 | 149 | 330 | 16 | 71 |
| v | 68 | 71 | | | 8 | 74 | 44 | 129 | 37 | 54 |
| θ | 9 | 14 | | 1 | | 1 | 16 | 128 | 15 | 1 |
| ð | | 1 | | 7 | | | | | | 7 |
| h | | | | | | 1 | | | 315 | 44 |

## MIT

| | s | z | š | ž | f | v | θ | ð | h |
|---|---|---|---|---|---|---|---|---|---|
| i$^y$ | 28 | 43 | 3 | 1 | 11 | 18 | 2 | 6 | 6 |
| e$^y$ | 14 | 12 | 20 | 2 | 2 | 5 | 1 | | 1 |
| ɔ | 9 | 6 | 2 | | 12 | | 5 | | 2 |
| o$^w$ | 14 | 18 | 5 | 2 | 2 | 16 | 4 | 1 | 2 |
| u | 18 | 25 | 3 | | 8 | 9 | 5 | 10 | 8 |
| a$^y$ | 15 | 17 | 1 | | 2 | 8 | | 4 | 2 |
| ɔ$^y$ | 7 | 5 | | | | | | 1 | |
| a$^w$ | 6 | 1 | | | | | | | |
| ɝ | 10 | 7 | 1 | | 5 | 6 | 4 | 2 | 1 |
| ɪ | 50 | 64 | 32 | 2 | 21 | 22 | 7 | 24 | |
| ɛ | 27 | 3 | 7 | 2 | 4 | 25 | | 3 | |
| æ | 17 | 15 | 6 | 1 | 17 | 15 | 3 | 1 | |
| ɑ | 11 | 1 | 2 | 5 | 2 | 2 | | | |
| ʌ | 16 | 2 | 3 | | 4 | 7 | 2 | 7 | |
| ʊ | | | | | | | | | |
| ɚ | 7 | 36 | 2 | | 9 | 4 | 1 | 7 | 8 |
| ə | 67 | 54 | 3 | | 28 | 55 | 6 | | 14 |
| ɨ | 43 | 17 | | | 2 | 2 | 1 | | |

| | s | z | š | ž | f | v | θ | ð | h |
|---|---|---|---|---|---|---|---|---|---|
| n | 60 | 45 | 7 | | 12 | 1 | 3 | 34 | 11 |
| m | 6 | 20 | 1 | | 8 | 1 | 2 | 11 | 4 |
| ŋ | 10 | 5 | 2 | | 5 | 2 | 2 | 4 | 8 |
| m̩ | | 2 | | | | | | | |
| n̩ | 2 | 4 | | | 1 | 1 | | 1 | |
| l̩ | 6 | 17 | | | 6 | 1 | 1 | 1 | |
| l | 7 | 13 | 1 | | 11 | 3 | 5 | 5 | 4 |
| r | 15 | 16 | 2 | | 8 | 3 | 2 | 9 | 1 |
| w | | | | | | | | | |
| y | | | | | | | | | |

| | s | z | š | ž | f | v | θ | ð | h |
|---|---|---|---|---|---|---|---|---|---|
| b | 2 | 4 | 1 | | 1 | 1 | 1 | 1 | 1 |
| d | 15 | 17 | 1 | | 15 | 3 | 4 | 15 | 11 |
| g | 1 | 11 | 1 | 1 | 1 | | 1 | 1 | 1 |
| p | 13 | 1 | | | 2 | 1 | 1 | 3 | 1 |
| t | 77 | 1 | 2 | | 12 | 2 | 1 | 21 | 12 |
| k | 53 | 1 | 4 | | 6 | 1 | 1 | 4 | 2 |
| ǰ | 1 | | 1 | | 1 | | 1 | 2 | 1 |
| s | 7 | 1 | 7 | | 10 | 1 | 2 | 7 | 7 |
| z | 19 | | 9 | | 19 | 2 | 2 | 5 | 16 |
| š | 3 | 1 | 1 | | 3 | 1 | 1 | | |
| ž | 1 | | | | | | | | 1 |
| f | 6 | | | | 1 | 1 | 3 | 1 | |
| v | 4 | 9 | 1 | | 6 | 1 | | 11 | 5 |
| θ | 3 | 1 | | | 2 | 1 | 1 | | |
| ð | 1 | | | | 2 | 1 | | 4 | 1 |
| h | | | | | | | | | |

## APDB

| | s | z | š | ž | f | v | θ | ð | h |
|---|---|---|---|---|---|---|---|---|---|
| i$^y$ | 374 | 435 | 52 | 9 | 170 | 194 | 56 | 102 | 158 |
| e$^y$ | 194 | 121 | 248 | 15 | 36 | 72 | 13 | 10 | 22 |
| ɔ | 105 | 55 | 19 | | 147 | 1 | 45 | 5 | 14 |
| o$^w$ | 174 | 184 | 59 | 17 | 24 | 165 | 51 | 29 | 33 |
| u | 193 | 216 | 38 | 19 | 88 | 91 | 53 | 131 | 109 |
| a$^y$ | 154 | 211 | 11 | | 53 | 94 | 3 | 42 | 23 |
| ɔ$^y$ | 67 | 45 | | | 2 | | | 7 | |
| a$^w$ | 55 | 22 | 2 | | 2 | 5 | 7 | 7 | 7 |
| ɝ | 129 | 97 | 17 | | 51 | 68 | 44 | 23 | 13 |
| ɪ | 647 | 901 | 290 | 21 | 252 | 251 | 152 | 168 | 5 |
| ɛ | 332 | 43 | 67 | 23 | 47 | 286 | 12 | 33 | 1 |
| æ | 206 | 263 | 67 | 9 | 166 | 180 | 28 | 14 | 1 |
| ɑ | 119 | 11 | 16 | 36 | 18 | 19 | 2 | 8 | 1 |
| ʌ | 238 | 52 | 30 | | 58 | 83 | 25 | 111 | |
| ʊ | | 2 | 5 | | 1 | | | | |
| ɚ | 100 | 346 | 23 | | 86 | 40 | 10 | 83 | 95 |
| ə | 797 | 639 | 66 | | 325 | 825 | 66 | 2 | 174 |
| ɨ | 452 | 162 | 12 | | 34 | 46 | 7 | | 1 |

| | s | z | š | ž | f | v | θ | ð | h |
|---|---|---|---|---|---|---|---|---|---|
| n | 718 | 451 | 83 | | 165 | 43 | 37 | 367 | 180 |
| m | 93 | 213 | 9 | | 81 | 9 | 26 | 97 | 58 |
| ŋ | 94 | 61 | 17 | | 48 | 18 | 23 | 54 | 82 |
| m̩ | | 18 | 1 | | | | | 1 | |
| n̩ | 25 | 36 | | | 9 | 7 | 1 | 10 | 2 |
| l̩ | 64 | 169 | 4 | | 67 | 12 | 10 | 11 | 24 |
| l | 123 | 152 | 16 | | 140 | 52 | 39 | 57 | 58 |
| r | 191 | 157 | 25 | 1 | 101 | 30 | 27 | 110 | 60 |
| w | | | | | | | | | 5 |
| y | | | | | | | | | |

| | s | z | š | ž | f | v | θ | ð | h |
|---|---|---|---|---|---|---|---|---|---|
| b | 32 | 36 | 7 | | 8 | 12 | 7 | 7 | 8 |
| d | 191 | 233 | 25 | | 174 | 54 | 43 | 191 | 214 |
| g | 10 | 131 | 8 | 7 | 10 | | 8 | 8 | 10 |
| p | 152 | 7 | 11 | | 23 | 7 | 7 | 35 | 14 |
| t | 910 | 7 | 41 | | 158 | 24 | 18 | 272 | 233 |
| k | 604 | 7 | 81 | | 66 | 11 | 8 | 49 | 47 |
| č | 32 | 7 | 7 | | 16 | 7 | 7 | 14 | 14 |
| ǰ | 11 | | 7 | | 9 | | 7 | 16 | 11 |
| s | 92 | 7 | 61 | | 127 | 14 | 21 | 95 | 107 |
| z | 227 | 1 | 80 | | 185 | 26 | 22 | 129 | 204 |
| š | 22 | 7 | 7 | | 22 | 7 | 9 | 1 | 3 |
| ž | 7 | | | | | | | 1 | 7 |
| f | 62 | | 7 | | 10 | 7 | 21 | 17 | 14 |
| v | 72 | 98 | 10 | | 58 | 13 | 2 | 152 | 71 |
| θ | 35 | 7 | 2 | | 20 | 9 | 8 | 16 | 15 |
| ð | 7 | 3 | | | 15 | 7 | | 30 | 7 |
| h | | | | | | | | | |

# STUDIES FOR AN ADAPTIVE RECOGNITION LEXICON[1]

Michael Cohen, Gay Baldwin, Jared Bernstein, Hy Murveit, Mitchel Weintraub
Speech Research Program
SRI International
Menlo Park, CA 94025

## ABSTRACT

In the past year, SRI has undertaken a series of empirical studies of phonological variation. The goal has been to find better lexical representations of the structure and variation of real speech, in order to provide speaker independence in speech recognition. Results from these studies indicate that knowledge of probabilities of occurrence of allophonic forms, co-occurrence of allophonic forms, and speaker pronunciation groups can be used to lower lexical entropy (i.e., improve predictive ability of lexical models), and possibly, therefore, achieve rapid initial adaptation to a new speaker as well as ongoing adaptation to a single speaker.

## INTRODUCTION

As the number of words in the lexicon grows, the speech recognition problem gets more difficult. In a similar way, as more possible pronunciations for each word are included in the lexicon, the recognition problem gets more difficult because there are more competing hypotheses and there can be more overlap between the representations of similar words.

One important goal of a lexical representation is to maximize coverage of the pronunciations the system will have to deal with, while minimizing overcoverage. Overcoverage adds unnecessary difficulty to the recognition problem. One way to maximize coverage while minimizing overcoverage is to explicitly represent all possible pronunciations of each vocabulary word as a network of allophones. An example of such a network is shown in figure 1 for the word "water". This network represents eight possible pronunciations, some of which are fairly common (e.g., [W AO DX ER]), and others somewhat rare (e.g., [W AA T AX]). Experience suggests that, to assure coverage, it will be necessary to include many pronunciations for each word, including those which happen relatively rarely.

In reality, speech is more highly organized. There is more predictive knowledge available than in a model that simply represents independent equiprobable choices with no interaction or influence between different parts of a model and with no ability to use information from other parts of an utterance or previous utterances by the current speaker. In current systems which use allophonic models, each node represents an independent set of equiprobable choices.

The goal of the research described in this paper is to explore ways in which a lexical representation can better reflect the structure of real speech data, so that the representation will have more predictive power, and thus improve recognition accuracy. A better understanding of the issues involved may lead to methods for rapid adaptation to a new speaker, as well as ongoing adaptation to a single speaker during a single session.

In order to explore these issues, we chose to model (as a single utterance) a pair of sentences containing 21 words for which we had a large data set. The patterns of variation found for this 21-word microcosm should indicate what kinds of structures will be needed in a larger lexicon. The data used were transcriptions of the two dialect sentences for the 630 speakers in the TI-AP database.

We have performed a series of four studies that explored four types of phonological structure, and ways of representing this structure in a lexicon. In the first study, we simply looked at the gain in predictive ability of a phonological model which incorporates knowledge of the probabilities of the various possible word pronunciations. The second study explored the co-occurrence of allophonic forms, and ways in which knowledge of these co-occurrences can be automatically computed into a phonological model. The third study explored the possibility of grouping speakers into a small number of pronunciation clusters, and looked for demographic and other predictors of these pronunciation clusters. The fourth study was designed to compare intra-speaker variation to the variation within the pronunciation clusters defined by the third study.

To evaluate our data, and compare representations, we used entropy as a measure of the predictive power of a representation, or difficulty of the recognition task given a particular representation. The entropy of a representation, developed from or "trained" on some large set of data, reflects both how well the representation captures significant structure in the data and how much predictive power is gained by modelling this structure.

The four studies are described in the following four sections, followed by a general discussion and conclusions.

## PRONUNCIATION PROBABILITIES

The goal of the first study was to determine how much speech recognition accuracy could be improved by incorporating knowledge of pronunciation probabilities into a phonological language model. An important goal of any lexical representation is to provide coverage of the pronunciations that the system will have to to deal with, including relatively rare pronunciations. This makes the recognition problem more difficult because there are more competing hypotheses and can be more overlap between word models. One way to deal with this problem is to include probabilities for pronunciations in the lexical model. In this way, including somewhat rare pronunciations will increase coverage without hurting performance. It will

[1] This research was sponsored by Defense Advanced Research Projects Agency Contract N00039-85-C-0302. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

allow recognition of these unusual pronunciations, avoiding confusion with other more common pronunciations of similar words. For example, consider the allophone string

|DH AX B IH G W AA DX AX B IH L Z|

This string contains, as a substring, the sequence |W AA DX AX|, which corresponds to one of the paths through the network for the word "water" shown in figure 1. This is a relatively infrequent pronunciation of the word "water". An alternative hypothesis for this same substring could be the pair of words "wad of", for which this pronunciation is relatively common. (Suggesting the phrase "The big wad of bills" rather than "The big water bills".) Appropriate probabilities associated with these pronunciations could allow a system to make a more intelligent choice. Such a model should help recognition accuracy significantly, provided that the probabilities used are accurate for the domain in which the system will be used, and especially if the probability distributions are significantly different from the default equi-probable models.

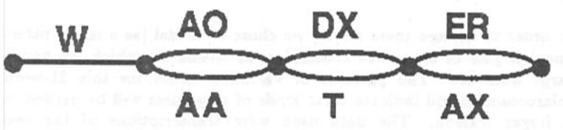


**Figure 1. Allophone network for the word "water".**

The data used in this study were transcriptions of the two dialect sentences for the 630 speakers in the TI-AP database. Originally, the allophonic forms used for each of 58 phonemes in the two test sentences as produced by the 630 speakers were transcribed by Margaret Kahn, Jared Bernstein, or Gay Baldwin. The transcriptions were done carefully using a high fidelity interactive waveform editor with a convenient means to mark and play regions in a high resolution image of the waveform. Spectrographic and other analytic displays were also easily available, though most of the work was done by ear and by visual inspection of the waveforms. Subsequently, a subset of 18 of these segments was chosen which we felt we could transcribe accurately and consistently; the 18 are disproportionately consonantal. This subset of 18 segments in the original 630 transcriptions were then re-checked and corrected by one individual (Gay Baldwin). The transcriptions of these 18 segments/utterance were compared to a subset of 156 speakers whose sentences had been independently transcribed at MIT as part of a related project. For this subset of 156 speakers, the number of transcription disagreements between SRI and MIT was about 5-10% for a typical phoneme.

Figure 2a shows the two dialect sentences, indicating the segments included in this study, along with the distributions of allophones found for each of these phonemes. Figure 2b shows the 18 node allophone model used to represent the possible pronunciations. Among these 18 phonemes, at the level of transcription we used, there are 14 two-way splits, two three-way splits, and a six and a seven-way split. The distributions vary from a 1%-99% split for canonical /t/ vs. flap in "water" to a 60%-40% split for the affricated vs. non-affricated /dy/ juncture in "had-your" to a 47%-53% split for a glottal gesture at the beginning of "oily".

The seven-way split for the juncture in "suit in" is the most unpredictable. The potentially variable events are the burst of the /t/, the occurrence of a glottal onset to "in" and the presence or absence of the vowel in "in". A third of the readers produced a very clear form that exhibited a t-burst and a glottal stop or glottalization at the onset of the /I/ in "in". A quarter of the readers flapped or produced a short /d/ into the vowel in "in". Nineteen percent of the utterances showed no burst for the /t/ but a glottal gesture into the /I/, while 18% showed the same burstless /t/ with glottal gesture, but released the gesture directly into the nasal, deleting the /I/. Three percent of the readers (18 people) released the /t/ with a burst right into the /I/, 3 speakers (0.6%) had a clear t-burst, but the glottal gesture goes right into the nasal with the /I/ deleted. One speaker (0.2% of the sample) produced the "suit in" juncture as a /d/ with a velic release into the nasal (as in a word like "sudden"). The distributions are surprising only as reminders of how little quantitative data on the relative frequency of occurrence of allophones is available. What experienced phonetician could have estimated the proportion of these forms in reading? It's no wonder that speech recognition lexicons would have whatever allophonic options they allow unspecified as to relative likelihood.

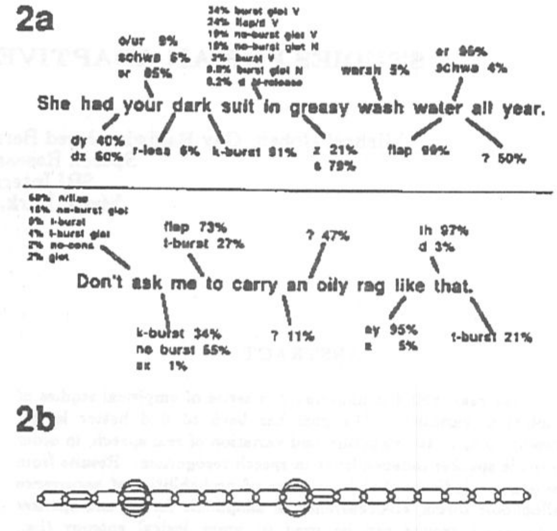


**Figure 2.**
**a) observed percentages of allophonic forms.**
**b) 18 node utterance model.**

The approach used in this study was to compute the probabilities of each of the transitions in the 18-node allophone model (figure 2b) from a large database of speech. The entropy of this model was computed, and compared to the entropy of a similar model without probabilities estimated from data, in which case all transitions from a node are considered equiprobable. Information theoretic entropy, H, of an arbitrary string, S, in the language was computed as:

$$H(S) = -\sum_n \sum_t P(t)\log_2 P(t) \tag{1}$$

where n ranged over all of the the nodes in the utterance model, t ranged over all of the transitions from the current node, and P(t) is the probability of transition t. This is the same as:

$$H(S) = -\sum_s P(s)\log_2 P(s) \tag{2}$$

where s ranges over all of the strings in the language [McEliece, 1977].

The entropy measured for the model with equi-probable transitions is 22.6 bits, and for the model with empirically estimated probabilities is 13.0 bits. This represents an increase of 42.4% in predictive ability (knowledge or source of constraint) for the model with trained probabilities. Presumably, this further constraint should translate into improved recognition accuracy.

## CO-OCCURRENCE OF ALLOPHONIC FORMS

The goal of the next study was to explore co-occurrence relationships in allophonic variation. A co-occurrence relationship is one in which the probability of the occurrence of a particular variant is conditioned on the presence or absence of some other variant in another part of the utterance. Knowledge of such co-occurrence relationships can be used to increase predictive power about allophonic variation.

The data used in this study was the same as that used in the previous study, except that the realization of /k/ in "dark" was excluded, since we decided we had insufficient confidence in our transcriptions of that phoneme. All possible pairings of the remaining 17 phonemes (136 pairs) were tested for co-occurrence relationships. The two examples in figure 3 demonstrate the technique. For each pair of segments, counts of all combinations of variants for the two forms were entered into a matrix. Chi-square tests were performed on these matrices at the 97.5% confidence level.

The example in figure 3a illustrates the analysis of glottal (or no glottal) at the beginning of "all" and "oily". The table shows that, of the 630 speakers, 230 used a glottal gesture (either a full glottal stop or a weaker gesture seen as several irregular glottal periods, both symbolized here as [ʔ]) at the beginning of both "all" and "oily". One hundred eighty-four speakers didn't use [ʔ] before either word, 65 speakers put [ʔ] just on "oily", and 151 just on "all". The chi-square is significant at the 97.5% level, indicating that this pattern of co-occurrence of glottals at the beginning of "all" and "oily" is rather unlikely to happen by chance if we assume that the two events are independent. In other words, speakers who used [ʔ] before "all" were more likely to use [ʔ] before "oily" as well. Similarly, if [ʔ] was omitted before "all", it was less likely to be found before "oily". This case of co-occurrence is not surprising, because both forms could be considered to result from the same phonological rule.

The co-occurrence matrix in figure 3b shows a dependent relationship between forms that are phonologically heterogeneous. In this case speakers who use [t] rather than flap in "to" show a strong tendency to use (rather than omit) [ʔ] before "an". This might be interpreted as evidence for a higher level fast-speech (or lax style) "macro-rule", which increases the likelihood of several types of phonological rules. One goal of our work is to establish a method by which such functional rule groups can be found (or dismissed). For now we just present preliminary data that show non-independence between pairs of forms over this sample of utterances.

Figure 4 shows which of the 136 possible co-occurrences actually had chi-squared values that indicated non-independence. The confidence level for the chi-squared value was 97.5%, meaning that of the 136 chi-squares calculated, one could expect about four artifactually non-independence between co-occurring forms. Of these 37, about 15 involve pairs that have a clear phonological relation, (r-lessness in "your", "dark", "water"; [ʔ] in "all", "an", "oily"; flapping in "water", "to", "don't ask", "suit in"; etc.). Most of the remainder show dependencies between variants in more remotely related phonological contexts. The number of dependencies is obviously considerable, and suggests that macro-level relationships — dialect region, utterance speed, style, sex-linked variation — are pervasive enough to be useful in improving predictions of forms for automatic speech recognition.

**3a**

| | | all-? | |
|---|---|---|---|
| | | 0 | ? |
| oily-? | ? | 65 | 230 |
| | 0 | 184 | 151 |

**3b**

| | | an-? | |
|---|---|---|---|
| | | ? | 0 |
| to-t | burst | 39 | 129 |
| | flap/d | 28 | 434 |

Figure 3. Co-occurrence Examples:
a) onsets in "all" and "oily".
b) onsets for "an" and "to".

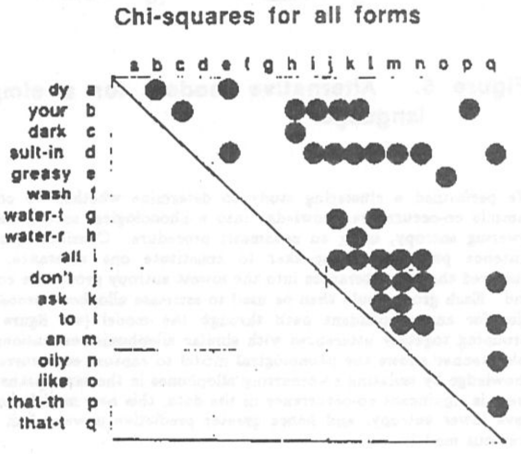


Figure 4. Significant Chi-squares for form pairs.

There could be significant advantage in finding a way to compile knowledge about the co-occurrence of allophonic forms into the phonological model. This would allow a form of within-utterance adaptation to take place automatically. An example of how this might be done is shown in figure 5. Figure 5a shows a probabilistic language model for a language consisting of strings of two symbols, the first symbol being "A" half the time and "B" the other half of the time, and the second symbol evenly divided between "C" and "D". There is additional structure to this language, in the form of co-occurrence. When the first symbol is "A", the probability is 90% that the second symbol will be "C", and 10% that it will be "D". When the first symbol is "B", the distribution is reversed. The entropy of such a model can be calculated as 1.47 bits. When a model representing the same language is configured as in figure 5b, without representation of the co-occurrence, the entropy is two bits, one bit for the choice at each node. Configuring the model to reflect co-occurrence knowledge has resulted in more than 25% lower entropy. Clearly, the model in 5a can do a better job of predicting incoming strings in the language than that in 5b.

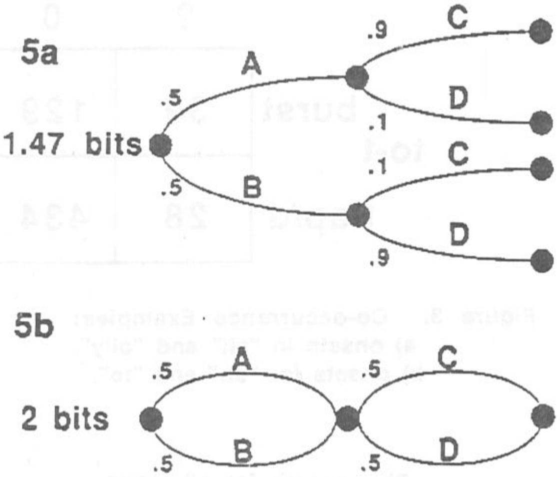


**Figure 5. Alternative models for a simple language.**

We performed a clustering study to determine whether we could compile co-occurrence knowledge into a phonological model, hence lowering entropy, using an automatic procedure. Considering each sentence pair from a speaker to constitute one utterance, we clustered the 630 utterances into the lowest entropy groups we could find. Each group could then be used to estimate allophone probabilities for an independent path through the model (see figure 6). Grouping together utterances with similar allophonic realizations in this manner allows the phonological model to capture co-occurrence knowledge by isolating co-occurring allophones in the same paths. If there is significant co-occurrence in the data, this new model should have lower entropy, and hence greater predictive power, than the previous model.

The clustering technique used was a combination of hierarchical clustering and the iterative Lloyd algorithm [Duda and Hart, 1973]. For each specific number of clusters desired, the data were clustered into that number of groups using an agglomerative hierarchical clustering technique, and then these clusters were used as the seeds to

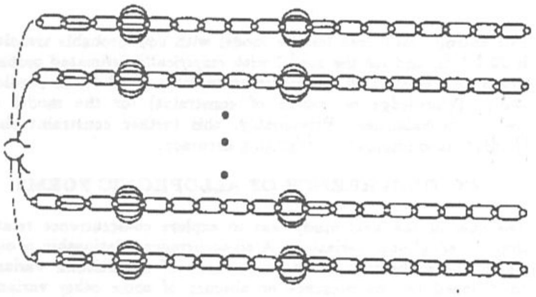


**Figure 6. Allophone network for the 2-sentence utterance, showing clusters as separate paths.**

the Lloyd algorithm. Each step of the hierarchical clustering algorithm involves merging the nearest pair of distinct clusters. Initially, each utterance forms a singleton cluster, and the procedure continues until the desired number of clusters is reached. At each step, the nearest pair of clusters was defined as that pair whose merging would result in a model with the lowest conditional entropy H(S|c), which was computed as:

$$H(S \mid c) = \frac{1}{N}\sum_{i=1}^{n} M(i) H(S \mid i) \qquad (3)$$

*where* $N$ = *total number of utterances in the sample* (630),

$n$ = *current number of clusters*,

$M(i)$ = *number of utterances in cluster* $i$, and

$H(S \mid i)$ = *entropy of a string* $S$ *in cluster* $i$.

Hence, H(S|c) is defined as the weighted average (weighted by cluster size) of the entropies of the individual clusters, which is the same as the entropy of a string, given that you know which cluster the string falls into. Though the real objective of this procedure was to minimize H(S) rather than H(S|c), computing H(S) for the composite model at each iteration of the algorithm is computationally too expensive. Though H(S|c) is not guaranteed to be monotonically related to H(S), it should be in most cases.

In the second phase of clustering, the clusters found by hierarchical clustering were used as a seed to the iterative Lloyd algorithm, which continued until the improvement for one iteration was less than a threshold. Each iteration of the Lloyd algorithm involved the following:

1) For each utterance: compute H(S|c), as in equation 3, with this utterance as a member of each current cluster - remember the cluster for which H(S|c) is minimal.

2) Once the new cluster is chosen for all utterances, actually make the switches.

Typically, the Lloyd algorithm continued for 5-10 iterations, and the amount of reduction in H(S) over the clusters output from the hierarchical clustering procedure was another 1-2% lower than the unclustered model.

The results of the clustering study are shown in figure 7. The higher curve H(S) is for the composite model given 10, 20, and 30 clusters. The results show that the entropy of a phonological model can be lowered 10-15% by modelling the co-occurrence of allophonic forms. Furthermore, this co-occurrence can be modelled for any sufficiently large data set by running a standard clustering algorithm, without the need to explicitly determine what the co-occurrences are. The significance of the lower curve (H(S|c)) will be discussed below in the section on speaker groups.

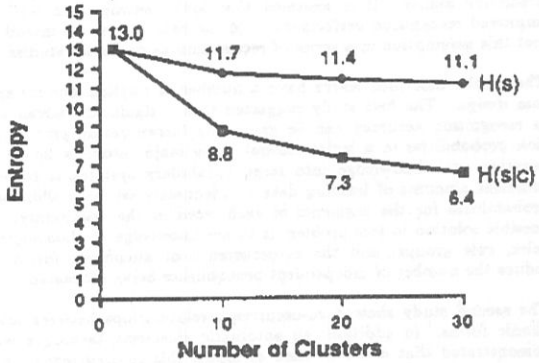


**Figure 7. Entropy of utterance model as a function of the number of clusters.**

We also tested whether demographic factors and speech rate could be used to predict allophonic forms. These results are shown in figure 8. Chi-squares (at the 97.5% confidence level) were computed to test for independence between region (each speaker was identified with one of seven geographic regions or as an "army brat"), age (by decade), race, sex, education (HS, BS, MS, or PhD), and speech rate, vs. form. As can be seen, the results show significant non-independence between all of the demographic factors vs. form and rate vs. form. This indicates that all of these factors are significant predictors of allophonic occurrences. For example, people from New England tend to say r-less "your", and people from the South tend to say "greasy" with a [z].

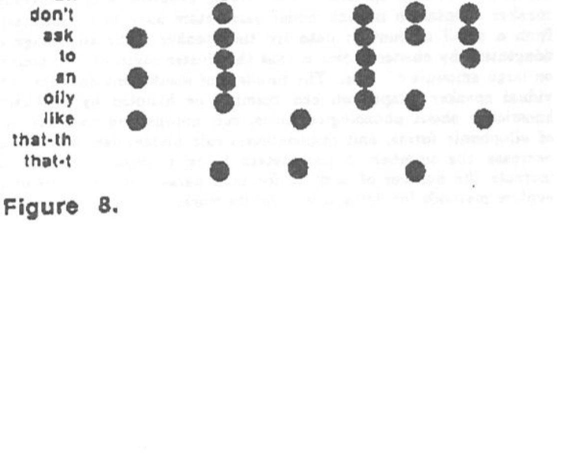


Figure 8.

## SPEAKER GROUPS

The lower curve in figure 7 shows the conditional entropy of the model given the cluster, computed as in equation 3. This result indicates that if the appropriate cluster for the incoming utterance is known in advance, entropy can be lowered 30-50% from the unclustered model. The question then arises of how well can we predict the cluster for an incoming utterance. This question, in turn, raises a number of additional questions:

1) What explicit predictors of cluster membership are available? (e.g., sex, region of origin, speech rate, etc.)

2) How consistently does a speaker stay within one cluster? (i.e., If a speaker stays in the same cluster with reasonable consistency, then rapid adaptation to a new speaker may be accomplished by choosing the appropriate cluster after some experience with this speaker, or choosing an appropriate weighting function over the clusters.)

3) How can we classify a speaker into the appropriate cluster, or choose the appropriate weighting function over clusters for this speaker at the current time?

4) When during a recognition session should a new cluster be chosen, or a new weighting function be computed? (e.g., when speech rate changes, when performance drops, only when a new speaker comes along, etc.)

The studies described in this section were designed to address the first two questions.

In order to test for predictors of cluster membership, we performed chi-squares at the 97.5% confidence level, testing for non-independence between cluster vs. form, cluster vs. all of our demographic factors (age, race, region, sex, and education), and cluster vs. speech rate. There was significant non-independence between cluster and all allophonic forms except for the /t/ in "water", as well as for all demographic factors and rate. The lack of significance for /t/ in "water" is not surprising since, out of our sample of 630 speakers, only five of them aspirated the /t/.

In order to test the consistency with which speakers remain in clusters, we gathered a new set of data, consisting of speakers repeating the same sentences many times. Four speakers were recorded in three sessions each, with recording sessions for the same speaker a week apart. The recordings were made in a sound-treated room, using a close talking microphone and a Nagra tape recorder. Each recording session consisted of eight readings of the same two sentences used in the experiments described earlier, interspersed in a set of seven filler sentences. The first five repetitions were uninstructed (i.e., "normal reading"). At the sixth repetition, the subjects were instructed to read very quickly, at the seventh slowly and carefully, and at the eighth normally. From listening to the recordings, it is our judgement that the fast readings were, indeed, extremely fast, and the slow and careful readings were extremely slow and careful. Since the uninstructed readings were fairly fast, the differences between the slow and uninstructed readings were more dramatic than those between the fast and uninstructed readings. The final data set consists of 96 repetitions of the two sentences, 24 from each speaker, with 72 repetitions uninstructed or "normal", 12 fast, and 12 slow and careful.

The same 18 phonemes used in the earlier experiments were phonetically transcribed, with the aid of the tools described earlier, by Michael Cohen, and checked by Jared Bernstein and Gay Baldwin. Each of the 96 utterances were then classified into the clusters based on the 630-speaker data, as described in the previous section. We chose to classify them into the 10-cluster version so that each cluster would be based on a large number of utterances (approximately 63). Each utterance was classified into the cluster with the centroid with minimal Euclidean distance to the utterance. Table 1 shows the number of utterances for each speaker classified into each cluster. As can be seen, most of the utterances for each speaker tend to be classified into two or three clusters. Eleven of the 12 slow utterances were classified into cluster two. The fast utterances did not tend to fall into any one cluster.

**Table 1. Classification of speaker utterances according to pre-existing clusters**

| Speakers | Clusters | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| JB | 0 | 0 | 3 | 0 | 1 | 2 | 3 | 1 | 0 | 14 |
| JK | 0 | 2 | 6 | 1 | 5 | 0 | 0 | 1 | 9 | 0 |
| KC | 0 | 1 | 2 | 1 | 2 | 0 | 1 | 0 | 0 | 17 |
| PR | 0 | 11 | 8 | 0 | 0 | 0 | 3 | 2 | 0 | 0 |

These results indicate that although speakers fall into the same clusters with some consistency, choosing a single cluster for a speaker is inadequate. A more reasonable approach may be to choose a weighting function over all of the clusters. Furthermore, cluster membership seems to be somewhat dependent on speech rate.

## INTRA-SPEAKER VS. INTRA-GROUP VARIATION

The results of the previous section suggest a method of adaptation by choosing appropriate sets of (or weights for) clusters for a speaker. The study described in this section addresses the question of whether or not it is useful to try to further adapt to the individual speaker once the clusters are chosen. We have addressed that question by comparing the amount of variation within a single speaker to the amount of variation within a single cluster. If there is considerably less variation within a speaker than within even a single cluster, then there may be ways to further adapt to the individual speaker.

The data used for this experiment included both the 630-speaker data described earlier, and the four speaker multi-repetition data described in the previous section. We compared the entropy of a model trained for a single speaker in the multi-repetition data set to the entropy of a cluster from the 630-speaker set. Only the 18 uninstructed utterances for each speaker were used from the multi-repetition data, because the 630-speaker data were recorded without instruction. The comparison was made with the 10-cluster version of the 630-speaker data so that each cluster would be based on an adequate amount of data. In order to be able to make a fair comparison, it was necessary to compare the entropy of models trained on the same number of speakers, so we sampled the large clusters from the 630 speaker set by randomly choosing a cluster, and then randomly choosing the appropriate number of speakers from the cluster. This was done 1000 times, and the mean entropy of the 18-member clusters were computed. The mean entropy of the 18-member clusters from the 630-speaker data was 8.39, and for a single speaker from the multi-repetition data was 6.86, approximately 18% lower. This suggests the possibility of significant individual speaker adaptation beyond the choosing of appropriate clusters.

## DISCUSSION

The studies described in the previous four sections have demonstrated some types of structure in the phonological variation observed in a data set consisting of two sentences (21 words) read by many speakers. In addition, we have shown some types of lexical representations that might be used to capture this structure. Representations were compared by measuring their entropy, or predictive ability. It is assumed that lower entropy can lead to improved recognition performance. In the near future, we intend to test this assumption in a series of recognition performance studies.

The results described above have a number of implications for system design. The first study suggested that a significant advantage in recognition accuracy can be gained by incorporating pronunciation probabilities in a lexical model. The major problem in incorporating such knowledge into large vocabulary systems is finding sufficient amounts of training data to adequately estimate allophone probabilities for the segments of each word in the vocabulary. A possible solution to this problem is to use knowledge of phonological rules, rule groups, and the co-occurrence of allophonic forms to reduce the number of independent probabilities being estimated.

The second study showed co-occurrence relationships between allophonic forms. In addition, an automatic clustering technique was demonstrated that could be used to model this co-occurrence for a data set without explicit knowledge of what these co-occurrences are. This result suggests that lexical representations can be improved by including a small number of sets of word models, each trained on an appropriate cluster of a large data set. When scoring sequences of word pronunciation hypotheses for an utterance, each sequence would only include one set of word model probabilities.

The last two studies suggest methods of adaptation to a new speaker, as well as ongoing adaptation within a session with a single speaker. In figure 7, H(S|c) is shown to be considerably lower than H(S). This suggests that predicting the appropriate cluster for an utterance can reduce entropy considerably by allowing the search to be confined to the model of a single cluster.

The third study, which explored the consistency with which a speaker remains in a cluster, suggests that predicting the cluster for an utterance cannot be achieved solely by speaker adaptation, since a speaker will not stay in a single cluster consistently. However, the third study does suggest that H(S|c) can be approached by choosing an appropriate weighting function over all the clusters, given some experience with a speaker. Furthermore, these results suggest that knowledge of speech rate can be used to improve prediction of the appropriate cluster for an utterance. Ongoing adaptation might be achieved by periodically recomputing the weighting function. We have not explored the question of when, or how often, should this weighting function be recomputed.

The results of the fourth study, comparing intra-speaker to intra-cluster entropy, show greater consistency within a single speaker than within the clusters found in the previous studies. This suggests that speaker adaptation can be improved beyond the choice of clusters by further refinement of model parameters, based on extended experience with a speaker. The major problem with individual speaker adaptation is that model parameters have to be estimated from a small amount of data for the speaker. The advantage of adaptation by cluster choice is that the cluster could be well trained on large amounts of data. The problem of insufficient data for individual speaker adaptation can possibly be handled by exploiting knowledge about phonological rules, rule groups, the co-occurrence of allophonic forms, and implicational rule hierarchies, in order to decrease the number of parameters being estimated, as well as increase the number of samples for each parameter. We intend to explore methods for doing this in future work.

## CONCLUSIONS

We have performed a series of four studies, with the following results:

1) Incorporating empirically determined probabilities of allophonic forms into a phonological model can significantly reduce model entropy, and possibly improve recognition accuracy.

2) There is significant co-occurrence of allophonic forms within an utterance, and automatic clustering procedures can be used to compile knowledge of these co-occurrences into a phonological model, without need to explicitly determine what the co-occurrences are. Incorporating these co-occurrences into the phonological model can significantly lower entropy and allow a form of within-utterance adaptation, possibly improving recognition accuracy.

3) Speakers tend to fall into phonological groups. Rapid adaptation techniques might work by choosing either a set of clusters or weighting function over all clusters for a speaker given a small amount of experience with that speaker. Ongoing adaptation may possibly be achieved by periodically rechoosing a cluster set or recomputing the weighting function.

4) Individual speakers vary less than speaker clusters, and therefore, further adaptation to an individual speaker could be useful. This may require the exploitation of knowledge about phonological rules, rule groups, implicational rule hierarchies, and the co-occurrence of allophonic forms.

## REFERENCES

[1] Duda, R. and Hart, P., "Pattern Classification and Scene Analysis", John Wiley & Sons, 1973, pp. 225-237

[2] McEliece, R., "The Theory of Information and Coding: A Mathematical Framework for Communication" in "Encyclopedia of Mathematics and its Applications, Volume 3", Rota, G., ed, Addison-Wesley, 1977, pp. 15-34.

# 7 References

Atwood, E. Bagby (1980), *The Regional Vocabulary of Texas*, The University of Texas Press, Austin.

Bailey, Richard W. and Robinson, Jay L. (1973), *Varieties of Present-Day English*, MacMillan.

Bronstein, Arthur J. (1960), *The Pronunciation of American English*, Appleton-Century-Crofts, New York.

Davis, Lawrence M. (1983), *English Dialectology: An Introduction*, The University of Alabama Press.

Department of Linguistics, The Ohio State University (1982), "Language Files," Advocate Publishing Group, Reynoldsburg, Ohio.

Fisher, William M., Zue, Victor, Bernstein, Jared, and Pallett, David S. (1987), "An acoustic-phonetic data base," presented at the 113th meeting of the Acoustical Society of America, *J. Acoust. Soc. Am., Suppl. 1, Vol. 81.*

Gove, Philip B., Ed. (1966), *Webster's Third New International Dictionary, Unabridged* William Benton, Publ., Encyclopedia Britannica, Inc., Chicago.

Hultzen, Lee S., Allen, Joseph H.D., Jr., and Miron, Murray S. (1964), *Tables of Transitional Frequencies of English Phonemes*, University of Illinois Press, Urbana.

ISO-9660 (1988), *Information Processing -- Volume and file structure of CD-ROM for information interchange*, Standard by the International Organization for Standardization.

Kassel, Robert H. (1986), *Aids for the Design, Acquisition, and Use of Large Speech Databases*, S.B. Thesis, MIT.

Kenyon, John S. and Knott, Thomas A. (1953), *A Pronouncing Dictionary of American English*, Merriam-Webster Inc., Springfield, Massachusetts.

Kuchera, Henry and Francis, W. Nelson (1967), *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island.

Kurath, Hans (1949), *A Word Geography of the Eastern United States*, The University of Michigan, Ann Arbor.

Pallett, D. S., Fisher, W. M., and Fiscus, J.G. (1990), "Tools for the Analysis of Benchmark Speech Recognition Tests," *Proc. ICASSP-90.*

Williamson, Juanita V. and Burke, Virginia M., eds. (1971), *A Various Language: Perspectives on American Dialects*, Holt, Rinehart, and Winston, New York.

Zue, Victor W, Cyphers, D. Scott, Kassel, Robert H., Kaufman, David H., Leung, Hong C., Randolph, Mark, Seneff, Stephanie, Unverferth, John E. III, and Wilson, Timothy (1986), "The Development of the MIT Lisp-Machine Based Speech Research Workstation," *Proc. ICASSP-86.*

★U S GOVERNMENT PRINTING OFFICE 1993-341-931/ 82624