A TELEPHONE SPEECH DATABASE OF SPELLED AND SPOKEN NAMES

Ronald Cole, Krist Roginski and Mark Fanty

Center for Spoken Language Understanding
Oregon Graduate Institute
Beaverton, Oregon, 97006

ABSTRACT

This report describes a telephone speech corpus collected at the Oregon Graduate Institute's Center for Spoken Language Understanding. Over four thousand people called in response to public requests. They were prompted by a recorded voice to say and spell their first and last names—with and without pauses, to say what city they grew up in and what city they were calling from, and to answer two yes/no questions. In order to collect sufficient instances of each letter, about 1000 callers also recited the alphabet. Each call is checked and transcribed by two people. In addition, a subset of the calls is being phonetically labeled.

INTRODUCTION

A successful telephone speech recognition system must deal with variations among speakers, telephone handsets, and channel characteristics. Modeling these sources of variability requires a corpus of speech samples provided by many speakers calling from different locations using different telephones.

We describe the status of a telephone speech corpus consisting of spoken letters and names provided by over 4000 speakers. This report describes procedures used to acquire the corpus, procedures used to transcribe and evaluate utterances, and the current status of the speech data.

The processed data will be made available to Universities for a nominal charge as it becomes available.

DATA ACQUISITION

A press release describing our research project and the need for volunteers produced newspaper, radio and television coverage. In addition, we posted requests for callers on several university bulletin boards and national computer newsgroups.

Speech was collected using a Gradient Technology Desklab connected via the SCSI port to a Sun 4/110 workstation. The device was programmed to answer

the phone, play digitized files requesting the speech samples, and digitize the callers' response for a designated period of time. Speech was sampled at 8000 samples per second at 14 bit resolution.

Protocol for First 3000 Calls

The caller heard the following instructions:

> Thank you for calling the OGI speech research laboratory. We are developing a computer system to recognize spelled names. To do this, we need to record samples of speech from many speakers. We will ask you to say and spell your last name, your first name, and to say the city and state you grew up in. The rest of the call will take about one minute. Please wait for the beep before speaking.
>
> - What city are you calling from?
> - What is your last name?
> - Please spell your last name.
> - Please spell your last name, with short pauses between letters.
> - Does your last name contain the letter A as in apple?
> - What is your first name?
> - Please spell your first name, with short pauses between letters.
> - What city and state did you grow up in?
> - Would you like to receive more information about the results of this project?
>
> Thank you for calling. We appreciate your help. If you would like to receive information about the results of this project, please leave your name and address at the tone.

Protocol for Calls 3000–4000

After 3000 speakers were recorded, the protocol was changed to guarantee more instances of each letter by asking speakers to recite the English alphabet. In addition, three additional words were added to the protocol. The new protocol consisted of the following instructions:

> Thank you for calling the OGI speech research laboratory. We are developing a computer system to recognize spelled names. To do this, we need to record samples of speech from many speakers. We will ask you to spell your last name, say where you are calling from and where you grew up, and to say the alphabet. Please wait for the beep before speaking.

- What city are you calling from?
- What is your last name?
- Please spell your last name.
- What city and state did you grow up in?
- Please say "apostrophe"
- Please say "capital"
- Please say "hyphen"

We will now ask you to say the alphabet. We need you to pause briefly between letters, like this: A B C D E F G. You may hang up when you are finished. Please begin speaking now.

TRANSCRIPTION AND EVALUATION OF UTTERANCES

Verification of the corpus consists of (a) listening to each utterance, (b) providing a precise transcription of what was said, and (c) making several judgements about the utterances. Transciption conventions provide additional information about the utterance. For example, when asked to say their last name, many callers also spell it. We distinguish between direct answers to the query and additional comments by bracketing the latter.

Each utterance is transcribed and evaluated by two different listeners. The procedure uses the LYRE interactive graphics program [1] that allows the transcriber to play selected portions of an utterance, and access other utterances produced by the same speaker. For example, if the caller's first or last name is ambiguous, the transcriber can listen to the spelled version of the same name before transcribing it. The two transcriptions of each utterance are compared automatically, and all differences are noted and resolved.

The transcription conventions include the following

- lip smacks are transcribed as "[ls]"
- breath noise is transcribed as "[br]"
- line noise is transcribed as "[ln]"
- filled pauses are transcribed as "[uh]"
- background noise is transcribed as "[bn]"
- coughs are transcribed as "[cough]"
- speech which does not directly answer the question is placed in brackets
- words which were cut off have an "-" appended (e.g. "my na-")

Transcription and evaluation is proceeding speaker by speaker, in the order in which the calls were received. The following procedure is used to process the calls:

1. The utterance is optionally chopped to remove lengthy non-speech intervals before and after the utterance. At least 100 msec is retained before and after the utterance.

2. The utterance is transcribed orthographically to indicate all speech events (including breath noise).

3. A set of judgements are made about each utterance. The listener notes the occurrence of any of the following: (a) extraneous speech; (b) environmental noise; (c) excessive breath noise; (d) did not follow instructions; (e) line noise; and (f) fluent spelling (no pause between any two letters during spelling, as seen on the waveform).

4. After individual utterances are processed, the following "global" judgments are made: (a) gender (male, female, unknown); (b) age (child, adult); (c) connection quality (poor, typical); (d) accent (not noticeable, South, East, other U.S., foreign); and (e) intelligibility (poor, typical).

5. Finally, a set of automatic measurements are performed on each utterance. These include its duration, the min and max sample, mean, and 10th and 90th percentile in dB measured over 10 msec windows in the utterance.

PHONETIC LABELING OF THE UTTERANCES

A subset of the utterances are being phonetically labelled by hand. As of this writing, these include 463 responses to the "what city are you calling from" question, 1305 responses to the "what city and state did you grow up in" and 101 responses to the "what is your last name" question. The set of phonemic labels is similar to that used with the TIMIT corpus [2,3], with the following distribution.

#h	3589	n	1923	pau	1687
ah	1434	l	1212	s	1134
ih	1051	k	956	tcl	837
ao-r	763	t	750	ae	696
kcl	682	eh	668	iy	667
q	579	ow	577	m	568
d	529	dcl	484	ao	464
ax	453	er	441	b	440
uw	435	r	414	y	412
g	400	w	393	v	385
gcl	355	ls	341	p	334
ix	313	f	303	aa	274
ey	269	sh	266	br	236
bcl	219	ng	217	ay	205
hh	191	z	183	glot	175
dx	165	nx	156	pcl	153
gx	143	ln	123	el	122
ch	110	jh	106	uh	79
th	71	cl	61	oy	61
pv	51	ax-h	49	unk	37
en	36	aw	27	dh	20
hv	17	ns	13	zh	2

Several labels were added to the standard set:

ao-r	r-sound in Portland, York, California
ls	lip smack
ln	line noise
br	breath noise
glot	glottalization
pv	prevoicing
ns	non-speech sound

The utterances are labeled, then printouts of the labels, time aligned with the speech spectrogram, are checked by an expert spectrogram reader to catch obvious mistakes.

CURRENT STATUS

There are currently 2,542 completed calls (1268 female, 1238 male and 22 unknown gender). There are a total of 2,500 last names (1,902 different last names); 1,495 first names (651 different first names); 886 different hometown cities; and 423 different "callfrom" cities. Some corpus statistics are provided below:

STATISTICS

Ten Most Common ...

First names		Last Names	
John	44	Smith	24
David	36	Johnson	21
James	21	Brown	12
Michael	20	Miller	11
Robert	20	White	10
Mary	18	Young	9
Steve	16	Jones	9
Barbara	14	Bell	9
Thomas	14	Anderson	8
Mark	14	Clark	8

City called from		Hometown	
Portland	475	Portland	199
Seattle	431	Seattle	160
Beaverton	74	Chicago	35
Bellevue	49	Vancouver	27
Denver	44	Los Angeles	25
Salem	42	New York	25
Vancouver	40	Denver	25
Pittsburgh	37	Minneapolis	22
Boston	37	Boston	18
Hillsboro	35	Salem	17

Responses of Yes/No Questions

yes	1416
no	1006
yes i would	76
sure	51
no thankyou	49
yes please	36
yes it does	35
no it does not	15
yeah	9
yes thank you	5

Distribution of Letters in Last Names

e	1023	h	402	w	172
r	808	m	319	p	171
a	787	c	308	y	169
n	738	d	285	f	119
o	673	b	254	v	72
s	633	k	226	j	61
l	606	u	215	z	51
i	487	g	184	x	22
t	480			q	9

ACKNOWLEDGEMENTS

Research supported by U.S. WEST, NSF and ONR.

BIBLIOGRAPHY

[1] Fanty, M., Pochmara, J. and R. Cole, "An interactive environment for speech recognition research," ICSLP 92.

[2] Lamel, L., Kassel, R. and S. Seneff, ""Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus", Proceedings of the DARPA Speech Recognition Workshop, pp. 100–110, 1986.

[3] Fisher, W., Doddington, G., and K. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specification and Status.", Proceedings of the DARPA Speech Recognition Workshop, pp. 93–100, 1986.