THE OGI MULTI-LANGUAGE TELEPHONE SPEECH CORPUS

Yeshwant K. Muthusamy, Ronald A. Cole and Beatrice T. Oshika †

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
19600 NW Von Neumann Drive, Beaverton, OR 97006-1999

† Department of Applied Linguistics
Portland State University
PO Box 751, Portland, OR 97207-0751

ABSTRACT

The OGI Multi-language Telephone Speech Corpus is designed to support research on automatic language identification and multi-language speech recognition. The corpus consists of up to nine separate responses from each caller, ranging from single words to short topic-specific descriptions to 60 seconds of unconstrained spontaneous speech. The utterances were spoken over commercial telephone lines by speakers of English, Farsi (Persian), French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese. We have completed the initial phase of our data acquisition effort: the recording and initial verification of utterances produced by 100 different speakers in each of the 10 languages. We describe the recording protocol, data collection procedure, ongoing corpus development, preliminary results of the statistical evaluation of the 10 languages, and plans to provide orthographic transcriptions of the speech.

INTRODUCTION

Research in multi-language recognition systems would be enhanced by the availability of public-domain, multi-language corpora that can be used to study languages and to develop, evaluate and compare multi-language recognition algorithms. Applications include automatic language identification, multi-language speech recognition, word-spotting, and speech-to-speech automatic language translation.

In this report, we describe the OGI Multi-language Telephone Speech Corpus designed to support research in these areas. The initial phase of the data acquisition effort, completed in June 1992, involved the recording and preliminary verification of utterances produced by 100 different speakers in 10 different languages. The second phase of the effort, now underway, involves orthographic transcriptions and several subjective judgments of each call made by native speakers of the individual languages.

CHOICE OF LANGUAGES

The languages currently in the corpus, English, Farsi (Persian), French, German, Korean, Japanese, Mandarin Chinese, Spanish, Tamil and Vietnamese, were selected based on a combination of linguistic considerations and the availability of native speakers in the United States.

These languages represent a range of unrelated languages (e.g., Vietnamese and Tamil and German) as well as languages from the same sub-family (e.g., Germanic languages such as English and German, Romance languages such as French and Spanish). The languages also include various prosodic features, e.g., Mandarin Chinese and Vietnamese are tonal languages, Japanese uses pitch-accents and syllabic mora. In addition to their linguistic characteristics, the languages represent important geographic and political regions, and many speakers of these languages can be found relatively easily in the U.S.

The speech corpus was originally collected to support research on automatic language identification. Since most approaches to language identification rely on discriminators based on patterns of sounds and sound classes, it is important that the corpus include pairs of languages that are phonologically similar and others that are quite distinct. For example, syllable patterns of Vietnamese and Chinese are similar, basically consonant-vowel (CV) or consonant-vowel-consonant (CVC) patterns, with a relatively limited consonant repertoire, and with a characteristic tonal contour associated with each syllable. In contrast, German and English have relatively elaborated syllable structures, potential clusters of half a dozen consonants between vowel nuclei, and no distinctive tonal contrasts at the syllable level. From the point of view of automatic language identification, Chinese and Vietnamese should be more confusable based on phonological sequences, and Chinese and German should be less confusable.

DATA ACQUISITION

Collection Campaign. Speaker participation was promoted under a "donate your voice to science" theme. Requests for callers were posted on several university bulletin boards and national computer network newsgroups. In addition, a press release describing the research project and the need for volunteers resulted in newspaper and radio coverage.

Equipment. Speech was collected using a Gradient Technology Desklab connected via a SCSI port to a Sun 4/110 workstation. The device was programmed to answer the telephone, play digitized files in each of the 10 languages 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.

Recording Protocol. The recording protocol was designed to obtain (a) fixed-vocabularies, (b) short topic-specific descriptions, and (c) samples of elicited free speech.

The fixed vocabularies were collected in response to the following prompts. The time allocated for each response is shown in parentheses:

1. What is your native language? (3 s)
2. What language do you speak most of the time? (3 s)
3. Please recite the seven days of the week. (8 s)
4. Please say the numbers zero through ten. (10 s)

The topic-specific descriptions were obtained in response to the following prompts:

1. Tell us something that you like about your hometown. (10 s)
2. Tell us about the climate in your hometown. (10 s)
3. Describe the room that you are calling from. (12 s)
4. Describe your most recent meal. (10 s)

Elicited free speech was obtained by asking callers to speak for 1 minute on any topic of their choice. They were given 10 seconds to organize their thoughts before the actual 1 minute recording, to minimize the number of long pauses and false starts in the free speech. The duration of each call was approximately 5 minutes, and resulted in a maximum of 126 seconds of speech.

Call Format. Callers received a brief greeting in English followed by a prompt, in each language, to select a language by pressing a digit from 0 through 9. All subsequent instructions and prompts were given in the target language. This procedure helped reduce the number of crank calls by non-native speakers.

CORPUS DEVELOPMENT

Development of this corpus has been divided into two phases. Phase I, which has been completed, consists of (a) **preliminary verification**: listening to each utterance and deleting prank or invalid calls (hangups); (b) **chopping**: removing excess noise at the beginning and end of each utterance; (c) **evaluation**: making several judgments about the quality and type of speech; and (d) **broad phonetic transcriptions**: providing time-aligned broad phonetic labels to a subset of the utterances.

Phase II, which has just begun, involves (a) **verification and evaluation** of the utterances by native speakers of the individual languages, (b) **orthographic transcriptions** of each utterance, and (c) **time-aligned fine phonetic transcriptions** (automatically generated from dictionary pronunciations) of a subset of the utterances for each language.

Phase I

Phase I tasks were carried out by trained laboratory assistants who are native speakers of English. An interactive graphics program was used to display the waveform, play selected portions of the utterance, and to log information into a text file[1].

Preliminary Verification and Evaluation. Each utterance was processed as follows:

- The utterance was chopped, if necessary, to remove the excess noise and/or silence before and after the speech. Care was taken to include at least 300 ms of "silence" before and after the speech. Audible lip-smacks and breath noise were always retained.

- Judgments were made about the quality and content of speech in each utterance. The listener noted the occurrence of any of the following: (i) American or British accents (applicable to English calls only); (ii) excessive breath noise; (iii) speech cut off at the beginning; (iv) speech cut off at the end; (v) environmental noise; (vi) caller did not follow instructions; (vii) caller not a native speaker; (viii) read speech; (ix) spontaneous speech; (x) extraneous speech; and (xi) speech in non-native language

- A set of automatic measurements was made on the utterance. These include its duration, the minimum and maximum sample values, the dc offset, and 10th and 90th percentile of the power (in dB) measured over 10 ms windows in the utterance.

The "caller not a native speaker" comment for languages other than English was made only if the speaker admitted to being a non-native speaker in response to the "native language" question. A more accurate determination of the number of non-native speakers in other languages will be made during Phase II. "Extraneous speech" refers to background speech produced by someone other than the caller.

[1] The speech tools used in the development of this corpus are described in detail in the Proceedings of this Conference in [1].

The laboratory assistants were trained to recognize the fixed vocabularies in each language and were able to detect incomplete responses and non-standard pronunciations of the days-of-the-week and the numbers.

In addition to these utterance-specific comments and measurements, the following ”global” judgments were made after listening to all utterances of a call: (i) gender (male, female and unknown); (ii) age (child, adult); (iii) connection quality (poor, average, good); and (iv) speaker intelligibility (poor, typical).

Broad Phonetic Transcriptions. As a first step in the acoustic-phonetic analysis of the languages, we have provided time-aligned broad phonetic transcriptions to selected utterances in each language. The speech is automatically segmented into 7 broad phonetic categories [2]: (i) vowels, (ii) fricatives, (iii) stops, (iv) closures (silence or background noise), (v) pre-vocalic sonorant, (vi) inter-vocalic sonorant, and (vii) post-vocalic sonorant. The segmenter output is then corrected by trained transcribers using our speech tools.

Phase II

Phase II of the development involves (a) verification of the consistency and fluency of the speech data and (b) orthographic and fine-phonetic transcriptions by native speakers.

Verification of the utterances by native speakers of the individual languages includes (a) confirming that each utterance was in fact spoken by a native speaker of that language, (b) verifying that the caller followed the instructions for that utterance, and (c) preliminary judgment of accents and dialects and influence of other languages, e.g., effect of English on speakers of other languages who may have been in the U.S. for a long time.

Orthographic transcriptions allow access to the database at the lexical and sentence levels, and make the corpus useful for the natural language community. Standard romanizations will be used initially for Chinese, Farsi, Japanese, Korean and Tamil. Transcriptions using original alphabets and character sets might be made if resources become available.

Eventually, more detailed phonetic transcriptions will allow us to pursue a fine-phonetic approach to automatic language identification.

We have begun orthographic transcriptions in English, German and Spanish, and will add languages as resources permit.

CURRENT STATUS

To date, we have received a total of 2485 calls. Of these, 1042 calls are in English, with an average of 144 calls in the remaining 9 languages. On the average, 22.0% of the calls have been rejected in each language, mainly because of hangups. A total of 1345 calls (29.5 hours of speech), 246 in English, and an average of 122 calls in the remaining 9 languages, have been judged as useful after our chopping and evaluation. Table 1 displays the distribution of raw calls, the number of usable calls that resulted from those raw calls, and the average amount of speech (in seconds) per caller, for all the 10 languages. Note that each call can produce a maximum of 126 seconds of speech. The lower numbers in column 4 of Table 1 represent the amount of speech actually obtained before the caller decided to terminate the call.

Broad phonetic transcriptions have been provided to 2 utterances per call for the first 25 valid calls in each language (total of 500 utterances).

Table 1: Distribution of Calls across 10 Languages

Language	*Raw Calls*	*Usable Calls*	*Avg Secs/Call*
English	299	246	89.70
Farsi	153	114	79.56
French	149	123	85.38
German	157	118	87.69
Japanese	147	107	79.47
Korean	148	111	71.12
Mandarin	174	133	66.44
Spanish	149	117	86.04
Tamil	188	150	67.84
Vietnamese	158	126	69.80

Table 2: Distribution of Calls by Gender and Age Judgments

Language	*Calls*	*Males*	*Females*	Gender?	*Adults*	*Children*
English	246	174	69	3	243	3
Farsi	114	91	22	1	114	0
French	123	88	35	0	121	2
German	118	73	45	0	118	0
Japanese	107	66	39	2	107	0
Korean	111	81	29	1	110	1
Mandarin	133	85	44	4	130	3
Spanish	117	74	41	1	116	1
Tamil	150	127	21	2	150	0
Vietnamese	126	81	45	0	126	0

CORPUS STATISTICS

Speaker Statistics

The distribution of calls by age and gender judgments is provided in Table 2. The ratio of male to female speakers was roughly 7:3 over all the 10 languages, and ranged from 1.6:1 for German to 6:1 for Tamil.

Utterance Statistics

- 22.9% of the elicited free speech utterances in English were judged to contain read speech

- Figures 1 and 2 show the average frequency of occurrence (per second of speech) of inter-vocalic sonorants and fricatives respectively, for the first 50 calls in each language.

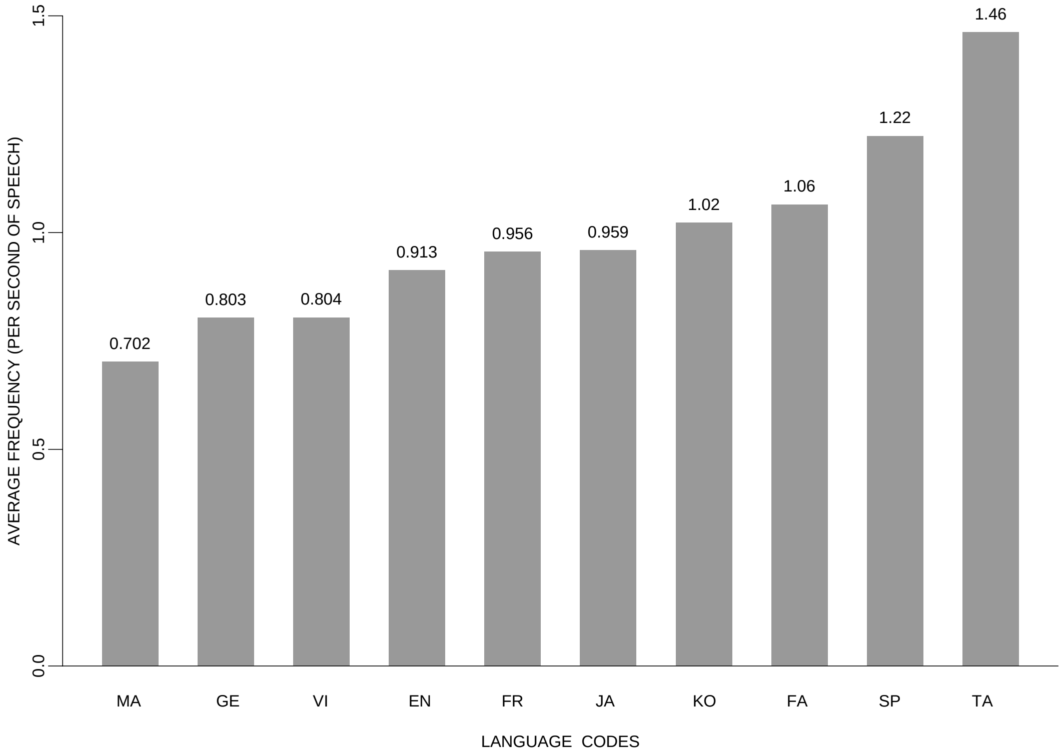

Figure 1: Average Frequency of Occurrence of Inter-vocalic Sonorants for 10 Languages

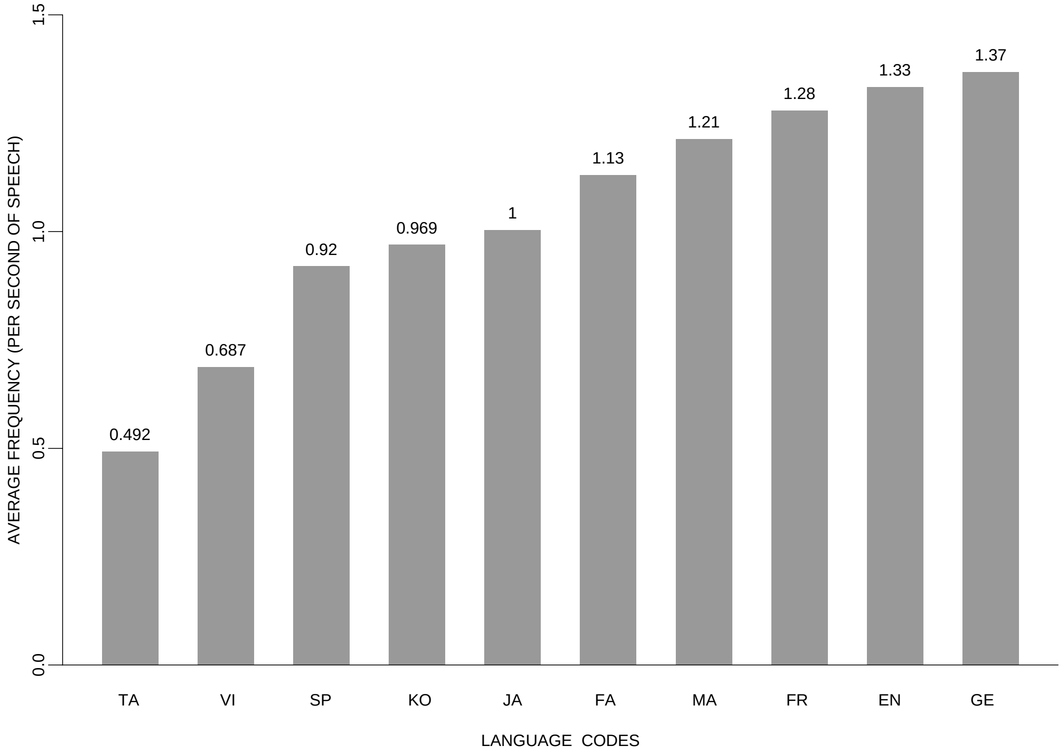

Figure 2: Average Frequency of Occurrence of Fricatives for 10 Languages

- The average speech rate (number of broad phonetic category segments ÷ utterance duration) for the first 50 calls in each language ranged from 8.02 segments/second for Vietnamese, to 9.56 segments/second for English, with a median of 8.98 segments/second.

USAGE AND DISTRIBUTION

Like the TIMIT corpus for continuous speech, we envision the OGI Multi-language Telephone Speech Corpus being used to study language differences and to develop and compare automatic language identification and other multi-language recognition algorithms. Also, smaller subsets of the corpus can be used for several well-defined tasks, such as multi-language digit recognition, recognition of language names, and multi-language days-of-the-week recognition.

To obtain a copy of this corpus, contact the second author.

ACKNOWLEDGEMENTS

The authors are grateful to Kelly Atkinson, Su Ge, Duc Ho, Françoise Paheau, Elena Pavia, Annette Reid, Ma-Ji Rhee, Shahnaz, Harini Srinivasan, Sydney Stephenson, and Suwako Watanabe for taking the time to record the prompts and instructions used in our data collection protocol. We also thank the 2000-plus anonymous volunteers who so graciously "donated their voices" to science.

REFERENCES

References

[1] M. A. Fanty, J. Pochmara, and R. A. Cole. An interactive environment for speech recognition research. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.

[2] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings International Conference on Spoken Language Processing 92*, Banff, Alberta, Canada, October 1992.