Multi-language Speech Database: Creation and Phonetic Labeling Agreement

Terri Lander, Beatrice Oshika, Ronald A. Cole, and Mark Fanty

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
Portland, Oregon USA
tlander@cse.ogi.edu

ABSTRACT

The focus of the paper is the evaluation of inter-labeler reliability on broad phonetic transcriptions when labelers do not necessarily know the language they are labeling. We provide an analysis of label disagreements, presenting results from six languages, English, French, German, Japanese, Spanish, and Vietnamese with a total of 2 minutes of continuous labeled speech. Labeler agreement across languages ranges from 41 percent with detailed label to label comparisons to 91 percent when less fine comparisons were made.

INTRODUCTION

This paper describes research on a large multi-language speech database being collected at the Oregon Graduate Institute (OGI). The Center for Spoken Language Understanding (CSLU) at OGI has been developing multi-language telephone speech corpora for the last 5 years. An earlier corpus [1] contained data from 11 languages with 90 speakers per language. Presently a 22 language corpus with over 200 native talkers per language is being collected with a wide representation of language types: Eastern Arabic, Cantonese, Czech, English, Farsi (Modern Persian), French, German, Hindi, Hungarian, Japanese, Korean, Malay, Mandarin, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil, and Vietnamese. The corpus consists of short responses to 21 questions plus extemporaneous responses up to 60 seconds long. The corpus will be donated to the National Institute of Standards and Technology and to the Linguistic Data Consortium.

Each call is verified by two native talkers who verified that the caller followed the instructions to each prompt, and made judgments as to regional accent, language competency (fluency), age of talker, telephone line quality, background noise and call completion.

Up to one minute of spontaneous speech and responses to questions are being transcribed at the orthographic level by two native talkers, with disagreement resolution. A standard method for transcribing continuous speech, including pauses and nonspeech sounds has been developed [2]. In addition, trained linguists will label two one-minute sections from each language at the broad phonetic level using Worldbet [3]. [1]

An earlier study [4] compared agreement of broad phonetic labels by both native and non-native talkers of five different languages. Label agreement between native speakers averaged 68%, while agreement between non-natives was much less consistent at 34%. This paper reports on results from labeled speech in six languages, and includes an analysis of phonetic categories on which labelers most disagree, with possible explanations of such variation.

TRANSCRIPTIONS

Transcription was supported by the OGI speech tools [5] which display the waveform and corresponding spectrogram. Transcribers were able to play any part of the waveform multiple times as needed. The labelers used Worldbet, an ASCII rendering of the IPA for broad phonetic transcriptions.

Worldbet attempts to represent phonetic and phonemic distinctions within a single level of transcription. Base symbols generally capture phonetic detail that might otherwise be described by rule, e.g., the Spanish stops /d/ and /t/ are transcribed in Worldbet as explicitly being dental: d[and t[. Diacritics are used to label allophonic variations. A nasalized vowel /i:/ in English would be i:_~ but nasalized vowels which are phonemic in the language, such as the French nasalized vowels, are transcribed A~ where nasalization is part of the base symbol, not a diacritic.

Little prior discussion went into specific labeling and segmentation conventions, although the transcribers did label and compare 10 seconds of speech per language to gain a basic familiarity with each language and speaker. Orthographic transcriptions produced by native speakers were also available to the phonetic labelers to assist in decisions about the choice of base symbol. These were useful when the transcribers were not familiar with the languages.

TRANSCRIBERS

The two labelers are trained in phonetics and acoustics. Both are native speakers of English and are familiar with Spanish. They have less or no knowledge of the other languages labeled. Both of the transcribers have had extensive experience labeling speech.

DATA

The data transcribed for this experiment were a subset of the OGI 22 Language Telephone Speech Corpus described in the introduction. Three 10-second segments of continuous speech were selected for English, German, French, Japanese, Spanish, and Vietnamese. The data selected were gender balanced.

Two ten-second segments of speech in each language (a total of 12 ten-second segments, or two minutes of speech) were labeled independently by the two transcribers.

ANALYSIS

Inter-transcriber agreement was measured in terms of the number of substitutions, deletions and insertions required to map one transcription to another. The "reference" transcription was chosen arbitrarily.

When computing the mapping, overlap in time and phonetic similarity were considered when deciding which segments were substituted, inserted and deleted. This occasionally resulted in a very slightly smaller accuracy than the optimal. However, it results in much more accurate and meaningful confusion matrices. Accuracy was computed as follows:

$$ACC = (ref - sub - ins - del)/ref$$

where ref, sub, ins, and del represent total number of reference segments, substitutions, insertions, and deletions, respectively.

The average accuracy for the set of files in each language was computed using the average number of reference segments, substitutions, insertions and deletions over both of the files.

Six scores were calculated, using the original labels and five less fine sets.

Original Labels To facilitate the anal-

[1] Phonetic label sets were developed by Dr. James Hieronymus, author of [3].

ysis, all non-speech labels were mapped to a single symbol, and adjacent, identical symbols were collapsed.(Table 1 Column 1)

Diacritic Stripping This is a reduced symbol set produced by stripping diacritic information but maintaining the base symbol (Table 1 Column 2).

Broad category We reduced labels into: vowel, plosive, fricative, approximant, nasal, and non-speech (Table 1 Column 3).

Vowel Agreement Additional analysis was performed that clustered vowels by place of articulation. Diphthongs were not included unless the place of articulation fell entirely within the space defined by the cluster. Three different vowel sub groupings were used; all non-vowel sounds were removed from the files so that the scoring algorithm would reflect errors in vowel category only:

1. high, mid, low (Table 2, Column A)
2. front, central, back (Table 2, Column B)
3. high-front, high-back, central, low-front and low-back. (Table 2, column C)

RESULTS

Table 1 displays results for three label comparisons. As expected, agreement improves as the distinctions within the symbol set are reduced.

Table 2 compares agreement between different vowel reductions based on place of articulation.

Table 3 displays the number of base symbols and the number of vowel symbols available per language.

Table 4 displays various usage patterns of original labels.

Table 1: *% Average transcriber agreement at three levels of precision: 1) original labels 2) diacritic stripped 3) broad category*

	1	2	3
Eng	55(143)	67(124)	83(143)
Fre	59(119)	60(97)	83(119)
Ger	41(122)	52(105)	74(122)
Jap	72(143)	77(118)	91(143)
Span	71(107)	78(94)	86(107)
Viet	60(104)	68(84)	84(104)
ave	59	67	84

Table 2: *% Average transcriber agreement for three levels of vowel reduction: A) high, mid low, B) front, central, back, C) high-front, low-front, central, high-back and low-back D) contains the average number of reference segments*

	A	B	C	D
EN	55	54	52	38
FR	62	73	62	45
GE	52	61	54	38
JA	75	78	77	57
SP	85	86	81	47
VT	52	62	56	27
ave	67	71	66	

DISCUSSION

As a follow up to [4] we wanted to do a more careful error analysis of labeler disagreement. In the present experiment, labeler agreement across languages ranges from 41 percent with detailed label to label comparisons to 91 percent when less fine comparisons were made. This compares to 33% and 83% in [4]. Perhaps using orthographies in addition to labeling and comparing test data prior to actual labeling helped to raise over all agreement.

Lower agreement with the full label set (Table 1) seems to result in part from convergence on "preferred" but differing sets of symbols. This happened with various symbols (see Table 4). L1 preferred the devoicing diacritic, using it 24 times more often than L2. L2 used the nasalization diacritic 22 times more often than L1. L1 used closure labels 44 times more often than L2.

Table 3: *Number of base symbols available (BL); number of vowels and diphthongs available (VL)*

	EN	FR	GE	JA	SP	VT
BL	63	47	72	70	41	64
VL	20	17	30	17	12	28

Table 4: *Specific examples of label divergences: the number of times each symbol (base label(b) or diacritic(d)) was used by each labeler (not necessarily simultaneously.)*

	L1	L2
closure(b)	222	178
schwa(b)	87	64
devoicing(d)	28	4
nasal(d)	3	35
centralize(d)	10	0

Over specificity factored in to some of the disagreements. L1 used an average of 6.1 (5%) more symbols per file than L2, using from -3 (Spanish) to 21 (English) symbols more than L2.

The variability in vowel comparisons (Table 2) seem to be related to the number of vowel labels available to transcribers for each language. Spanish and Japanese, both with relatively small vowel inventories, represented the greatest agreement. Although English and Japanese had the same number of vowels (Table 3), there were actually only 7 places of articulation represented in the Japanese vowel labels, as five of the Japanese vowels differ only in length.

Label inventory seems to influence agreement more than knowledge of the language, because although transcribers were familiar with Spanish and English, they agreed more often in Spanish, with its smaller label inventory.

In the future we plan to expand this experiment by labeling a larger set of languages, more speech per language, and a variety of speakers in each language. We also plan to further analyze the role played by the orthographies for non-native transcribers.

ACKNOWLEDGEMENTS

This research was funded by U S WEST, the Office of Naval Research, and the National Science Foundation.

REFERENCES

[1] Y.K. Muthusamy R.A. Cole, and B.T. Oshika, "The OGI Multi-language Telephone Speech Corpus", *The International Conference on Spoken Language Proceedings,* Banff, Alberta, Canada, Oct 1992, pp 895-898.

[2] T. Lander, S.T. Metzler, "The CSLU Labeling Guide", CSLU, Oregon, February, 1994.

[3] J.L. Hieronymus. "Ascii phonetic symbols for the world's languages: Worldbet". AT&T Bell Laboratories Technical Memo, 1994.

[4] R. Cole, B.T. Oshika, M. Noel, T. Lander, M. Fanty, "Labeler Agreement in Phonetic Labeling of Continuous Speech", Proceedings ICSLP94, Yokohama, Japan, September 1993, pp. 2131-2134.

[5] CSLU. "OGI speech tools user's manual," Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1993.