LABELER AGREEMENT IN PHONETIC LABELING OF CONTINUOUS SPEECH

Ronald Cole, Beatrice T. Oshika,
Mike Noel, Terri Lander, and Mark Fanty

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
20000 N.W. Walker Road, P.O. Box 91000, Portland, OR 97291-1000, USA

ABSTRACT

This paper analyzes inter-labeler agreement of label choice and boundary placement for human phonetic transcriptions of continuous telephone speech in different languages. In experiment one, English, German, Mandarin and Spanish are labeled by fluent speakers of the languages. In experiment two, German and Hindi are labeled by linguists who do not speak the languages. Experiment two uses a somewhat finer phonetic transcription set than experiment one. We compare the transcriptions of the utterances in terms of the minimum number of substitutions, insertions and deletions needed to map one transcription to the other. Native speakers agree on the average 67.52% of the time at the finest level of labeling, including diacritics. Non-native linguists agree 34.41% of the time. The implications of the results are discussed for evaluation of phonetic recognition algorithms.

1. INTRODUCTION

Phonetically transcribed continuous speech databases are important for understanding the phonological structure of fluent speech and for developing and evaluating segmentation and phonetic recognition algorithms for speech and language recognition. With the availability of phonetically labeled public domain corpora such as TIMIT [1], it has become standard practice to evaluate phonetic recognition algorithms in terms of hand labeled speech.

When evaluating recognition algorithms in terms of hand labeled speech, we implicitly believe that the phonetic labels produced by human transcribers are "correct." Is this a reasonable assumption? Perceptual studies support the notion of variability in human judgments of speech sounds; subjects presented with vowel sounds excised from TIMIT utterances agreed on the vowel category about 60-75% of the time when the vowel was presented in its left and right phonetic context [2], [3].

Given this perceptual variability, it is useful to examine the level of agreement among transcribers. We examined inter-transcriber reliability as a function of precision of available labels and the transcribers' familiarity with the language.

In experiment one, up to 50 seconds of extemporaneous speech from ten telephone calls in four languages—English, German, Mandarin and Spanish (approximately 30 minutes of continuous speech), were selected from the OGI Multi-Language Corpus [4]. Each call was transcribed independently by two transcribers. The transcribers were either native speakers of the language or considered very competent speakers. They each underwent extensive training procedures to learn the labeling tools, label sets, and labeling conventions. They were able to listen repeatedly to any length interval of speech and to view an associated waveform and spectrographic display. They were asked to mark segment boundaries, and to label the segments using OGI-bet [5].

A second experiment compared transcriptions of 15 two-second files of German and Hindi by two labelers trained in phonetics who are are not speakers of Hindi or German. The transcribers used Worldbet [6] and a set of segmentation conventions taken from [5] developed particularly for this study. The objective of this study was to compare the performance of transcribers with extensive phonetics background but limited familiarity with the specific languages.

2. EXPERIMENT ONE

2.1. Transcribers

The transcribers for each language were either CSLU staff or students. All were native speakers of the language they were labeling, or were considered very fluent in the language.

The English transcribers were VW (native) and TD (native). They are professional phonetic transcribers, each of whom has completed a spectrogram reading course offered by CSLU and has had extensive training and experience in labeling.

The German transcribers were KB (native) and AJ (native). KB has completed the spectrogram reading course, and AJ received extensive on-the-job training.

The Mandarin transcribers were LJ (fluent), YY (fluent), and ZH (native). Each completed the spectrogram reading course and received extensive training. Although there are three labelers, only two labeled each story.

The Spanish transcribers were TL (fluent) and AJ (native, see German). TL is trained in phonetics and has completed the spectrogram reading course along with practical training.

2.2. Transcription Procedure

Transcription was supported by the OGI Speech tools [7], which display the waveform and corresponding spectrogram.

An early version of the CSLU Labeling Guide [5] defined the label set and segmentation procedures. The labelers used OGIbet, which is a broad phonetic label set based on TIMIT. OGIbet offers additional phonetic detail beyond TIMIT by use of diacritics. In OGIbet, labelers had approximately 55 English labels, 62 German labels, 50 Mandarin labels and 42 Spanish labels at their disposal. These were all base labels, which do not include diacritics. Additionally, there were 12 nonspeech labels.

2.3. Data

The data transcribed for this experiment were a subset of the OGI Multi-language Telephone Speech corpus [4], the NIST standard for automatic language identification. This corpus contains several 50 second segments of continuous speech, referred to as "stories." Ten stories were randomly selected in each language, resulting in approximately 30 minutes of speech. Each story was transcribed by two labelers.

2.4. Analysis

Label selection. Inter-transcriber agreement was measured in terms of the number of substitutions, deletions and insertions required to map one transcription to another. The scoring algorithm required one sequence of labels to be the "reference" string, and computed the minimum number of substitutions, insertions and deletions that were needed to transform it to the second label sequence, the "hypothesis" string for each pair of transcriptions. We chose the reference string arbitrarily. 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 all the files.

In addition to the original analysis using the full label set, two more scores were calculated. The scores measured accuracy after mapping the labels to a less specific set. The additional two levels were:

- Reduced symbol set produced by stripping diacritics but maintaining the base symbol.
- Broad categories representing: vowel, closure, plosive, fricative, semi-vowel, nasal, and nonspeech.

Boundary placement. A non-trivial aspect to producing time-aligned transcriptions is the placement of the boundaries between labels. The boundaries represent the location, in time, where the speech represented by the bounded label begins and ends. Some of these locations are arbitrary and must be defined by convention. Agreement on boundary location for segments on which the labelers agreed is reported for each language.

2.5. Results

Table 1 shows the results of the analysis for each of the four languages. The "full" column represents the analysis without any label set reductions. The "base" column refers to the label set with diacritics removed. The "broad" column refers to the broad category analysis. The "segments" column shows the average number of reference segments for each story. Table 2 shows, for labels mapped to the same broad category, the number of milliseconds for which the boundaries differ.

Table 1. Transcriber agreement at three levels of phonetic precision where transcribers are native or fluent speakers of the language.

	Full	Base	Broad	Segments
English	69.67	70.79	89.06	512
German	60.98	64.69	80.78	533
Mandarin	65.61	77.90	86.75	410
Spanish	73.81	81.77	90.13	523

Table 2. Transcriber agreement on location of boundaries for experiment one.

	milliseconds			
	< 2	< 4	< 6	< 11
English	29%	55%	67%	79%
German	21%	46%	63%	79%
Mandarin	32%	58%	71%	83%
Spanish	20%	40%	53%	71%

3. EXPERIMENT TWO

The second experiment explored the transcription agreement of two labelers with extensive phonetics training using a more detailed phonetic label set. In contrast to the labelers in the first experiment, these transcribers are not native speakers of the languages they labeled.

3.1. Transcribers

The two transcribers were TL (also transcribed Spanish) and BO. Both are trained in phonetics, in spectrogram reading, and in the use of the OGI Speech tools.

3.2. Transcription Procedure

Transcription was done using the OGI Speech Tools [7]. The labelers used Worldbet, [6] which captures more phonetic detail in the base symbol set than in OGIbet. There were 69 German base labels to choose from (compared to 62 OGIbet) and 67 Hindi base labels. In addition, there were nine nonspeech labels available. The seven additional German Worldbet labels required finer discrimination of vowels and diphthongs.

3.3. Data

For this experiment a two-second excerpt was extracted from each of 15 German and Hindi stories. These segments came from the OGI Multi-language Telephone Speech corpus [4]. The segments were gender balanced, began and ended with silence, and had few intersegmental pauses.

3.4. Analysis

The same analyses were performed as in experiment one.

3.5. Results

Table 3 shows the agreement between the two labelers for both Hindi and German. Table 4 shows, for labels on which

Table 3. Transcriber agreement at three levels of phonetic precision where transcribers are not native or fluent speakers of the language.

	Full	Base	Broad	Segments
German	34.79	40.50	77.52	25
Hindi	34.03	42.22	82.87	26

Table 4. Transcriber agreement on location of boundaries for experiment two.

	milliseconds			
	< 2	< 4	< 6	< 11
German	32%	59%	69%	81%
Hindi	27%	56%	67%	79%

there is agreement, the number of milliseconds by which the boundaries differ.

4. DISCUSSION

In the first experiment, using native speakers, the average agreement using the full label set was 67.52%. Removal of diacritics raised this to 73.79%, mostly because of gains in Mandarin and Spanish. In Mandarin, diacritics were used to indicate tone, and disagreements on tone assignment increased the overall level of disagreement. More detailed analyses of tone labeling should be done in a future study. In Spanish, with its fewer phonemic vowels than English or German, labelers used more diacritics to indicate the phonetic vowel variability.

When phonetic labels were combined into broad categories, agreement between labelers reached 86.68% in experiment one. Of the remaining disagreements, 82.94% were insertions or deletions, indicating over 10% disagreement about the number of basic speech segments. The amount of agreement may improve if the alignment found by the scoring algorithm were guided by phonetic similarity, i.e., based on place of articulation as well as manner, which is planned for future studies.

The agreement for experiment two is even lower, 34.41% for the full label set comparison and 80.2% for the broad category comparison. The labelers in this experiment had more linguistic training, but did not speak the language they were labeling. A possible explanation is that phonemic expectation—knowing the expected sounds of the language—leads to a bias towards the expected labels and, hence, greater consistency.

Although the differences observed between experiments 1 and 2 are intriguing, comparison of the experiments is confounded by several factors; different languages, a change in label sets, and large differences in the amount of speech data and the duration of the speech excerpts labeled. Additional studies are needed to confirm the difference between native and non-native speakers using the same symbol set and more languages.

The boundary analyses showed that once labelers agree on the broad category of a sound, boundary placement is within 10 milliseconds an average of 78% of the time in experiment 1, and 80% of the time in experiment two. Segmentation does not seem to be affected by the fact that a labeler does not speak the language.

Table 5. Transcriber agreement of English within broad categories for base labels.

	count	correct	subs	dels
vowel	3118	59%	37%	3%
nasal	937	89%	7%	2%
semivowel	1125	79%	15%	4%
plosive	1293	90%	4%	4%
closure	1225	86%	9%	4%
fricative	1501	82%	12%	5%
nonspeech	1123	78%	7%	13%

Even given their preliminary nature, the results presented here, taken together with the results of human perceptual experiments, indicate that there is a great deal of uncertainty about the exact phoneme sequence realized by a given acoustic signal. The results clearly suggest that there is no single "correct" transcription of an utterance. Even professional labelers with extensive training on labeling conventions disagree about 30% of the time on transcriptions of their native language.

These results are interesting in light of the many experiments which compare phonetic recognition algorithms against hand labeled speech. Agreement with a single hand labeled transcription is certainly one indicator of recognition performance. However, it may be more insightful to examine machine recognition performance relative to human labeling, and to present these comparisons for specific broad categories. Analysis of English labels from experiment one, shown in Table 5, reveals that labeler agreement for base labels differs substantially as a function of broad categories. It would be interesting, therefore, to report machine recognition results for phonetically labeled corpora (e.g., TIMIT) by broad category, and compare these results with those produced by two or more human labelers.

Our future analyses are aimed at discovering these patterns of high and low reliability among human labelers for different languages. Whatever the outcome of these analyses, it is clear that care should be taken when defining the "correct" phoneme sequence for a spoken language corpus however it is derived.

5. ACKNOWLEDGMENTS

Without the dedicated efforts of the transcribers it would not have been possible for us to present this paper. Many thanks to Terri Durham, Vince Weatherhill, Anna Johansen, Kay Berkling, Zhihong Hu, Yonghong Yan, and Li Jiang for transcribing the data.

We would also like to thank Johan Schalkwyk for the maintenance of the CSLU speech tools during the transcription process.

6. REFERENCES

1. W. Fisher, G. R. Doddington, and K. Goudi-Marshall. "The darpa speech recognition research database: Specification and status," *Proceedings DARPA Speech Recognition Workshop*, pp 93-100, February 1986.
2. Ronald A. Cole and Yeshwant K. Muthusamy, "Perceptual Studies On Vowels Excised From Continuous

Speech," *Proceedings ICSLP*, 1992, pp 1091-1094.

3. Gary N. Tajchman and Marcia A. Bush, "Effects of Context and Redundancy in the Perception of Naturally Produced English Vowels," Proceedings ICSLP, 1992, pp 839-842.
4. Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," *Proceedings of the International Conference on Spoken Language Proceedings*, Banff, Alberta, Canada, October, 1992, pp 895–898.
5. Terri Lander, S. T. Metzler, *The CSLU Labeling Guide*, CSLU Oregon, February, 1994.
6. James L. Hieronymus, "Ascii phonetic symbols for the world's languages: Worldbet," AT&T Bell Laboratories Technical Memo, 1994.
7. CSLU. "OGI speech tools user's manual," Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1993.