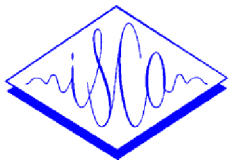
ISCA Archive
http://www.isca-speech.org/archive



Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

What did they actually say? Agreement and Disagreement among Transcribers of Non-Native Spontaneous Speech Responses in an English Proficiency Test

Klaus Zechner

Educational Testing Service
Princeton, NJ, USA
kzechner@ets.org

Abstract

This paper presents an analysis of differences in human transcriptions of non-native spontaneous speech on a word level, collected in the context of an English Proficiency Test. While transcribers of native speech typically agree at a very high level (5% word error rate or less), this study finds substantially higher disagreement rates between transcribers of non-native speech (10%-34% word error rate).

We show how transcription disagreements are negatively correlated to the length of utterances (fewer contexts) and to human scores (impact of lower speaker proficiency) and also seem to be affected by the audio quality of the recordings.

We also demonstrate how a novel multi-stage transcription procedure using selection and ranking of transcription alternatives by peers can achieve a higher quality gold standard that approaches the quality of native speech transcription.

1. Introduction

In order to analyze the spontaneous speech responses of non-native speakers in our English Proficiency Test (EPT), an important first step is to obtain accurate verbatim transcriptions thereof.

Moreover, verbatim transcriptions of speech serve an important role for automatic speech recognition (ASR), which we are also interested in here. In system training, they allow the recognizer to match the acoustic signal to the phone string (computed via the recognizer dictionary) and that way assign posterior probabilities to acoustic features (computed from the audio signal) given different phonetic contexts (acoustic model, AM). Also, the sequence of words is used to compute posterior probabilities for words to occur in a certain context of words already spoken (language model, LM). While a recognition system for read speech can just use the texts being read as transcripts, assuming the readers make few reading errors, the situation is different for spontaneous speech as in our English Proficiency Test, where candidates respond to short prompts[1] with speaking times of 15-60 seconds per prompt (medium or high-entropy speech[2]).

Sometimes, a sub-optimal training can be done by using an existing recognizer to first hypothesize words and then use these hypotheses (or parts thereof with higher confidence scores) as pseudo-reference for the genuine training ([1]). However, this method only works reasonably well if the word error rate (WER[3]) of the initial recognizer is already quite low.

In our situation, WER is typically around 50% since we have not only non-native speech, but speakers from widely varying native language backgrounds and from all levels of proficiency (from hardly understandable to almost native-like).

Therefore, not only for a general analysis of the spoken data, but also for ASR purposes, precise verbatim transcriptions of our EPT data are essential.

Intuitively, it makes sense to assume that transcribing this kind of speech is harder than native speech for the above-mentioned reasons. In an earlier transcription effort using related corpora we observed significant transcriber disagreement when doing the initial calibration to ensure every transcriber is familiar with the transcription guidelines.

The purpose of this current study is to follow up on these observations in a systematic way and to have a larger set of transcribers work on all medium and high entropy items of the EPT, to compute disagreement (measured in WER), and to find possible causes and remedies.

In this context it is interesting to note that very little attention has been given to the issue of agreement, coherence and validity of speech transcription by human transcribers. There is work in the earlier years of ASR of spontaneous speech (e.g., [2]) which observed word level disagreement of less than 5% in two spontaneous speech corpora, even with high noise levels in the signal. The

[1] A prompt is simply the "test question" provided by a native speaker that calls for a spoken response by the candidate.

[2] With "high entropy speech" we refer to spontaneous speech with highly unpredictable word sequences, in contrast to "low entropy speech", such as reading aloud, where the word sequence is highly predictable.

[3] WER is defined as usual as the ratio of all word errors (substitutions, deletions and insertions) and the length of the reference. In this paper we typically multiply WER by 100.0 and obtain "WER in percent".

authors observed that familiarity with the context was a big factor and also attention and motivation of transcribers. More recent work was done on the Buckeye corpus ([3], [4]) and here, agreement was found to be around 98% on a word-token basis.

It is probably because of these high levels of transcriber agreements in native speech that this issue has been under-explored in the past. In other areas, such as in phonetic or prosodic annotations, however, there are more studies on human agreement on native speech (e.g., [5], [6]).

For non-native spontaneous speech, which has a much shorter history of ASR research (about one decade or so), there have been no thorough studies on human transcriber agreement at the word level in the literature. We thus see this study as the first major contribution to this field. In particular, we here present and discuss a novel multi-stage approach for significantly improving inter-transcriber agreement.

The remainder of this paper is organized as follows: Section 2 provides an overview of the tasks of EPT, Section 3 describes the data used for the study, and in Section 4 the study is described in detail, along with agreement results. Section 5 summarizes the results and provides an outlook on future work.

2. English Proficiency Test

There are six task types in the EPT Speaking test, ranging from reading-aloud tasks to tasks that require short answers and tasks that require extended spontaneous responses of one minute. The tasks differ in both the dimensions of speaking skills measured and the possible score points. A brief description of the tasks, the respective response times and score ranges is provided in Table 1. Note that we include task type 1 only for completeness; we do not use this task type in our study which is only concerned with spontaneous speech.

Task type ID	Task description	Response length in seconds	Score range
1	Read 2 passages aloud	45 x 2	1-3
2	Describe a picture	45	1-3
3A	Respond to a survey	15	1-3
3B		15	1-3
3C		30	1-3
4A	Refer to information in a schedule	15	1-3
4B		15	1-3
4C		30	1-3
5	Respond to a voicemail	60	1-5
6	State an opinion	60	1-5

Table 1. Task characteristics of the English Proficiency Test. Task 1 has 2 passages to be read; tasks 3 and 4 consist of 3 responses each. (Higher scores on the scales correspond to higher proficiency.)

3. Data

We used a total of 540 speech responses from EPT task types 2-6. The responses came from 4 different forms[4] and we split the data into two batches with two forms each (270 speech files in each batch). The reason for this split was that the forms of Batch 1 were from a different EPT administration than those of Batch 2; we noted in an independent study that the audio quality of the Batch 2 responses was considerably lower than that of the Batch 1 responses which might affect the inter-transcriber agreement. Note that task types 3 and 4 have 3 parts each which are always combined in our evaluations below, as their combined duration is comparable to the other 3 tasks 2, 5 and 6.

4. Transcription study

4.1 Transcribers

We used a total of 14 transcribers for this study, all of whom were experienced human raters of high entropy non-native speech but not of the EPT. 11 of these took part in both batches, 2 of them only in batch 1 and 1 only in batch 2.

We assigned the files to the transcribers so that each task was transcribed by at least two transcribers. The survey and schedule sub-tasks were combined into one task. We assigned more transcribers to the task types 2, 5 and 6 since we expected more disagreement due to the longer average response times there, as opposed to tasks 3 and 4 which have response times of 15-30 seconds for each sub-task.[5] The overall workload for each transcriber was comparable – they all had to transcribe about 30 minutes of speech total.

4.2 Transcription guidelines

We used a rather simple set of transcription guidelines which state that all words spoken have to be transcribed, even if repeated or if fillers (e.g., uh, um), but they have to be words of English (i.e., not words from another language or neologisms). Transcribers could further mark words where they were unsure with a special symbol and should also mark longer stretches of silence or unintelligible speech.[6] For this study, to keep focused on the words, we ignore the latter annotations and treat "unsure words" as regular transcribed words.

Transcribers further obtained the prompts of the different test tasks so that they were able to familiarize themselves with the context of the responses.

4.3 Transcription phases

We devised three phases of transcriptions for each batch of files with the idea that the transcriptions would eventually converge to a "gold standard" at the end. In Phase 1, every transcriber had to produce the baseline

[4] A form is a collection of all test items of a test.

[5] As we will see later***CHECK***, though, it turns out that agreement is actually *lower* for shorter utterances.

[6] Non-English words, neologisms and the like were also subsumed under this category.

transcription for his/her set of files (30 files for tasks 2, 5 and 6 and 90 (shorter) files for the tasks 3 and 4).

Task	Tran-scriber pair	WER Phase 1	WER Phase 2	Indivi-dual tran-scriber Phase 2	WER gold vs. Phase 2
2	AN – AR	18.9	5.8	AN	2.0
	AN – BE	10.1	5.6	AR	5.0
	AR – BE	18.5	7.7	BE	4.1
3	BR - CA	13.0	8.9	BR CA	5.4 3.4
4	EI - GI	17.6	10.2	EI GI	3.9 6.3
5	JF – JW	10.3	5.4	JF	6.0
	JF – SH	11.2	8.5	JW	4.7
	JW – SH	12.7	7.2	SH	9.0
6	MK – MY	21.3	11.5	MK	5.3
	MK – ST	14.1	10.8	MY	8.5
	MY - ST	20.0	5.1	ST	3.0
Ave-rage	-	15.2	7.9	-	5.1

Table 2. Word error rates (in %) between transcriber pairs and transcribers vs. gold standard for different tasks in different transcription phases for Batch 1.

The inter-transcriber disagreements, measured as word error rates, were evaluated following this Phase 1.

In Phase 2, transcribers were presented with all baseline transcriptions of their peers (including their own) from Phase 1 in random order. They had to choose the one they felt was best and closest to the true audio, mark it, and then further improve their selected transcription. Again, WERs were computed between the transcribers of the same group.

Finally in Phase 3, the selected transcriptions from Phase 2 were again presented to all transcribers of a group in random order and they had to rank them according to the perceived correctness. Tied rankings were discouraged but allowed. No further edits were allowed in Phase 3, though. (In all phases, annotators had to listen to the original audio files.)

The ranks (r=1, 2, or 3, with r=1 being the highest rank) were converted into scores (s) with s=3-r. The scores were then summed up for all transcriptions and the highest scoring transcription was selected as "gold standard" transcription. In case of ties we selected randomly from the top scored transcriptions.

Also, WER computations comparing the gold standard with Phase 2 transcriptions were performed to see which transcribers were closest to the final gold standard and hence also produced the most accurate transcriptions for this task.

4.4 Results

Tables 2 and 3 show the results of our WER computations between all transcriber pairs of each transcriber group for the batches 1 and 2, respectively. (We are using NIST's `sclite` package for this purpose.)[7] We provide the WERs for the baseline (Phase 1), the improved transcriptions (Phase 2) and finally for the comparison between gold standard transcriptions and Phase 2 transcriptions for each annotator. In Batch 1, less than 5% of transcriptions had to be excluded from Phase 1 evaluations and less than 14% from Phases 2 and 3 due to empty responses or files not returned by the transcribers. This affected mostly task 2 (picture) for Phases 2 and 3 where one annotator returned only 10 of 30 files.[8]

For Batch 2, less than 5% of transcriptions had to be excluded from the Phase 1 evaluations (empty files, not returned files), less than 6% from Phase 2, and less than 11% from Phase 3.

Task	Tran-scriber pair	WER Phase 1	WER Phase 2	Indivi-dual tran-scriber Phase 2	WER gold vs. Phase 2
2	AN – AR	16.9	10.6	AN AR	6.8 3.9
3	BR – BE	34.0	11.3	BR BE	8.9 2.3
4	EI – TI	31.9	15.1	EI TI	8.0 7.0
5	JF – JW	13.8	10.8	JF	3.1*
	JF – SH	12.9	10.6	JW	5.8*
	JW – SH	12.5	8.1	SH	6.1*
6	MK – MY	24.4	8.1	MK	7.4
	MK – ST	16.1	16.3	MY	3.1
	MY - ST	21.6	11.6	ST	9.8
Ave-rage	-	20.5	11.4	-	6.0

Table 3. Word error rates (in %) between transcriber pairs and transcribers vs. gold standard for different tasks in different transcription phases for Batch 2.
* Due to annotation formatting errors, we could only consider 18 of 30 voicemail files for these evaluations.

Tables 2 and 3 show clearly that inter-transcriber disagreements for non-native spontaneous speech of the EPT are substantially higher (about 15%-20% WER after Phase 1) than what was observed for native speech previously (5% WER or less). After Phase 1, no task had a disagreement of less than 10% and one task's

[7] http://www.itl.nist.gov/iad/mig/tools/

[8] All transcriptions of one file ID were excluded if they were missing from at least one annotator to obtain matching files for the sclite WER comparison runs.

disagreement was as high as 34%, an order of magnitude higher than for native speech transcriptions.
These two tables, however, also show how a multi-stage transcription protocol can achieve significantly lower disagreements than a single-pass protocol would be able to. Average disagreement drops by almost 50% relative between Phase 1 and Phase 2, and by about 70% between Phase 1 and Phase 3. The average final disagreement between the gold standard transcriptions of Phase 3 and Phase 2 transcriptions was only around 5%-6%, which is approaching the disagreement rates found in native speech transcription.

In terms of transcription time, in both batches, a transcriber used about 14 hours total on average for all three phases to transcribe 30 minutes of speech (22.5 minutes for the picture task). Phases 1 and 2 took longer with 6 and 5 hours on average, respectively. Phase 3 which only involved the ranking and no transcription or editing was faster with 3 hours on average.

4.5 Factors influencing disagreement

We looked at three factors that might have an effect on or be correlated with human disagreement, measured in WER.

The most obvious factor concerns the audio quality of the speech samples. While we do not have human ratings of the samples in this study, we have 400 ratings available from the same data set, collected in a different context, 100 samples per form. Since the delivery mode of the 2 forms in Batch 1 was different from that in Batch 2 it is not surprising that the average audio quality score also differs. Using a 5-point discrete scale with 5 being the highest audio quality, the average for the 200 samples of the 2 forms used in Batch 1 was 3.75, while it was only 3.34 for the two forms used in Batch 2. Indeed, the average WERs are higher for Batch 2, compared to Batch 1, particularly for Phases 1 and 2 (last rows of Tables 2 and 3), and so we conjecture that the audio quality might have been an influencing factor in the higher disagreement of transcribers observed in Batch 2.

Factor 2 concerns the relationship between transcriber disagreement and human scores of speech samples. We would conjecture that higher scores and more proficient speakers would be easier to understand and easier to transcribe with fewer disagreements. This assumption is borne out in our evaluation where we correlated the average WERs of all 258 speech samples each in the Phase 1 evaluations of both batches with human rater scores. As we can see in Table 4, we obtain significant negative correlations for both batches.

The third and final factor we looked at was utterance length, i.e., how many words were spoken (transcribed) in a speech sample. The evaluation method was the same as for the human scores and the results are also reported in Table 4. (We averaged the length of utterances across all transcribers of a group). Again, significant negative correlations are observed for both batches, with a stronger correlation in Batch 2. This may be due to the lower audio quality here which may have caused an even higher disagreement for relatively short utterances with little context.

	Number of speech samples	Correlation with human scores	Correlation with utterance length
Batch 1	258	-0.382**	-0.198**
Batch 2	258	-0.344**	-0.392**

Table 4. Pearson r correlations between WERs of 258 speech samples of Phase 1 transcriptions (Batch 1, Batch 2) and (a) human scores and (b) utterance length in words. (** = r is significant at $p<0.01$)

5. Summary and future work

In this paper, we presented a study of human transcription agreement of non-native spontaneous speech in the context of the English Proficiency Test. We found that in the initial transcription phase, transcribers' disagreement as measured in word error rate is significantly higher than reported for native speech transcription, and can be higher than 30% for some tasks and transcriber groups.

We argue for a multi-stage approach in non-native speech transcription which yields significantly lower disagreement word error rates after three transcription phases, approaching agreement rates of native speech transcription.

Aside from yielding the most accurate transcription, the last (third) phase in our approach can also indicate whether transcribers are more or less accurate and reliable in their work.

Therefore, the gold standard transcriptions obtained in this study could also serve as a calibration corpus for future transcriptions of the EPT where transcribers first have to achieve a minimum agreement with these transcriptions before they can be certified.

In future work, we plan to extend the analysis of the transcription errors further and look into the effects of inaccurate transcriptions on automatic speech recognition.

6. References

[1] Kemp, T. & Waibel, A. (1998). Unsupervised Training of a Speech Recognizer Using TV Broadcasts. Proceedings of ICSLP-98, pp. 2207-2210. Sydney, Australia, December.

[2] Deshmukh, N., Duncan, R. J., Ganapathiraju, A. & Picone, J. (1996). Benchmarking Human Performance for Continuous Speech Recognition. Proceedings of ICSLP-96, pp. 2486-2489. Philadelphia, PA, October.

[3] Raymond, W.D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R. & Hilts, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. Proceedings of ICSLP-02, pp. 1125-1128. Denver, CO, September.

[4] Pitt M.A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication, 45 (1), pp.89-95.

[5] Pitrelli, J.F., Beckman, M.E. & Hirschberg, J. (1994). Evaluation of Prosodic Transcription Labeling Reliability in the Tobi Framework. Proceedings of the ICSLP-94, pp.123-126. Yokohama, Japan, September.

[6] Gut, U. & Bayerl, P S. (2004). Measuring the reliability of manual annotations of speech corpora, Proceedings of Speech Prosody 2004, pp. 565-568. Nara, Japan, March.