Automatic Estimation of Transcription Accuracy and Difficulty

*Brandon C. Roy**, *Soroush Vosoughi**, *Deb Roy*

The Media Laboratory, Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

bcroy@media.mit.edu, soroush@media.mit.edu, dkroy@media.mit.edu

Abstract

Managing a large-scale speech transcription task with a team of human transcribers requires effective quality control and workload distribution. As it becomes easier and cheaper to collect massive audio corpora the problem is magnified. Relying on expert review or transcribing all speech multiple times is impractical. Furthermore, speech that is difficult to transcribe may be better handled by a more experienced transcriber or skipped entirely.

We present a fully automatic system to address these issues. First, we use the system to estimate transcription accuracy from a a single transcript and show that it correlates well with inter-transcriber agreement. Second, we use the system to estimate the transcription "difficulty" of a speech segment and show that it is strongly correlated with transcriber effort. This system can help a transcription manager determine when speech segments may require review, track transcriber performance, and efficiently manage the transcription process.

Index Terms: speech transcription, inter-transcriber agreement, human-machine collaboration

1. Introduction

Annotating audio recordings with speech transcripts is an important task for many enterprises. However, as the size of audio corpora increase thanks to advances in recording and storage technology, efficient speech transcription methodologies will become increasingly important. While automatic speech recognition technologies continue to improve, they are inadequate for many tasks and human transcribers are needed.

Speech transcription approaches can be considered along several dimensions. One aspect is the level of annotation detail, and it is common to annotate speech at the phonetic, word and discourse level. The focus here is on word-level (orthographic) transcription of short speech segments. Purely manual approaches are often used for orthographic transcription and tools such as CLAN [1] and Transcriber [2] are popular, but these tools do not perform well for large-scale transcription tasks. In particular, transcription times are usually an order of magnitude longer than the actual audio being annotated [2, 3].

An active area of research focuses on how services such as Amazon's *Mechanical Turk*[1] can be leveraged to reduce the cost of annotation. Speech transcription using Mechanical Turk can attain quality competitive with conventional methods, but at a fraction of the cost [4]. Furthermore, different transcription goals may permit different levels of transcription quality. For example, if the goal is to train a speech recognizer, then more transcription errors may be tolerated if the yield is higher [5].

Mechanical Turk is a promising approach for many applications, but in our work there are certain privacy requirements that must be met. Specifically, we are studying early child language acquisition as part of the Human Speechome Project [6]. For the Human Speechome Project (HSP), the home of one of the authors (DR) was outfitted with a custom audio-video recording system with the goal of recording the first three years of his child's life. This naturalistic, ultra-dense, longitudinal corpus is helping to shed new light on early language learning [7, 8]. But given the privacy concerns with this corpus and the special training required to adequately transcribe a young child's early speech, Mechanical Turk is not a viable solution. Instead, the size of this corpus (roughly 120,000 hours of audio), the nature of the data and the privacy requirements have led us to develop a new semi-automatic transcription system called BlitzScribe [9].

2. Semi-automatic speech transcription with BlitzScribe

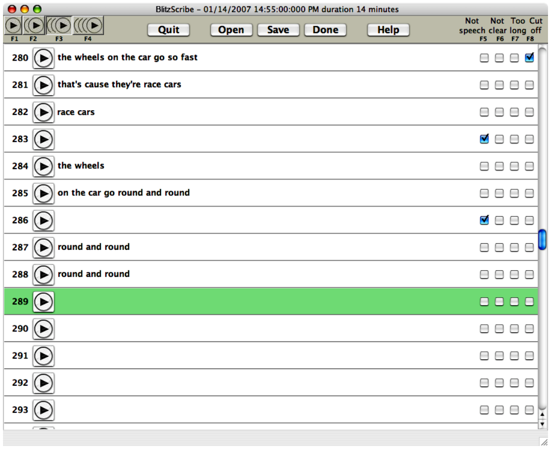
BlitzScribe is a human-machine collaborative system that enables human transcribers to transcribe significantly faster than purely manual approaches. BlitzScribe works by first automatically identifying and segmenting speech in audio, and then presenting it to the human transcriber in a streamlined user interface. The human transcriber plays a segment, transcribes and advances to the next segment using only the keyboard, enabling them to proceed quickly and efficiently and eliminating the need to spend time browsing for and selecting speech using the mouse. Since the speech is automatically detected, the annotator may encounter non-speech while transcribing. Such non-speech segments are marked naturally in the course of transcription by simply leaving them blank and advancing to the next segment. Both the transcribed speech and non-speech segments are then fed back into the system to retrain the speech detection component, thereby adapting and improving the system as a whole. Using BlitzScribe yields up to a six-fold speedup over other purely manual approaches [9]. This corresponds to transcriber time of roughly twice the audio duration. BlitzScribe is particularly effective on *unstructured* audio, where a significant portion of the audio consists of non-speech. While a purely manual tool usually requires scanning all the audio to identify speech, transcriber time in BlitzScribe depends primarily on the amount of speech. The BlitzScribe user interface is shown in figure 1.

3. Estimating transcription accuracy

Speech transcription can be challenging, and spontaneous speech captured in a naturalistic environment can be particularly difficult to transcribe. Such speech is often mixed with background noise and other non-speech sources, can include

*The first two authors contributed equally to this work.

[1] http://www.mturk.com



BlitzScribe – 01/14/2007 14:55:00:000 PM duration 14 minutes

		Not speech F5	Not clear F6	Too long F7	Cut off F8
280	the wheels on the car go so fast	□	□	□	☑
281	that's cause they're race cars	□	□	□	□
282	race cars	□	□	□	□
283		☑	□	□	□
284	the wheels	□	□	□	□
285	on the car go round and round	□	□	□	□
286		☑	□	□	□
287	round and round	□	□	□	□
288	round and round	□	□	□	□
289		□	□	□	□
290		□	□	□	□
291		□	□	□	□
292		□	□	□	□
293		□	□	□	□

Figure 1: *BlitzScribe user interface*

multiple overlapping speakers, and can vary in loudness and speaking style. In the Speechome Corpus, one of the primary speakers is a young child in the early stages of language learning, while another is a non-native speaker. Given these challenges and the size of the transcription task, a major concern is monitoring and maintaining transcription quality.

One approach is for an expert to review a portion of the transcripts. Another approach is to have multiple transcribers transcribe the same segments, and then automatically compare their transcriptions. The assumption is that if multiple transcribers have produced the same transcript, they have likely converged on the correct transcription. This approach was taken in [9] using the NIST `sclite` tool [10] and the Sphinx-4 [11] implementation of the alignment algorithm. This tool is often used to evaluate automatic speech recognition (ASR) performance, where errors in the hypothesis transcript produced by the ASR system are identified relative to the reference transcript. Here, rather than a hypothesis and reference transcript for a pair of transcribers, a symmetric agreement value was calculated by treating each transcript as first the hypothesis and then the reference and averaging the scores. This approach is very useful for monitoring transcriber performance and detecting problems, but it has the shortcoming that obtaining inter-transcriber agreement scores requires significantly more human effort – at least twice as much effort per segment – and thus can be quite costly.

Instead of relying on inter-transcriber agreement, we apply an automatic system to estimate transcription accuracy using only a single transcript per segment. The main component of the system is an acoustic matcher that computes the match between an audio segment and the transcript. The matcher is built using HTK [12] and works as follows. Given a speech segment and its transcript, the Viterbi algorithm in HTK is applied to align the transcript to the speech segment on the phoneme level, using the CMU pronunciation dictionary. Each candidate alignment for each phoneme comes with a score that measures the acoustic similarity of the HMM of the phoneme in the transcript to the corresponding audio, normalized by the number of frames. Since each transcript comes with a speaker ID label that is generated automatically, we use a speaker-specific acoustic model to obtain the *acoustic alignment* scores. The scores for all the phonemes in the segment are then averaged to get a single score for the segment. This score measures how well the transcript matches the audio at the phonetic level. Roughly, the idea is that segments where the transcript is a poor fit to the audio, due either to an incorrect transcript or low audio quality, are more likely to contain transcription errors. This pipeline is shown in figure 2.

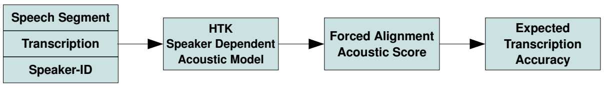


Figure 2: *Pipeline for calculating the expected accuracy of a speech transcription.*

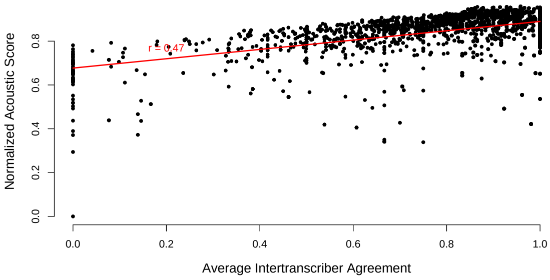


Figure 3: *Acoustic alignment score vs. average inter-transcriber agreement.*

Figures 3 and 4 show the relationship between the acoustic alignment score and the inter-transcriber agreements. In figure 3, the inter-transcriber agreement for a transcriber is the average of their pairwise inter-transcriber agreements against the other transcribers who have transcribed a segment. The strong correlation ($r = .47$, $p < .001$) shows that the acoustic alignment score is a very good predictor of the average inter-transcriber agreement value for a transcription. Thus, our alignment score could be used as a proxy for inter-transcriber agreement without requiring any additional transcription effort. Figure 4 shows a somewhat different relationship. Here, the *difference* of the acoustic alignment scores for a pair of transcripts on the same segment are plotted against the inter-transcriber agreement value for those same transcripts. In this case, we obtain a strong correlation of $r = 0.5$, $p < .001$. Finally, the average acoustic alignment score for a transcriber can be used to evaluate overall transcriber performance, as shown in figure 5. This figure shows that T3 has the highest accuracy score while T2 has the lowest, indicating that T3 is likely a more accurate transcriber.

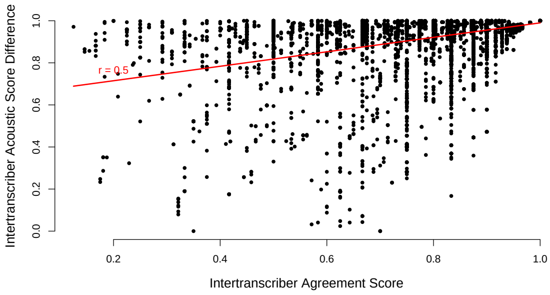


Figure 4: *Difference of acoustic alignment scores between a pair of transcribers vs. their inter-transcriber agreement.*

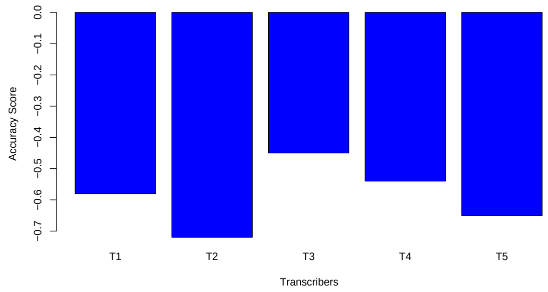


Figure 5: *The average acoustic alignment score for each transcriber.*

4. Predicting transcription difficulty

In a large-scale transcription effort, efficiently distributing work to the transcription team is important. For a fixed budget, one might wish simply to transcribe as much speech as possible. Or, one might wish to have a more experienced transcriber work with the more challenging audio. Both of these goals could be served if we could estimate the difficulty of transcribing a speech segment in advance of actually transcribing it. This section details an experiment using unconstrained acoustic recognition on the speech segments as a predictor of transcription difficulty.

Actual transcription difficulty was measured by recording the transcriber activity associated with each segment. The intervals during which a segment was selected, the start and stop times for playback, the number of plays, the intervals during which the transcriber was actively editing a segment (determined by monitoring keyboard activity) and the number of edit intervals were logged. A visualization of these transcriber logs is presented in figure 7, which shows segments being selected, played and edited. Note that in some cases a segment is played multiple times, or there are multiple edit intervals. We define the *user activity interval* as the interval over which the transcriber is either playing or editing a segment. We take this quantity to be the amount of work required for a segment. Since the actual audio segment length varies, we take the ratio of work to segment length to reflect the "difficulty" of the segment.

An important confound to address are situations where the transcriber spends a long time on a segment but is not actively working, such as during a break. Fortunately, in these situations the ratio of user activity for a segment to the time the segment was selected will be very low. For example, in some cases a segment is selected for 10 minutes although they only spent 10 seconds actively working. By thresholding on this value we can filter out these cases. Figure 6 shows a histogram of the number of segments in the transcription log binned by the transcriber activity percentage. Most of the segments show that the transcribers were actively working, but by using this histogram we can choose an appropriate cutoff point for segments where the transcriber was either distracted or not working.

The question we wish to answer in this section is whether a purely automatic method can predict the transcription difficulty of a segment before it is transcribed. Our method applies HTK's [12] automatic speech recognition tools to the audio segment to obtain an acoustic recognition score. The acoustic score captures how well the acoustic model fits the given audio. If there was a lot of noise, if the speech was in another language, or the words were not spoken clearly one would expect a low acoustic score. As with the accuracy measure, Viterbi is run on speech segments, but this time uses a speaker independent acoustic model. Since there is no transcript associated with the speech segment, the system essentially performs automatic speech recognition using a complete English dictionary, yielding *acoustic difficulty* scores similar to those generated by the aligner described earlier. This pipeline is depicted in figure 8. The acoustic difficulty score measures how well an untranscribed speech segment matches a speaker-independent acoustic model for English. Figure 9 shows a scatter plot of the acoustic difficulty score against the actual transcriber difficulty. This plot was generated by including only those segments for which the transcriber was actively working for more than 75% of the time the segment was selected. The high correlation ($r = 0.64$ and $p < 0.001$) implies that acoustic difficulty of a segment can be used to predict how difficult it will be to transcribe the segment, and thus how long the transcription will take. As mentioned, if the transcriber takes a break while a particular segment is selected, the activity percentage for that segment will

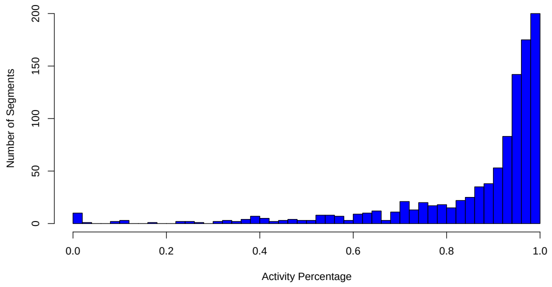


Figure 6: *The number of segments binned by how actively the transcriber worked on the segment. A low activity percentage for a segment may be due to the transcriber taking a break from transcription and thus should not be included in the analysis.*

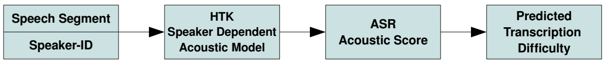


Figure 8: *Pipeline for calculating the predicted transcription difficulty for an audio segment.*

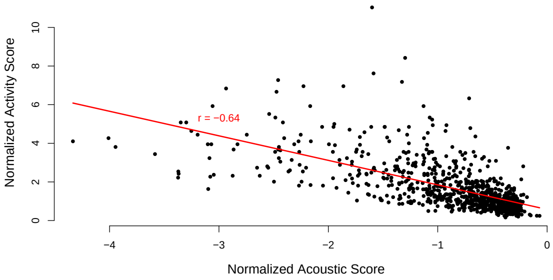


Figure 9: *Scatter plot of acoustic score of a segment against the transcriber "activity" on that segment. Segment for which the transcriber activity relative to segment selection time are greater than 0.75 are included here, resulting in a correlation of* $r = -0.64$, $p < .001$

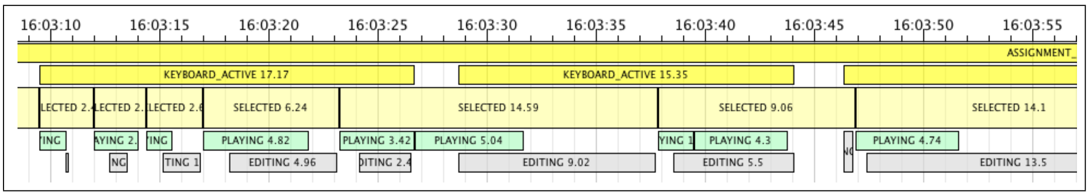


Figure 7: *User activity log showing segment selections, playbacks and edits. Note that playback and editing overlap, and in some cases a segment is played and edited over multiple intervals.*

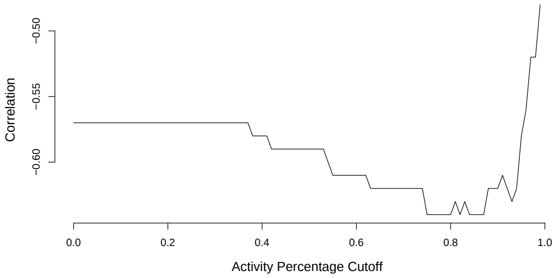


Figure 10: *Acoustic difficulty score vs. actual transcriber difficulty r-values at different activity percentage cutoff thresholds.*

decrease significantly and our model of transcription difficulty will break down. Figure 10 shows how the correlation varies as a function of this activity percentage cutoff.

5. Discussion

We have shown how an automatic system, built from HTK's forced-alignment and speech recognition components, can automatically estimate the accuracy of transcriptions and the effort required to transcribe speech segments. Using this system, we can estimate the accuracy of a given transcript without manual review or the additional effort required to perform inter-transcriber agreement. Though the acoustic alignment score is noisy, it is a good proxy for inter-transcriber agreement and vastly cheaper to obtain. Furthermore, this accuracy measure can be used to evaluate individual transcriber performance. Transcripts with a low acoustic alignment score could be identified in order to re-transcribe them or simply remove them from the dataset.

In addition, this system enables us to automatically predict how hard a given segment will be to transcribe. This gives the transcription manager control over a large-scale transcription task by allowing them to prioritize which speech should be distributed to the transcription team. If the goal is to transcribe as much speech as possible, then perhaps the easiest utterances should be transcribed first. Alternatively, blocks of speech that are easier, on average, could be distributed to new transcribers while more challenging audio is distributed to experienced transcribers. In the future, we may explore how other variables such as number of edits and number of playbacks relate to overall transcription time and how well they correlate with the acoustic difficulty score. Overall, we see this system as an important and useful tool for managing a large-scale transcription effort.

6. Acknowledgements

Many thanks to Karina Lundahl and the transcription team, as well as Matt Miller for his work on the speaker identification system. Special thanks to Ruquia Asghar, Ruchi Gadodia, Drew Laude, Sara Natale, and Halle Ritter for their helpful comments.

7. References

[1] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Lawrence Erlbaum Associates, 2000.

[2] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: Development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, January 2001.

[3] D. Reidsma, D. Hofs, and N. Jovanović, "Designing focused and efficient annotation tools," in *Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands, 2005.

[4] M. Marge, S. Banerjee, and A. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *ICASSP*, March 2010.

[5] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *NAACL*, 2010.

[6] D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, M. Levit, and P. Gorniak, "The Human Speechome Project," in *Proceedings of the 28th Annual Cognitive Science Conference*. Mahwah, NJ: Lawrence Earlbaum, 2006, pp. 2059–2064.

[7] B. C. Roy, M. C. Frank, and D. Roy, "Exploring word learning in a high-density longitudinal corpus," in *Proceedings of the 31st Annual Cognitive Science Conference*, 2009.

[8] S. Vosoughi, B. C. Roy, M. C. Frank, and D. Roy, "Effects of caregiver prosody on child language acquisition," in *Fifth International Conference on Speech Prosody*, Chicago, IL, 2010.

[9] B. C. Roy and D. Roy, "Fast transcription of unstructured audio recordings," in *Proceedings of Interspeech*, Brighton, England, 2009.

[10] J. Fiscus. (2007) Speech recognition scoring toolkit ver. 2.3 (sctk). [Online]. Available: http://www.nist.gov/speech/tools/

[11] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Sun Microsystems, Tech. Rep. 139, November 2004.

[12] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Cambridge University Engineering Dept, 2001.