Crowdsourcing Transcription Beyond Mechanical Turk

Haofeng Zhou
School of Information
University of Texas at Austin
haofzhou@utexas.edu

Denys Baskov
Department of Computer Science
University of Texas at Austin
dbaskov@utexas.edu

Matthew Lease
School of Information
University of Texas at Austin
ml@ischool.utexas.edu

Abstract

While much work has studied crowdsourced transcription via Amazon's Mechanical Turk, we are not familiar with any prior cross-platform analysis of crowdsourcing service providers for transcription. We present a qualitative and quantitative analysis of eight such providers: 1-888-Type-It-Up, 3Play Media, Transcription Hub, CastingWords, Rev, TranscribeMe, Quicktate, and SpeakerText. We also provide comparative evaluation vs. three transcribers from oDesk. Spontaneous speech used in our experiments is drawn from USC-SFI MALACH collection of oral history interviews. After informally evaluating pilot transcripts from all providers, our formal evaluation measures word error rate (WER) over 10-minute segments from six interviews transcribed by three service providers and the three oDesk transcribers. We report the WER obtained in each case, and more generally assess tradeoffs among the quality, cost, risk and effort of alternative crowd-based transcription options.

Introduction

Amazon's Mechanical Turk (AMT) has revolutionized data processing and collection practice in both research and industry, and it remains one of the most prominent paid *crowd work* (Kittur et al. 2013) platforms today. However, AMT provides relatively low-level support for quality assurance and control, as well as mechanisms for tackling more complex or collaborative tasks.

While AMT helped launch the crowd work industry eight years ago, research on crowd work has continued to focus on AMT near-exclusively. Such focus risks letting AMT's particular vagaries and limitations unduly shape our crowdsourcing research questions, methodology, and imagination too narrowly for AMT, "...writing the user's manual for MTurk ... struggl[ing] against the limits of the platform..." (Adar 2011). We are not familiar with any prior cross-platform analysis of crowdsourcing service providers for transcription.

To address this lack of knowledge of crowdsourced transcription beyond AMT, we present a qualitative and quantitative analysis of eight transcription service providers: 1-888-Type-It-Up (formerly Verbal Fusion), 3Play Media, Transcription Hub, CastingWords, Rev, TranscribeMe, Quicktate, and SpeakerText. We also compare to three transcribers with varying hourly-rates from oDesk, an online labor marketplace focusing on more specialized forms of labor than AMT.

Vakharia and Lease (2013) present a qualitative cross-platform evaluation of seven crowdsourcing platforms, assessing distinguishing features at large and their relevance to researcher needs and open problems. In contrast, we focus specifically on transcription (of spontaneous speech), which leads us to evaluate a different set of platforms. Moreover, since we are focusing on a specific task, we are able to provide quantitative as well as qualitative evaluation.

Spontaneous speech used in our experiments is drawn from USC-SFI MALACH collection of oral history interviews of WWII Holocaust witnesses (Byrne et al. 2004). After informally evaluating pilot transcripts from all providers, our formal evaluation measures Word Error Rate (WER) over 10-minute segments from six interviews transcribed by three service providers and three oDesk transcribers. We report WER vs. transcription price, along with other considerations.

For this "heavily accented, emotional and elderly spontaneous speech" (Byrne et al. 2004), oft-recorded in noisy environments, mean WER for the 2006 ASR transcripts is reported as 25% (Pecina et al. 2008). Anecdotally, we have found these ASR transcripts often too difficult to understand at all. Our investigation of crowdsourcing mechanisms for efficient human or human-assisted transcription is motivated by the real need to produce human readable transcripts for challenging spontaneous speech data in practice.

Our contributions include: 1) a qualitative snapshot in time of current crowdsourcing transcription providers and offerings beyond AMT, reviewing alternatives and providing a reference point for future studies; 2) a quantitative assessment of WER vs. cost for spontaneous speech transcription across multiple providers; and 3) discussion of important tradeoffs among quality, cost, risk and effort in crowd transcript work. For example, crowdsourcing research is often "penny-smart, pound-

foolish" in myopic focus on direct transcription costs without considering setup and oversight costs, which can easily dwarf the former in practice.

Any description of present commercial services risks becoming quickly dated with regard to specific details. In the near-term, we hope to provoke more use and study of transcription beyond AMT, and to help inform those using or designing crowd-based transcription methods. In addition, we believe this work will provide a valuable "snapshot in time" of the industry and capabilities in 2013, as a useful reference point for later, retrospective comparison to future developments.

Related Work

As with other applications of crowdsourcing, lower cost vs. traditional practice is a key factor driving crowdsourced transcription as well. Research has almost exclusively focused on Amazon Mechanical Turk (AMT).

Marge et al. (2010b) investigated AMT's capability to support various speaker demographics, and found that the accuracy of results was acceptable, was less influenced by payment than expected, and concluded that AMT was a good resource for transcription. The influence of pay rate and whether the speaker is male/female is also examined by Marge et al. (2010a). They found that higher waged transcriptions had faster turnaround time, with varying payment yielding similar accuracy.

Evanini et al. (2010) considered two more challenging tasks: spontaneous speech and read-out-loud. Each HIT consisted of a batch of 10 responses, paying $3 per batch or $0.30 per transcription. Individual workers achieved 7% WER in the read-out-loud condition and 9.7% in the spontaneous condition, vs. experts achieving 4.7% and 8.1%, respectively. Merger methods achieved better results. LCS and Rover performed better in the read-out-loud condition. Lattice performed best for spontaneous speech, with WER 22.1%.

Audhkhasi et al. (2011) consider 1000 Spanish broadcast clips, transcribed by 5 transcribers each. Gold standard transcriptions were unavailable, so evaluation used Rover tests as gold transcription data. They found that out of 19 workers, WER ranged from 10 to 20% (26 hours of audio split over 1000 clips), with 13% average WER. They found that by combining 5 transcripts using an unweighted Rover algorithm, WER decreased to 2.5% and sentence error rate (SER) to 19.7%. Additional measurements using reliability metrics further decreased WER to 2.3% and SER to 18.9%.

Additional has work addressed the problem whether the cost had effect on final quality of the transcription and how to use it more efficiently. Novotney et al. (2010) proposed a procedure of maintaining the quality of the annotator pool without needing high quality annotation. They also found that higher disagreement did not have a significant effect on the performance, and best resource allocation was to simply collect more data (i.e. quality over quantity). Parent et al. (2010) describe a multi-stage model that when "gold-standard" quality control was used with double cost, the results could achieve close to NIST expert agreement. Lee et al. (2011) also proposed a two-stage transcription task design for crowdsourcing with automatic quality control. Gruenstein et al. (2009) introduced their work on collecting orthographically transcribed continuous speech data via AMT. Williams et al. (2011) studied transcription of difficult speech. They predict reliability to balance precision, recall and cost. Recent work has investigated real-time transcription which could be extended to crowd transcribers (Lasecki et al. 2012).

Evaluation Criteria

This section defines the criteria developed to qualitatively characterize and differentiate crowdsourced transcription providers. Criteria were defined inductively via open-ended review of considered service providers.

- **Base Price.** Base price for transcription is typically billed per minute of audio. Additional fees are typically assessed for features like higher quality, time stamps, multiple speakers and/or speaker identifications, difficult audio, specialized vocabulary (e.g., legal or medical), or faster turnaround. We discuss bulk-order discounts as below the base price.
- **Accuracy.** Often an informally or formally-guaranteed level of accuracy is offered, and typically with tiered pricing for higher quality.
- **Transcript Format.** Providers offer output transcripts in one or more formats which may vary in their match to user needs. These include: Distribution Format Exchange Profile (DFXP), HTML, JavaScript Object Notation (JSON), Adobe PDF, Rich-Text Format (RTF), SubRip subTitle (SRT), text, MS Word, and XML. Pricing may also be impacted by desired format.
- **Time stamps.** Time stamps provide helpful timing tags in the transcription to align the text with the audio. Providers vary in offering time stamps after specific time intervals or at paragraph boundaries.
- **Speaker Identification/Changes.** Speaker identification provides a text label in the transcript which identifies speaker changes when the audio comes from multiple speakers. Speakers may be identified generically by *Speaker1* and the like, or some providers offer to label speakers according to their actual names (which may be determined from the audio or which the user must provide when placing the order).
- **Verbatim.** Providers typically produce *clean verbatim* excluding filled pauses and dysfluency (speech repairs) (Lease, Johnson, and Charniak 2006) from transcripts (e.g., as done in copy-editing transcripts for print to improve readability (Jones et al. 2003)). Some providers also offer *true verbatim*, typically at a surcharge, which include word for word everything said in the transcript, with no such copy-editing. This may be valuable to researchers interested in analyzing these phenomena, and for training/evaluating automatic systems in simulation of perfect ASR.

- **Turnaround Time.** Platforms often offer tiered pricing for how long users should expect to wait before receiving transcripts. Many platforms offer a standard turnaround time guarantee, as well as faster turnaround time options for a surcharge. Interestingly, some platforms provide a channel for the user to negotiate directly with the workers themselves.
- **Difficult Audio Surcharge.** Difficult audio, arising from poor recording equipment, background noise, heavy accents, simultaneously talking, etc., often requires a surcharge or involves tiered pricing for higher quality or faster turnaround time. Determining whether or not audio is difficult is somewhat subjective, may vary by provider, and may be difficult to determine without provider guidance.
- **Distinguishing Features.** Other distinguishing features of each provider are discussed when the provider is first introduced. For example, a provider may offer "interactive transcripts" which digitally link the transcript and audio to continuously align text and audio, rather than providing only incremental time stamps and no digital linkage.

Qualitative Evaluation

We analyzed eight transcription providers: 1-888-Type-It-Up, 3Play Media, CastingWords, Rev, Quicktate, SpeakerText, TranscribeMe, and Transcription Hub. These providers were chosen based on informal web searches and reading about online offerings. As oDesk is a general purpose online contracting site, we do not assess it here along with transcription providers.

1-888-Type-It-Up

Type-It-Up (`www.1888typeitup.com`), formerly Verbal Fusion (VF), does not split audio into chunks for different transcribers, but rather has one person produce the transcript and a more senior transcriber review it. It is the only provider we saw offering a 99.9% guaranteed accuracy option, and one of the few providers with an entirely US-based workforce. Transcript turnaround is also offered on weekends and holidays for urgent needs. Technical/specialized transcription is offered at a premium. They offer a 10 minute free trial and a flexible, special instructions request box.

Base Price & Accuracy: Three price tiers include "draft" at \$1/min, while a double-checked second tier with guaranteed 99% accuracy is offered at \$2/min. For \$3/min, 99.9% accuracy is guaranteed. Lower prices are also available for bulk orders but require negotiation with the platform representative.

Transcript Format: Formats include: HTML, PDF, RTF, text, and MS Word. More complicated formats, such as XML, require a \$.50/min surcharge.

Time Stamps: Time stamps are offered at varying intervals/pricing: every 15 min (\$.25/min), every 10 min (\$.30/min), 5 min (\$.35/min), 1-4 min (\$.50/min), 30s (\$1/min), 20s (\$1.5/min), 15s (\$2/min).

Speaker Identification/Changes: Speaker changes can be noted by line breaks and/or bold/unbold speaker names. For a surcharge, depending on the recorded voice format (interview, presentation, etc.), speakers can be named.

Verbatim: Clean verbatim is standard, with true verbatim offered at at a surcharge of \$.50/min.

Turnaround time: 7-10 (business) days. Tiered time/surcharge options for faster turnaround include: within 5 (business) days (\$.25/min), 4 days (\$.50/min), 3 days (\$1/min), 2 days (\$2/min), and 1 day (\$3/min). Additional pricing options include delivery within 24 hours, same day, on Saturday/Sunday, or on holidays.

Difficult Audio: Heavily-accented speech incurs surcharges ranging from \$.50-\$1/min.

Summary: Flexible features include time stamps, speaker identification, transcript formats, turnarounds, special instructions, and accuracy guarantees. US-based workforce may mean higher English transcript quality at higher cost. However, options are somewhat pricey, including a surcharge for heavily-accented speech.

3Play Media

3Play Media (`www.3playmedia.com`) utilizes hybrid transcription, with 75% automatically transcribed and the rest manually. This perhaps explains why their standard turnaround time is relatively quick. The platform is distinguished by not charging for time stamps and speaker identification, and by offering an "interactive transcript" (defined earlier). They offer technical/specialized transcription. A 10 minute free trial is offered, and users can submit special instructions/vocabulary to transcribers.

Base Price: The base price is \$2.50/min with bulk pricing options for prepaid hours: \$2.35/min (100-249 hours); \$2.30/min (250-499 hours); \$2.25/min (500-999 hours); and \$2.15/min (1000 or more hours).

Accuracy: 99% guaranteed. A "flawless" transcript, with accuracy up to 99.9%, is available at a premium.

Transcript Format: HTML, JSON, PDF, text, MS Word, and XML.

Time Stamps: Freely available at any time interval.

Speaker Identification/Changes: Custom speaker identifications and changes available.

Verbatim: Clean verbatim is default, with an option to include editing flags in the transcript or let transcribers replace them with best guesses.

Turnaround time: 4 (business) days is standard, with options including: 2 days (\$.75/min), 1 day (\$1.50/min), or 8 hours (\$2.50/min).

Difficult Audio Surcharge: \$1/min surcharge.

Summary: Time stamps and speaker identifications are: flexible and free of charge. A large variety of transcript formats is offered. Can handle technical/special transcription. One of few platforms that is US based, with a corresponding fairly expensive base price.

Casting Words

Casting Words (`castingwords.com`) offers live tracking, allowing users to view transcription progress. Technical/special transcription is offered.

Base Price: $1/min.

Accuracy: No accuracy guarantee is offered, though quality transcripts are informally promised.

Transcript Format: Formats include HTML, RTF, text, and MS Word.

Time Stamps: Time stamps at speaker changes are offered at $.10/min. If individual speech is long enough, time stamps are inserted at paragraph breaks.

Speaker Identification/Changes: Speakers are labeled by name.

Verbatim: Clean verbatim is default, true verbatim requires a surcharge.

Turnaround Time: Standard turnaround time at the base price is not specified, with options including: within 6 days ($1.50/min), or within a day ($2.50/min).

Difficult Audio: There is no surcharge. With manageable difficulty, transcripts may be delayed; with high difficulty, transcription will be aborted and the user refunded their money minus cost.

Summary: Fairly low base price with live transcription progress tracking and opportunity for special instructions. There is no difficult audio quality surcharge, but also no accuracy guarantee.

Quicktate

Quicktate (`quicktate.com`) prices transcription by words rather than minutes, which complicates price comparisons with other providers. A free trial is offered ($5 credit), and transcripts are available in Evernote format for interaction. Technical/specialized transcription needs are offered (e.g., medical transcription has different pricing). Users can choose to have a global transcriber, or for a premium, a U.S. transcriber.

Base Price: A global transcriber under "pay as you go" costs $0.0325/word. With 175 prepaid words/week ($1.25), cost per word falls significantly to $0.007/word (with overruns at $0.012/word). With 750 prepaid words per week ($5) package, cost is $0.0067/word (with overruns at $0.011/word). U.S. transcribers are more expensive; the pay as you go option is $0.02/word. With 160 prepaid words per week ($1.99), cost falls to $0.022/word (with overruns at $0.0325/word). With 775 prepaid words ($6.99), cost falls to $0.009 (with overruns at $0.024/word).

When longer than 10 minutes, cost can be effectively priced at the overrage rate. As an upper-bound on speaking rate, one might assume 160 words/minute, the fastest speed reported in (Williams 1998). At $0.011/word (international overrun with largest prepaid package), 120-160 words/min would equate to $1.32-1.76/min. Similarly, at $0.024/word (U.S. transcriber overrun with largest prepaid package), 120-160 words/min would equate to $2.88-3.84/min.

Transcript Format: Transcripts can be sent via email in text (not as a text file attachment). Transcripts are also available in Evernote format.

Time Stamps: not available.

Speaker Identification/Changes: Speakers are labeled by their names.

Verbatim: Clean verbatim only is offered.

Turnaround Time: There is no specific guarantee, though they offer fast turnaround times.

Summary: Interactive transcripts are offered via Evernote, as well as technical/special transcription needs and preference for U.S. transcribers. However, pricing is complicated, format is limited, and no specific accuracy nor turnaround guarantee is offered.

Rev

Similar to Type-It-Up, Rev (`www.rev.com/transcription`) assigns audio to a single transcriber, rather than splitting up the audio for multiple transcribers, with senior transcriber review. Despite this lack of parallelization, Rev offers 48-hour turnaround at 98% accuracy for multi-speaker audio at only $1/min. They also accept special instructions and offer technical/specialized transcription.

Base Price: $1/min.

Accuracy: 98% guaranteed.

Transcript Format: MS Word.

Time Stamps: $.25/min surcharge, every 2 minutes.

Speaker Identification/Changes: Speakers are labeled as "Speaker 1" and so on.

Verbatim: They exclude filled pause terms (fillers) such as "uhh" and "umm" but do not correct disfluency. Fillers can be kept for a $.25/min surcharge.

Turnaround Time: Standard turnaround time is within 48 hours for non-difficult, under 60 minute audio. Longer and/or difficult audio may take longer.

Difficult Audio: There is no surcharge, though transcripts may be delayed.

Summary: Rev offers competitive pricing, a 98% accuracy guarantee, relatively quick turnaround, no difficult audio surcharge, and a special instructions box. However, there is only one format option and limited speaker identification options.

SpeakerText

While SpeakerText (`speakertext.com`) seems to be focused on video transcription, for the same price it transcribes audio as well. A free 5 minute trial is offered.

Base Price: $2/min with bulk pricing options for prepaid hours: $1.20/min (0-49 hours), $1.12/min (50-99 hours), $1.08/min (100-249 hours), $1.03/min (250-499) hours, and $0.98/min (500+ hours).

Accuracy: Accuracy is not specified, though the platform offers "Guaranteed Accuracy".

Transcript Format: DFXP, HTML, SRT, text, and XML.

Time Stamps: Time stamps are offered for XML format only, anywhere from sub-second to 4s.

Speaker Identification/Changes: Not offered.

Verbatim: Clean verbatim only.

Turnaround Time: No guarantee.

Summary: The base price is somewhat expensive relative to available features offered, though cost decreases considerably with bulk pricing. Speaker identification/changes are not marked, and there is no guaranteed turnaround time or specified accuracy.

TranscribeMe

Similar to 3Play Media, TranscribeMe (`transcribeme.com`) utilizes hybrid transcription. They digitally enhance audio, perform ASR, and segment audio and transcripts for the transcribers to work on. Transcripts are double checked by a quality assurance team. Specialized professional transcribers are available for special requirements. This service also reports being mobile friendly, with an app for iPhone (and one promised for Android) that allows ordering transcripts "on the go".

Base Price: $1/min for single speaker, $2/min for multiple speakers. Bulk pricing and monthly subscription plans offer reduced prices.

Accuracy: 98% guaranteed accuracy.

Transcript Format: HTML, PDF, and MS Word.

Time Stamps: Provided at each speaker change or every 2 minutes with a single speaker.

Speaker Identification/Changes: Speakers are labeled as s1, s2, etc.

Verbatim: Clean verbatim by default, though filler words can be retained by request. Disfluency is not corrected; audio is transcribed as is.

Turnaround Time: 48 hours typically; audio that cannot be digitally enhanced may take longer.

Summary: 98% guaranteed accuracy with relatively quick turnaround time. Time stamps and speaker labels are free, with no difficult audio surcharge. Flexible transcription formats. In addition to free trial, offers mobile transcript orders. Speakers cannot be labeled by name.

Transcription Hub

Transcription Hub (`www.transcriptionhub.com`) offers technical/specialized transcriptions and accepts special instructions. A 5 minute free trial is offered.

Base Price: $.75/min.

Transcript Format: MS Word, though a different format can be requested free of charge.

Time Stamps: For an additional $.15/min, time stamps are inserted every 2 minutes.

Speaker Identification/Changes: Speakers can be identified generically or by user-supplied labels.

Verbatim: Default verbatim is clean, though true verbatim is available, or other custom verbatim styles.

Turnaround Time: 15 days is standard. Options include: 5 days (additional $.20/min), 2 days (additional $1.25/min), or 1 day (additional $1.70/min).

Difficult Audio Surcharge: $.75/min.

Summary: Lowest base price, but with no accuracy guarantee and longest base turnaround time. Flexible transcript format. Speaker identification is also flexible and free of charge. Accepts special instructions.

Discussion

Every provider has relative strengths and weaknesses, of varying importance to different users. Some may have a low budget and have no need for near perfect accuracy. Others may only go for best guarantees. In a low pricing category, TranscribeMe and Rev may be most similar. Both platforms offer 98% accuracy; both offer time stamps and speaker identification in similar manner; both have relatively quick standard turnaround. However, Rev offers same $1/min transcription price for multiple speakers, while TranscribeMe only offers such price for a single speaker transcription (and double cost for multiple speakers, unless you order in bulk). However, while time stamps in Rev are $.25/min, TranscribeMe time stamps as well as speaker changes are free. TranscribeMe offers relatively larger transcript format variety, while Rev offers only MS Word format.

Perhaps the most similar service to Transcription Hub would be CastingWords. Transcription Hub offers the lowest base transcription price of $.75/min, while on the CastingWords' price is $1/min. However CastingWords time stamps are 5 cent cheaper. CastingWords does not charge for difficult audio quality, while Transcription Hub charges $.75/min for difficult audio. CastingWords also emphasizes that their transcription process and quality control is made strictly by humans. This might be suggestive of better accuracy. Transcription Hub, however, offers more dynamic transcription formatting and speaker identification.

Quantitative Evaluation

After informally evaluating pilot transcripts from all providers, we selected 1-888-Type-It-Up, CastingWords, and Transcription Hub for this quantitative evaluation due to low cost, though 1-888-Type-It-Up is the most expensive due to surcharges on multi-speaker and difficult speech. CastingWords and TrancribeMe have the same base price ($1/min). We also hired three oDesk transcribers at various price points for further comparative analysis; listed prices in oDesk varied dramatically (roughly$3-$250/hr). We measure WER over the first 10-minutes from each of six interviews.

oDesk (`www.odesk.com`) provides a general purpose online labor marketplace whose focus on specialized and higher-skilled forms of labor (i.e., contractors) distinguishes it from relatively unskilled work often posted to Amazon's Mechanical Turk. We posted jobs, and workers bid on them. Once accepted, work can be monitor via oDesk's "Work Diary" system, and workers can update the status of their work (whether they are transcribing, proofreading, etc). Once finished, oDesk collects collects 10 percent of the total payment. Regarding pricing, note that oDesk prices per hour of work, whereas transcription services price per hour of audio.

Data Preparation

Spontaneous speech interviews (audio and reference transcripts) used in our evaluation come from the USC-

SFI MALACH English corpus (LDC2012S05[1]). Speech data was collected under a wide variety of conditions ranging from quiet to noisy (e.g., airplane overflights, wind noise, background conversations and highway noise) on tapes, and then the tapes were digitized and compressed into MP3 format. Each LDC interview includes a 30 minute section taken from a longer full interview. Due to the way the original interviews were arranged on the tapes, some interviews were clipped and had a duration of less than 30 minutes. LDC transcripts are in XML format, with a DTD file for validation.

We asked service providers to transcribe the first 10 minutes from of interviews 00017, 00038, 00042, 00058, 00740 and 13078. On oDesk, we first filtered out workers in the market with less experience and low testing scores according to their profiles, and finally requested 3 workers with different prices. OD1 ($5.56/hour) was from Philippines. OD2 ($11.11/hour) indicated he was an English teacher in Australia. OD3 ($13.89/hour) claimed to be a US transcriber. Note that the hourly rates of these selected oDesk transcribers are far below the service provider rates (e.g., $1/min = $60/hour). After they accepted our contracts, we assigned interviews 00017 and 00038 to OD1, 00740 and 13078 to OD2, and 00042 and 00058 to OD3.

Table 1 prices reflect the cheapest minute-based base price rate from the qualitative evaluation section (our evaluation intentionally focused on the budget-oriented user without need for more expensive accuracy guarantees). However, while we expected 1-888-Type-It-Up to charge an additional $0.5/min for multiple speakers, we were surprised to be billed afteward for an additional $0.50/min difficult audio fee for all interviews, and for interview 00042 a higher $1/min, reporting heavily accented speech. We include these charges in calculating the hourly rate shown. However, because it does not bear on transcript accuracy, we exclude an unanticipated $3/min we were additionally charged for requesting XML format. For comparison, we note that Transcription Hub did not notify or charge us for difficult audio, and CastingWords notified us of 00038 having difficult audio, giving us the option of whether we wanted to pay a surcharge for expedited processing.

For alignment with LDC reference transcripts and WER measurement, we used Sphinx4[2] with a minor change to its `NISTAlign` class to ignore final punctuation (e.g. considering "book." and "book" to match).

Pre-processing

The `NISTAlign` class takes raw text as input; not only did reference LDC transcripts need to be pre-processed, but so did each variant transcript format from crowdsourced transcribers. While LDC data included a DTD file for XML files, there were still some variances, especially on the tagging of speaker switching. For example, 00038 put the speaker id in the speech text, like <spk2>; while in 13078, the speaker was represented as an element, and during the Turn switching it was quoted as an attribute in the Turn element. We filtered out speaker changes for WER evaluation, but we did not remove the tags which identify the background, noise and description text like "unintelligible" and "silence".

CastingWords (CW) provided results in text format, but we needed to filter out speaker labels and time stamps. Transcription Hub (TH) transcripts in MS Word format labeled speakers and time stamps, which we removed manually. 1-888-Type-It-Up(VF) used its own data format: an XML file but more like an HTML format. oDesk output in MS Word format, had different document layouts from each of OD1-3. And so on.

Another significant difference between reference and crowdsourced transcripts was the representation of numbers, including date, time, and other quantitative information. For example, in most SPs' transcripts, years such as 1936 were in digital format, while in LDC, it was expressed as "nineteen thirty six". We manually changed all numbers into such text format.

In hindsight, we should have determined exactly where each 10-minute segment ended in the LDC reference transcript, and used this same reference transcript for evaluation across providers. Instead, we adopted a more complicated sliding window technique to optimally align each crowdsourced transcript with the entire reference transcript. This introduced some noise in our WER scoring, which we quantify later in detail.

WER Accuracy and Error Analysis

After alignment, we found WERs shown in Table 1. The first number in each cell indicates our initial WER results, approximately 20-30%. This was very surprising, leading us to perform detailed manual error analysis and identify a variety of ways in which these initial WER results were artificially high. In reviewing differences between aligned LDC and crowdsourced transcripts, we identified the following groups:

- **Background**: noise, unclear words marked in reference text, also some emotion words like LAUGH, CRY, as well as COUGH
- **Partial words**: ending with "-" in reference text
- **RefError**: rare errors in the LDC transcripts.
- **Fillers**: disfluent words like "uh", "em", as well as phrases like "you know", "you see", "I mean"
- **Repetition**: repetition reflecting disfluency, e.g., "to the to the", was preserved by LDC but removed by providers for "clean verbatim" transcripts.
- **Repairs**: disfluency where the speaker self-corrected was preserved by LDC but not service providers.
- **Spelling**: orthographic differences such as "A-L-I-C-E" vs. "A L I C E", "every day" vs. "everyday", "old fashioned" vs. "old-fashioned", "that is" vs. "that's", British English vs. US English, use of ellipses between words without spacing, etc.

[1] `http://www.ldc.upenn.edu`

[2] `cmusphinx.sourceforge.net/sphinx4/`

Service Provider with Price Rate	Interview Transcripts						Avg. WER by Service Provider	Accuracy/$ Ratio
	00017	00038	00042	00058	00740	13078		
CastingWords (CW) ($60/hr per audio)	31.356 9.707 (0.154)	33.198 17.005 (0.881)	23.273 14.885 (0.822)	28.624 15.976 (0.814)	16.833 11.643 (1.996)	26.452 14.129 (2.119)	26.623 13.891 (1.131)	1.435
Transcription Hub (TH) ($45/hr per audio)	30.233 **8.450** (0.155)	34.628 18.405 (1.022)	29.129 18.308 (1.221)	33.433 18.399 (1.197)	18.071 9.036 (2.495)	28.874 14.588 (2.116)	29.061 14.531 (1.368)	1.899
1-888-Type-It-Up (VF) (avg $125/hr per audio)	28.874 9.524 (0.151)	26.819 **11.051** (1.011)	18.543 **11.175** (0.662)	23.921 **11.658** (0.454)	12.559 **6.212** (2.296)	24.072 **10.977** (2.120)	22.465 **10.099** (1.116)	0.719
oDesk Worker1 (OD1) ($5.56/hr per work)	31.144 10.510 (0.155)	29.787 16.884 (1.098)	-	-	-	-	30.465 13.697 (0.626)	**15.522**
oDesk Worker2 (OD2) ($11.11/hr per work)	-	-	-	-	20.066 12.226 (2.591)	28.495 14.973 (2.597)	24.281 13.600 (2.594)	7.777
oDesk Worker3 (OD3) ($13.89/hr per work)	-	-	34.415 22.545 (1.623)	37.983 19.228 (1.734)	-	-	36.199 20.886 (1.678)	5.696
Avg. by Interview	30.402 9.548 (0.154)	31.108 15.836 (1.003)	26.340 16.728 (1.082)	30.990 16.315 (1.050)	16.883 9.779 (2.345)	26.973 13.667 (2.238)	28.183 14.451 (1.419)	-

Table 1: WER between LDC transcriptions (gold) and results from SPs. Each cell has 3 numbers: the original WER, the reduced WER, and the WER introduced by name errors (braced by parenthesis). The right-most column is accuracy vs. cost ratio based on the reduced WER, computed by (100-WER)/hourly price. The lowest reduced WER observed for each interview is marked in bold, and the highest WER observed for each interview is underlined.

- **Post-Error**: errors in converting date and number from numeric to text form.
- **Alignment**: alignment limitations from Sphinx or in our use of it, typically occurring at the end of alignment string (e.g., our sliding window method).
- **Named-Entity**: unfamiliar names were difficult for transcribers, as expected.
- **Miscellaneous**: other errors, typically true errors made by crowd transcribers.

Many of these "errors" should be excluded from WER measurement. Fillers, Repetitions, and Repairs all reflect disfluency, common in spontaneous speech as speakers form their utterances on the fly (Lease, Johnson, and Charniak 2006). Service providers explicitly remove these for "clean verbatim", as well as Partial Words, unless directed otherwise. Post-Error and Alignment issues reflect limitations of our own processing. Background and RefError reflect artifacts of the LDC reference text. Finally, orthographic variants should also be accepted without specific guidelines specifying orthographic transcription norms.

This leaves crowdsourced transcribers responsible for Miscellaneous errors, and reasonably responsible for Named-Entity errors (though we expect such is difficult and likely could be further improved through specialized transcription requests).

Returning to Table 1, the 2nd number in each cell is the "reduced WER" which only includes Miscellaneous and Named-Entity errors. The 3rd number in the cell is the WER for Named-Entity errors only. Compared with the initial WER, each cell improved by approximately 10-20%. We observe most reduction in measurement error for interview 00017, around 20%. If we further look into the alignment, e.g. CastingWords' transcript, we see 407 deletions, insertions and substitutions. The top-3 errors are 53 Background, 83 Filler, and 69 Partial, which constitute of nearly half of the 407 errors.

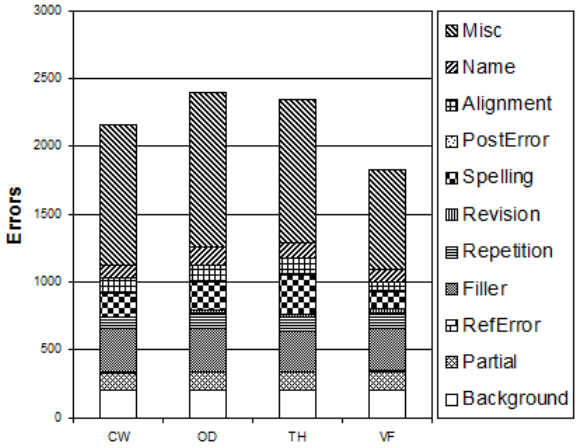

Figure 1: Errors in CW, OD, TH and VF. True errors (Miscellaneous and Named-Entity) occupied half of the total errors. Filler is the most among those false errors.

Figure 1 break-downs error types across all alignments for each service provider. True errors ("Miscellaneous") with Named-Entity errors provide 50% of the total errors for all SPs. Fillers dominate initial measurement errors, followed by differences in orthography and background noise captured in the reference transcripts.

VF shows the fewest total errors, yielding the lowest WER in alignment. The 3 OD transcribers generated the most errors (though we note again their far lower

price). Because we did not ask every OD transcriber to transcribe every interview, our analysis of their relative error rates must allow for some interviews being more difficult than others. That said, OD2 appears to have performed much better than OD1 and OD3. OD2 had approximate 300 errors per transcript, while OD1 and OD3 had at least 400 errors for each alignment.

Discussion

Usually, when users talk about crowd transcription, what they mostly care about is the quality and the cost, especially, the price rate. Relative to price, Transcript Hub appears preferable since each dollar buys approximately (100-14.531)% WER / \$45 = 1.899 accuracy/\$, vs. CastingWords' 1.453 and Type-It-Up's 0.719, as described in the right-most column of Table 1. Although 1-888-Type-It-Up achieved the most accurate transcriptions, it has the lowest margin of cost where each dollar only contributes 0.719 accuracy. Similarly, though OD2 is the most efficient worker in oDesk as mentioned above, his accuracy gain is still lower than OD1 since OD1 requested only half of OD2's price and provided closer WER (13.697 vs. 13.600). So by this simple analysis, cheaper oDesk is a far better deal.

However, this picture is incomplete. CW, TH and VF's price rates (\$45-60/hr) are much higher than those of oDesk (\$5.56-13.89/hr), because they are commercial companies who take the responsibility of quality and delivery, as well as risk management. They not only ensure the transcription is completed and on schedule, but they assume the management costs to accomplish this which are otherwise born by an oDesk or AMT customer (though presumably lower on oDesk than AMT).

Crowdsourcing studies have rarely accounted for such management costs when reporting savings, though such real costs of crowdsourcing "in the wild" could easily result in higher total costs vs. traditional transcription practices. When we recruited oDesk transcribers, while their price rates were much lower, we had to carefully communicate with individual workers to negotiate price, clarify requirements, and monitor work. We also had to take the risk that the workers might miss the target date. Overall, crowdsourcing research would benefit tremendously by finding ways to assess these tradeoffs more transparently and holistically when evaluating and motivating alternative practices to the community.

Acknowledgments. This research was supported in part by DARPA Award N66001-12-1-4256, and IMLS grant RE-04-13-0042-13. Any opinions, findings, and conclusions or recommendations expressed by the authors do not express the views of any of the supporting funding agencies.

References

Adar, E. 2011. Why I Hate Mechanical Turk Research (and Workshops). In *CHI Human Computation Workshop*.

Audhkhasi, K.; Georgiou, P.; and Narayanan, S. S. 2011. Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics. In *IEEE ICASSP*, 4980–4983.

Byrne, W.; Doermann, D.; Franz, M.; Gustman, S.; Hajic, J.; Oard, D.; Picheny, M.; Psutka, J.; Ramabhadran, B.; Soergel, D.; et al. 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Speech and Audio Processing* 12(4):420–435.

Evanini, K.; Higgins, D.; and Zechner, K. 2010. Using amazon mechanical turk for transcription of non-native speech. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 53–56.

Gruenstein, E.; Mcgraw, I.; and Sutherl, A. 2009. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *the Speech and Language Technology in Education (SLaTE) Workshop*.

Jones, D. A.; Wolf, F.; Gibson, E.; Williams, E.; Fedorenko, E.; Reynolds, D. A.; and Zissman, M. A. 2003. Measuring the readability of automatic speech-to-text transcripts. In *International Speech Conference (InterSpeech)*.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The Future of Crowd Work. In *Proc. CSCW*.

Lasecki, W.; Miller, C.; Sadilek, A.; Abumoussa, A.; Borrello, D.; Kushalnagar, R.; and Bigham, J. 2012. Real-time captioning by groups of non-experts. In *ACM symposium on User interface software and technology (UIST)*, 23–34.

Lease, M.; Johnson, M.; and Charniak, E. 2006. Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech and Language Processing* 14(5):1566–1573.

Lee, C., and Glass, J. 2011. A transcription task for crowdsourcing with automatic quality control. In *12th International Speech Conference (InterSpeech)*, 3041–3044.

Marge, M.; Banerjee, S.; and Rudnicky, A. 2010a. Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 99–107.

Marge, M.; Banerjee, S.; and Rudnicky, A. I. 2010b. Using the amazon mechanical turk for transcription of spoken language. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 5270–5273. IEEE.

Novotney, S., and Callison-Burch, C. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Proceedings of NAACL-HLT*, 207–215.

Parent, G., and Eskenazi, M. 2010. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT)*, 312–317.

Pecina, P.; Hoffmannová, P.; Jones, G. J.; Zhang, Y.; and Oard, D. W. 2008. Overview of the clef-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval*. Springer. 674–686.

Vakharia, D., and Lease, M. 2013. Beyond AMT: An Analysis of Crowd Work Platforms. Technical report, University of Texas at Austin. arXiv:1310.1672.

Williams, J. D.; Melamed, I. D.; Alonso, T.; Hollister, B.; and Wilpon, J. 2011. Crowd-sourcing for difficult transcription of speech. In *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 535–540.

Williams, J. R. 1998. Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1447–1451.