Don't Count on ASR to Transcribe for You: Breaking Bias with Two Crowds

Michael Levit, Yan Huang, Shuangyu Chang, Yifan Gong

Microsoft, USA

{mlevit|yanhuang|shchang|ygong}@microsoft.com

Abstract

A crowdsourcing approach for collecting high-quality speech transcriptions is presented. The approach addresses typical weakness of traditional semi-supervised transcription strategies that show ASR hypotheses to transcribers to help them cope with unclear or ambiguous audio and speed up transcriptions. We explain how the traditional methods introduce bias into transcriptions that make it difficult to objectively measure system improvements against existing baselines, and suggest a two-stage crowdsourcing alternative that, first, iteratively collects transcription hypotheses and, then, asks a different crowd to pick the best of them. We show that this alternative not only outperforms the traditional method in a side-by-side comparison, but it also leads to ASR improvements due to superior quality of acoustic and language models trained on the transcribed data.
Index Terms: Unbiased speech transcription, crowdsourcing, acoustic and language modeling

1. Introduction

Recent years have witnessed a surge in speech recognition quality with automated systems coming close to, or even exceeding, human performance on established benchmarks [1, 2, 3]. With word error rates (WERs) so low, ability to measure them reliably gains new importance. However, accurate transcription of speech is difficult due to its highly ambiguous nature.

Using ASR results as hints to help guide transcription is commonly practiced in the ASR industry for model development and evaluation. On the other hand, our experiments showed that transcribers tend to adopt the ASR result instead of authoring a new transcription not just in truly ambiguous cases but sometimes even when the ASR result is wrong. This has two implications: first, inaccurate model evaluation due to transcription bias towards ASR result generated by the current production system; second, sub-optimal acoustic and language models trained on erroneous transcriptions.

To address this issue, we propose an iterative two-stage crowdsourcing strategy that takes advantage of the automated system's expertise while minimizing the bias. This is achieved by only exposing ASR transcriptions in ambiguous cases and blending them with other human-generated alternatives. The approach is iterative and saves time and costs with a mechanism that dynamically determines the number of opinions needed for each utterance based on the perceived difficulty level. In a human side-by-side transcription quality comparison study, the proposed transcription approach notably outperforms the traditional method. Furthermore, it leads to improved acoustic and language models with 3-4% WER reduction.

The rest of this paper is organized as follows: Section 2 discusses transcription bias and its impact on ASR system evaluation and model development; Sections 3 and 4 introduce and refine the two-stage iterative crowdsourcing framework and Section 5 describes experiments on subjective evaluation as well as acoustic and language model training.

2. Transcription Bias

Manual transcriptions of speech are commonly used to train and evaluate stochastic models for speech recognition. Speech scientists like to act on the assumption that these transcriptions are unconditionally trustworthy (and even dub them "gold references"). Validity of this assumption is reasonable for certain applications such as read texts with available references, but in other scenarios such as third-party transcriptions of spontaneous speech or voice search queries, it is rather misconceived. Indeed, human transcribers, and in particular unskilled crowd judges, are prone to problems such as insufficient domain expertise (e.g. lacking familiarity with certain named entities) and avoidable mistakes (e.g. typos). Besides, the task of third-party transcriptions often grapples with true ambiguity (*"call Chris"* or *"call Kris"*?) due to lacking discourse context and/or not being able to get into the mind of the utterance originator. In some cases, speakers could even have difficulties transcribing their own speech (*"you are right"* or *"you're right"*?).

By breaking such ties and offering transcription alternatives that human transcribers might not have enough expertise to guess on their own, ASR-assisted transcription setups reduce transcription time and improve quality [4, 5]. However, we argue that such setups can misguide transcribers: by virtue of being primed with generally reliable hints, transcribers get habituated to seeing correct ASR hypotheses and trust these more than they actually deserve, especially when acoustic conditions are adverse or homophone alternatives are possible. An anecdotal evidence collected internally illustrates the above problem. One member of our team was asked to listen to a large number of voice-search queries and short messages accompanied by ASR recognition hypotheses and select 100 clear and unambiguous examples. After that, another team member was asked to transcribe these without looking at the ASR results. We then compared the results and observed that more than a quarter of transcriptions had mismatches that we attributed to the priming effect that ASR hypotheses exercised on the first colleague. Not only does this evidence undermine our faith in measured word error rates, the above semi-supervised approach can be dangerously misleading when one tries to develop improved ASR models and compare their accuracy against baselines, in case those baselines were used to assist transcribers. Our experiments showed that under realistic conditions, while testing new ASR deployment candidates, WER reductions of up to 10% relative can be entirely masked by the transcriber bias due to exposure to the baseline ASR's recognition hypotheses.

In addition, in regard to transcribed training material, we collected evidence that such bias could adversely affect the quality of acoustic and language models trained on it, due to reinforcement of certain recognition mistakes. Thus, we find ourselves faced with the dilemma of choosing between bias and domain expertise.

3. Two-stage Iterative Crowdsourcing

Crowdsourcing has been demonstrated to be an effective tool for many annotation and transcription tasks [6, 7]. While various manual and automatic techniques have been offered to improve general quality [8, 9, 10, 11], the bias problem can be expected to affect behavior of unskilled judges even more strongly than professional transcribers. In order to enjoy the advantage of having ASR in the loop without falling victim to the priming bias, we have conceived a two stage transcription pipeline that, for each utterance, first collects unassisted lexicalized transcriptions from human judges and one or more ASR systems, and then lets a different crowd select the best one from a randomized list of alternatives. Judges in each stage are supported by a variety of tools from search engines with integrated spellers to white lists and rule-based format- and syntax checkers.

Furthermore, to save costs, we do not ask for a fixed number of judgments at once, but rather request one opinion at a time. By comparing next opinion to the already available ones (in the beginning, only one or several ASR hypotheses are available) and measuring degree of disagreement among them, the decision is made as to how to proceed next for this utterance.

An illustration of the above algorithm is shown in Figure 1. The pipeline starts by providing each of the utterances with two or more alternative automatic recognition results (1). In the spirit of [12], we prefer recognition setups that are different from each other but of comparable quality. In practice, when deciding whether to deploy a new ASR instead of the production ASR, recognition hypotheses from both systems can be used. During the first transcription stage, one judge at a time listens to the audio and provides lexical transcription for it (2). Judges are allowed (though not encouraged) to mark utterances as too difficult to transcribe. If a reliable consensus regarding utterance's difficulty is achieved, the utterance is dismissed. As new transcription hypotheses arrive for each utterance, we compile them into a single cumulative distribution (3) assigning each hypotheses h a probability $p(h)$ derived from recognition confidences (for ASR) and judge reliability estimates. We then look at the value of this distribution's normalized entropy

$$E(p) = \frac{1}{\log J} \sum_h p(h) \log p(h)$$

where J is the total number of judges so far, including ASR systems. If this entropy is low $E(p) < \theta_1$, then the highest probability hypothesis in the set is promoted to be the sole output of the system (4a). If the entropy is high $E(p) > \theta_2$ but the maximum number of iteration (we set it to 5) has not yet been exceeded, next iteration for this utterance is initiated (4c). Otherwise, the collected hypotheses (automatic and human generated) are ranked by their probabilities and the N highest scoring of them passed to the second stage for selection (4b). An even smaller subset of the $N' \leq N$ highest scoring hypotheses, along with the corresponding probabilities, is used to seed the distribution of transcription alternatives for this stage. We now ask one judge (of the second crowd) at a time to listen to the audio and pick the best of the listed alternative transcription hypotheses (5) or even provide a new one, if deemed necessary. The normalized entropy of the obtained cumulative distribution (6) is, again, used to decide how to proceed. In this scenario, the possible outcomes are: to continue with iterations (7b) or to accept one or several highest-scoring transcription hypotheses as the final outcome (7a) along with the associated weights as measures of their reliability. Note that at all times, the order in which alternatives are shown to the judges in the second stage is randomized to avoid bias towards the first item on the list due to limited attention span. A number of heuristics is employed to make decisions more plausible. For instance, we require that no hypothesis is returned unless at least one human judge produced it in the first stage or selected it in the second stage. Thresholds θ_i were set empirically: $\theta_1 = 0.2$ and $\theta_2 = 0.3$.

4. Scoring Hypotheses and Judges

The algorithm above relies on probability distributions for transcription hypotheses in the presence of other alternatives. Instrumental in producing these distributions are ratings of individual judges that are accumulated and normalized into hypotheses' posteriors. Specifically, if we denote the momentary rating of judge j as $R(j)$, posterior probability of hypothesis h can be computed as:

$$p(h) = \frac{\sum_{j:h_j=h} R(j)}{\sum_j R(j)}$$

For the ASR "judges", this rating is just the normalized recognition confidence. For human judges, ratings are based on several factors such as judge's prior rating before the current transcription job and his agreement with peers. Specifically for the first stage, let $\#_t(j)$ be the number of hypotheses that judge j has provided for a given corpus by the end of iteration t. Furthermore, let $M_t(j)$ be the number of utterances for which j's transcription hypotheses agreed with the majority of participating judges, and $S_t(j)$ the number of utterances for which no other judge provided the same hypothesis as judge j, but there was another hypothesis that a majority of judges agreed upon. Then, j's rating $Q_t(j)$ based on the corpus at hand can be updated as:

$$Q_t(j) = \frac{\#_t(j) + M_t(j) - S_t(j)}{2\#_t(j)}$$

and the overall rating $R_t(j)$ after t iterations is obtained via interpolation with prior rating $P(j)$:

$$R_t(j) = \lambda_t Q_t(j) + (1 - \lambda_t) P(j)$$

where interpolation coefficient is assumed to be a function of $x := \#_t(j)$ and its fraction to the number of prior judgments that $P(j)$ was based on $y := \#_t(j)/\#(j)$:

$$\lambda_t = k_1 * \left(\frac{1}{1 + e^{-k_2 x}} - 0.5\right)\left(\frac{1}{1 + e^{-k_3 y}} - 0.5\right)$$

with empirically set constants: $k_1 = 2.5$, $k_2 = 0.1$, $k_3 = 2$. In addition, for judges of the first stage, the rating $P(j)$ is updated using $R_T(j)$ (T being the last iteration of the first stage) but also by taking into account the support that his hypotheses received from judges of the second stage. Apart from that last nuance, ratings of second stage judges are computed similarly.

5. Experiments and Results

5.1. Experimental Setup

All of our experiments have been conducted on spoken utterances from the Microsoft Cortana scenario, covering a wide range of domains and topics from chit-chat and voice search to command-and-control and short message dictation. Each utterance has been transcribed twice: once by a randomly selected single professional transcriber with direct access to ASR recognition hypothesis (baseline), and the other time using the transcription method described in Sections 3 and 4 leveraging large

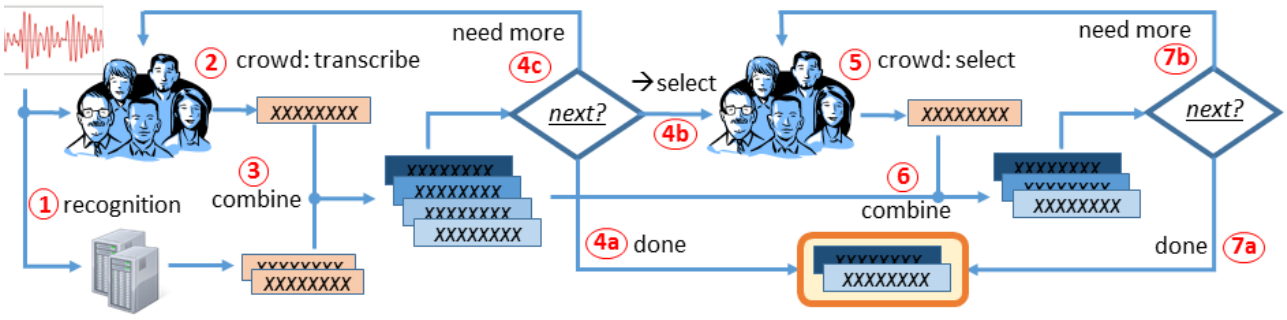

Figure 1: *Iterative two-stage crowd-transcription pipeline.*

crowds of unskilled judges. For the purpose of the investigations in this paper, we only focused on the single highest scoring transcription hypothesis for each utterance produced by the second pipeline. Crowd transcriptions have been carried out using Microsoft UHRS crowdsourcing platform [13]. The platform supports – and our method benefited from – the standard set of quality control measures ranging from qualification tests to occasional probes with and without feedback to the judges.

We ran two groups of evaluation experiments to see whether our transcription pipeline achieves its goal of producing high fidelity, unbiased lexical transcriptions. In the first set of experiments, we looked at plausibility of the obtained transcriptions from three perspectives: asking human experts to pick the better of the two, probing for bias and correlating WERs against the transcriptions to human assessment of recognition results. A small corpus of 1940 Cortana utterances was randomly selected for this purpose.

The second set of experiments investigated the effect of our transcription pipeline on the quality of language and acoustic models trained on it. For this purpose, we have put together a pair of training/test corpora of 130K/8872 utterances transcribed with either the standard transcription pipeline (S1 baseline) or the proposed new transcription pipeline (C2). We would train our language and/or acoustic models either on S1-TRAIN or C2-TRAIN and evaluate them on S1-TEST or C2-TEST.

5.2. Subjective Evaluation

For a subjective comparison of the two competing transcription pipelines we first focused on the 232 of the 1940 utterances (about 12%) that the pipelines produced different results for. An independent group of professional transcribers was asked to perform blind side-by-side comparison and select the best of the two alternatives given audio. In the end, 50% of cases were resolved in favor of the crowdsourcing approach, 36% in favor of the legacy pipeline (adjusted to compensate for automatic data normalization artifacts discovered after the fact), and the rest was deemed of equal quality.

The idea behind our next experiment was that reliable WER estimates would correlate more strongly with perceived recognition quality. To see if this holds in our case, we took recognition results from two ASR systems that were used to seed the crowdsourcing pipeline as explained in step (1) of the algorithm. Their recognition hypotheses were then evaluated side-by-side by several language experts whose task was to assess whether one ASR's hypothesis was much better, somewhat better or about the same as the other. The per-utterance verdicts were converted to numbers (-2, -1, 0, 1, 2) and correlated with the pairwise differences of per-utterance WERs measured against transcriptions of either the baseline or suggested transcription pipeline. On the 496 utterances for which the two ASRs disagreed with each other, the crowdsourcing approach produced Pearson coefficient of 0.76, whereas the baseline pipeline's number was 0.68.

We then looked at the absolute corpus-level WER numbers according to the two transcription methods. For the ASR system that **was not** used to assist S1-transcribers, WER numbers for both transcription pipelines exhibited an impressive agreement (11.38% vs 11.34%). However, for the (better quality) ASR that **was** used to assist these transcribers, we observed a significant WER difference: 8.4% against the baseline transcriptions and 9.4% against transcriptions of the crowdsourcing pipeline. Amounting to more than 10% relative, this difference quantifies the effect of priming bias and illustrates how it is eliminated in our new transcription framework.

Having observed how a crowd of unskilled workers might produce more reliable transcriptions than a single professional transcriber, we naturally inquired how a professional crowd would perform with our pipeline. Somewhat surprisingly, we have observed in a series of preliminary experiments that only an insignificant reduction in the total number of required transcriptions could be achieved. However, the measured WER of a number of ASRs was lower by about 4% relative. In the absence of bias, this is an indirect indicator that crowds of professionals, though more expensive, are capable of providing even better reference transcriptions.

5.3. Effect on Language Model Training

To measure the impact of the proposed transcriptions on language model (LM) training, we trained a pair of 4-gram Kneser-Ney LMs on both versions of the 130K utterance training set: one with the single-opinion professional baseline transcriptions (S1-TRAIN) and the other one with our proposed two-stage crowdsourcing transcriptions (C2-TRAIN). It is understood that by current industry standards,130K utterances (corresponding to 70 hours of audio) is a fairly small training set. Nonetheless, this setup still allows us to perceive the generalizable trend.

Each of these LMs was combined with a pre-trained production-quality acoustic model (AM) to form two competing ASR systems. We evaluated them on the 8872 utterance test set either according to S1-TEST or C2-TEST. Table 1 presents the WER comparison results.

While training an LM on unbiased transcriptions has little effect on the S1-TEST set, testing on C2-TEST appears to give the unbiased transcriptions about 4.2% advantage. Since, following conclusions in Section 5.2, C2 test set can be considered more reliable, we conjecture that our new transcription pipeline

Table 1: *Effect of transcription method on LM training measured in relative WER(%).*

	S1-TEST	C2-TEST
S1-TRAIN	14.27	14.79
C2-TRAIN	14.24	14.17
WER reduction	0.2	**4.2**

is also advantageous from the LM training point of view. As for the overall smaller WERs measured on the S1-TEST set, our analysis showed that direct exposure to ASR results in the baseline pipeline was, yet again, the primary cause for the difference. However, note that the bias is significantly smaller in this setup, because transcribers saw recognition results of an ASR using the same AM as in these experiments, but not the LM.

5.4. Effect on Acoustic Model Training

A similar study was then conducted to understand the impact of the proposed transcription on AM training. A pair of long-short term memory acoustic models (LSTM) [14] was trained using the S1-TRAIN and C2-TRAIN versions of the same training set of 130K utterances and corresponding audio (70 hours). We bootstrapped the training with the cross-entropy criterion (CE) and then proceeded to sequence training (SEQ). Similarly, each of these LSTM AMs was used together with a pre-trained production-quality LM to form two competing ASR systems.

As before, we evaluated these systems on S1-TEST and C2-TEST. Table 2 presents the WER comparison results. Results on both cross-entropy LSTM and sequence LSTM models are presented. Our conclusions are the following: The AMs trained on the unbiased transcriptions (C2-TRAIN) consistently outperform the corresponding AMs trained on the baseline transcriptions (S1-TRAIN) no matter whether they are tested on C2-TEST or S1-TEST. The LSTM trained on C2-TRAIN yields 3.75% and 5.83% WER reduction for cross entropy and sequence training respectively. This agrees with our earlier study which showed that SEQ training is more sensitive to transcription errors and therefore can potentially benefit more from improved transcription quality [15]. Furthermore, we experimented on other AM types, such as fully connected deep neural network acoustic models, and observed similarly expected trends [16].

Table 2: *Effect of transcription method on AM training measured in relative WER(%).*

	S1-TEST	C2-TEST
S1-TRAIN (CE)	15.21	16.43
C2-TRAIN (CE)	14.95	15.81
WER reduction (CE)	1.74	**3.75**
S1-TRAIN (SEQ)	14.44	15.78
C2-TRAIN (SEQ)	14.17	14.86
WER reduction (SEQ)	1.88	**5.83**

5.5. Effect on ASR system

Finally, we wanted to see how much improvement can be achieved when both, acoustic and language models are trained with the proposed unbiased crowdsourced transcriptions. To that objective, we compared two ASR systems with both models trained either on S1-TRAIN or on C2-TRAIN. It should be noted that the setup of this experiment is different from Sections 5.3 and 5.4 in that neither acoustic nor language models are of production quality anymore, but are trained on 130K utterances, thus leading to less accurate outcomes. In that combination, the WER drops from 20.75% for the S1-TRAIN baseline to 20.13% for the models trained on C2-TRAIN, as measured on C2-TEST. There is no change on the less reliable S1-TEST.

6. Discussion

The presented results suggest that the two-stage iterative crowdsourcing transcription pipeline not only eliminates transcriber bias but also delivers superior quality training material despite relying on unskilled crowd. Nonetheless, the above investigations are likely to raise more questions than they provide answers. For instance, it is not entirely clear how much of the advantage will be preserved as the size of the training material increases several orders of magnitude, as is nowadays common for most industrial grade acoustic and language models.

We are also painfully aware of the heuristic nature of our methods used to define posterior probabilities of competing transcription hypotheses and update judge ratings. While the presented methods appear to work well in practice, we are convinced that better and theoretically cleaner alternatives can be devised, such as Bayesian models for judge modeling [17]. Even more important, in our opinion, is optimizing decision policy regarding the next step once a new transcription has been received for an utterance. As of now, for the Cortana domain, an average of 2.5 opinions (in both stages together) is collected for each utterance, which makes it slightly more expensive than professional single-opinion transcriptions. By optimizing this policy we expect to not only reduce noise in the final output but also lower the number of transcription rounds, making the entire process cheaper.

Finally, one could point out that the two compared transcription pipelines produce different numbers of transcriptions due to dismissal of different audio. However, we did run experiments that have shown that this difference (which amounts to a bit more than 1% of all utterances) did not have a significant effect on the results.

7. Future Work and Conclusion

We have identified an important shortcoming of common ASR-assisted transcription methods, namely transcription bias through exposure to ASR recognition hypotheses. It was shown that this bias can account for around 10% of the measured error rates and mask improvements by alternative ASR systems. A two-stage iterative crowdsourcing approach was then proposed that solves the above problem while producing high quality inexpensive transcriptions. These transcriptions are not only preferred to the standard transcriptions by language experts, they also exhibit better correlation with other quality metrics. Furthermore, it was demonstrated that using transcription obtained with the new pipeline for training of language and acoustic models resulted in higher recognition accuracy. The improvements were moderate but consistent and statistically significant. While only highest scoring transcription hypotheses were used for this training, our next step is to investigate how pluralities of weighted transcription alternatives from the crowd can be used for that purpose. Indeed, we expect an effect similar to improvements observed for training on recognition lattices [18]. Finally, by focusing on the decision making policy we hope to reduce transcription costs without affecting its accuracy.

8. References

[1] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," *CoRR*, vol. abs/1512.02595, 2015. [Online]. Available: http://arxiv.org/abs/1512.02595

[2] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," 2016. [Online]. Available: http://arxiv.org/abs/1610.05256

[3] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," 2017. [Online]. Available: http://arxiv.org/abs/1703.02136

[4] L. Rodríguez, F. Casacuberta, and E. Vidal, "Computer assisted transcription of speech," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2007, pp. 241–248.

[5] T. Bazillon, Y. Esteve, and D. Luzzati, "Manual vs assisted transcription of prepared and spontaneous speech." in *LREC*, 2008.

[6] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08, 2008, pp. 614–622.

[7] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 254–263.

[8] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, "Soylent: A Word Processor with a Crowd Inside," in *Proc. of the 23 Annual ACM Symposium on User Interface Software and Technology*, 2010.

[9] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *ICASSP*, 2010, pp. 5270–5273.

[10] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data," in *SLT*, 2010, pp. 312–317.

[11] C.-y. Lee and J. R. Glass, "A transcription task for crowdsourcing with automatic quality control," in *Interspeech*, 2011, pp. 3041–3044.

[12] K. Audhkhasi, A. Zavou, P. Georgiou, and S. S. Narayanan, "Empirical link between hypothesis diversity and fusion performance in an ensemble of automatic speech recognition systems," in *Interspeech*, Aug. 2013.

[13] R. Patel, "Crowdsourcing at microsoft," http://research.microsoft.com/en-us/um/redmond/events/fs2012/presentations/Rajesh_Patel.pdf, 2012.

[14] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech2014*, 2014, pp. 338–342.

[15] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised gmm and dnn acoustic model training with multi-system combination and confidence re-calibration," in *Interspeech*, August 2013.

[16] Y. Huang, y. wang, and Y. Gong, "Semi-supervised training in deep learning acoustic model," in *Interspeech*, September 2016.

[17] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14, 2014, pp. 155–164.

[18] V. Kuznetsov, H. Liao, M. Mohri, M. Riley, and B. Roark, "Learning n-gram language models from uncertain data," in *Interspeech*, 2016.