Empirical Link Between Hypothesis Diversity and Fusion Performance in an Ensemble of Automatic Speech Recognition Systems

Kartik Audhkhasi, Andreas M. Zavou, Panayiotis G. Georgiou, and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL), Electrical Engineering Department
University of Southern California, Los Angeles, CA, USA.
{audhkhas, zavou}@usc.edu, {georgiou, shri}@sipi.usc.edu

Abstract

Diversity is crucial to reducing the word error rate (WER) when fusing multiple automatic speech recognition (ASR) systems. We present an empirical analysis linking diversity and fusion performance. We transcribed speech from the first 2012 US Presidential debate using multiple ASR systems trained with the Kaldi toolkit. We used the N-best ROVER algorithm to perform hypothesis fusion and measured N-best diversity by the average pairwise WER. We make three key observations. We first note that the WER of the fused hypothesis decreases significantly with increasing diversity of the N-best list. This decrease is greater than the decrease in WER of the oracle hypothesis in the list. N-best lists from systems trained on different data sets are the most diverse and give the lowest WER upon fusion. We then observe that the benefit of diversity depends on the choice of the fusion scheme. We show that confidence-weighted ROVER is able to better exploit diversity than unweighted ROVER and gives lower WERs. We finally explain the above observations by a simple linear relation linking diversity to the ROVER WER. This relation depends on the fusion scheme and also reveals the tradeoff between diversity and average WER of hypotheses in the N-best list.

Index Terms: automatic speech recognition, diversity, system combination, ROVER, ensemble methods

1. Introduction

System combination is a widespread practice in automatic speech recognition (ASR) research. Many large-scale projects such as DARPA GALE [1], TRANSTAC [2], EARS [3], and CALO [4] have utilized ensembles of ASR systems. The fused ensemble system often provides lower word error rate (WER) than the individual systems and better generalization to unseen audio data. Diversity of the ASR systems is a desirable characteristic to achieve lower WER upon fusion as noted in [1–4] and other works. Figure 1 illustrates a non-diverse and a diverse 3-best list. The diverse 3-best list in Figure 1(b) contains hypotheses with complementary errors and can thus benefit more from a suitable fusion scheme.

Many works have explored formally the benefits of ensemble diversity in machine learning [5–9]. Researchers have also developed a range of methods for training diverse ASR systems, mainly guided by empirical investigations. The simplest approach is to use diverse data sets and acoustic features for training the ASR systems. Other works [10–14] use machine learning techniques such as bagging [15], boosting [16, 17] and random forests [18] to train diverse ASR systems. Breslin [19] provides a comprehensive review of these techniques. Diverse hypotheses are also prevalent in other speech and language processing tasks. These include transcription of speech data by workers on crowd-sourcing services such as Amazon Mechanical Turk [20–26]. The advantage herein is that while any single transcription by itself may not be perfect, an appropriate combination of diverse set of transcriptions can boost accuracy.

This research was supported by NSF, NIH, and DARPA.

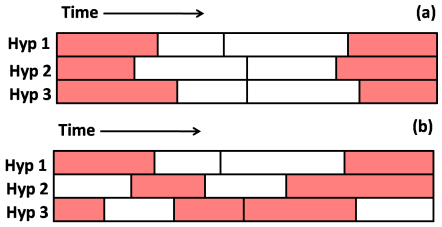

Figure 1: This figure shows a set of (a) non-diverse and (b) diverse 3-best lists. Red (dark) segments indicate a word error with respect to a reference transcript while the white (light) ones show correct recognition. Both (a) and (b) have the same number of errors but the hypotheses in (b) make more complementary errors. A suitable combination scheme can thus reduce the WER by using the 3-best list in (b).

Not many previous works have explicitly addressed the link between hypothesis diversity and fusion performance despite significant interest in training and fusing diverse ASR systems. This paper empirically investigates this important link. We transcribed speech from the first 2012 US Presidential debate using multiple ASR systems trained with the Kaldi toolkit [27]. We then conducted fusion experiments using the N-best ROVER algorithm [28]. We used the average pairwise WER for estimating the diversity of an N-best list. Our key observation is that

$$\text{ROVER WER} \approx \text{Avg. N-best WER} - \gamma(\text{N-best Diversity}) \quad (1)$$

where $\gamma > 0$ is a constant which depends on the N-best list and the ROVER scheme used, as discussed in Section 3.1. $-\gamma$ approximates the change in ROVER WER with a unit change in N-best diversity keeping the average N-best WER constant. The linear expression on the right-hand side of (1) provides an accurate estimate of the ROVER WER with nearly perfect correlation and extremely low mean squared error. It also shows the tradeoff between diversity and average N-best WER and explains the following observations in this paper:

1. The WER of the fused ROVER hypothesis reduces with increasing diversity of the N-best list. This decrease is greater than the decrease in WER of the oracle (lowest

WER) hypothesis in the list. N-best list from a single ASR is the least diverse and gives the least reduction in WER over the 1-best WER upon fusion. N-best lists from systems trained on different data sets are the most diverse and give the greatest benefit in WER upon fusion.

2. Not all ROVER schemes are able to exploit the diversity in an N-best list. A word confidence-aware fusion gives significantly lower WER than the vanilla word frequency-based ROVER.

The next section presents details of our experimental setup - the ASR system training using Kaldi, the testing set, and the variants of the N-best ROVER algorithm we used in this paper. Section 3 discusses the experimental results and presents a detailed analysis of the impact of diversity upon fusion performance. We conclude the paper in Section 4.

2. Experimental Setup

We first give the details of the various ASR systems we trained using the Kaldi toolkit.

2.1. ASR System Training Using the Kaldi Toolkit

The Kaldi toolkit [27] provides state-of-the-art open-source tools for training ASR systems. It offers many advantages over other ASR toolkits (e.g. HTK [29] and Sphinx [30]) such as tight integration with finite state transducers (FSTs) using the OpenFST toolkit [31], generic and extensible design, Apache 2.0 license, and recipes for various standard data sets. We used data from the Wall Street Journal (WSJ) [32], the HUB4 English broadcast news [33], and the ICSI meeting [34] data sets.

We used the default Kaldi training recipe for WSJ to train all ASR systems. This recipe uses the CMU pronunciation dictionary and its phone set. It first computes Mel frequency cepstral coefficients (MFCCs) over 25 msec long speech frames with a 10 msec shift. It then trains monophone models using the Viterbi-EM algorithm. The recipe next aligns the training data using these models and uses the resulting alignments to train triphone models (M1). We used 2000 leaves for decision tree clustering and 10000 total Gaussians.

We obtained the second system (M2) by using the alignments from the M1 models to perform Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transformation (MLLT) for increasing the discrimination between the various phones. We then discriminatively adapted the M2 acoustic models using the Maximum Mutual Information (MMI) [35] criterion which resulted in the final system (M3). Table 1 summarizes the three systems.

ASR System	Training Steps
M1	Triphone Viterbi-EM training
M2	M1 $\rightarrow$ LDA $\rightarrow$ MLLT
M3	M2 $\rightarrow$ MMI

Table 1: This table summarizes the training steps for the three ASR systems used in this paper for each of the WSJ, HUB4, and ICSI data sets.

2.2. First 2012 US Presidential Debate Corpus and Generic Presidential Debate Language Model

We downloaded the audio and reference transcriptions for the first 2012 United States Presidential debate between President Barack Obama (BO) and Governor Mitt Romney (MR) from the National Public Radio (NPR) website[1]. This debate was held at the Magness Arena of the University of Denver in Denver, Colorado and was hosted by Mr. Jim Lehrer (JL). The audio is approximately 90 minutes long.

We performed speaker diarization on this audio using the system from [36] which uses voice activity detection followed by correlation-based segmentation and hierarchical clustering. This diarization system generated three clusters corresponding to JL, BO, and MR. We did not manually correct the obtained cluster labels. We then segmented the audio into clips up to 10 sec long for faster decoding. The number of utterances for JL, BO, and MR were 107, 282, and 253 respectively. We cleaned the debate transcripts by removing punctuation marks and mapping numbers to words (e.g. "$5 trillion" to "five trillion dollars"). We did not add any out-of-vocabulary (OOV) words from the transcripts to the ASR pronunciation dictionary.

Text data from the three ASR data sets is inappropriate for training the language models due to domain mismatch. We thus obtained transcripts of all US Presidential and Vice-Presidential debates from 1960-2012[2] excluding the testing set. These transcripts contain approximately 0.5 million words and were used to train a 4-gram LM with back-off using SRILM [37]. Table 2 gives the WER of the 1-best (Viterbi) hypothesis for each speaker in the testing set using various systems. We observe that the systems trained on the HUB4 data set perform the best. The M3 systems give the lowest WER in most cases. The debate audio, transcripts, and LM are available here[3].

2.3. N-best ROVER Fusion Schemes

The ROVER algorithm first aligns the hypotheses by minimizing the total cost of word insertion, deletion, and substitution needed to make all hypotheses identical. The resulting word confusion network contains a sequence of confusion bins each represented as sets of confusing words between time segments.

Let w_{ij} denote the j-th unique word in the i-th confusion bin with confidence $c_{ij} \in [0,1]$ and frequency $f_{ij} \in [0,1]$. ROVER picks a word w_i^* for this bin using the following decision rule:

$$\textit{Weighted ROVER: } w_i^* = \arg\max_{w_{ij}} \Big[\alpha f_{ij} + (1-\alpha) c_{ij}\Big] \quad (2)$$

where $\alpha \in [0,1]$ is the weight given to the word frequency f_{ij}. ASR system confidence estimation is a challenging problem and has been the subject of several works [38–41]. Confidence estimation systems often are not perfect, and hence not desirable for the analysis of the diversity-fusion performance link that we are exploring in this paper. Hence, we initially use an oracle confidence estimator for our experiments to get a lower bound on the analysis (we consider automatically estimated confidence scores in Section 3.2). We align the hypothesis word sequence with the reference transcript and set the confidence score of the j-th hypothesis word w_j as

$$\textit{Oracle Confidence: } \quad c_j = I(w_j \text{ is correct}) \quad (3)$$

where I is the indicator function. This oracle confidence finds the lower bound on the WER after fusion using N-best ROVER.

[1] http://www.npr.org/2012/10/03/162258551/transcript-first-obama-romney-presidential-debate

[2] http://www.debates.org/index.php?page=debate-transcripts

[3] sail.usc.edu/data.php

1-best WERs									
	HUB4			WSJ			ICSI		
	MR	BO	JL	MR	BO	JL	MR	BO	JL
M1	31.8	34.2	46	39.3	47.7	55.5	40.1	44.2	54.1
M2	29	30.4	44.1	37.1	44.3	59	35.1	41.3	54.8
M3	28.2	29.6	43.6	36	43.2	58.5	33.9	39.6	53.4

Table 2: This table summarizes the testing set 1-best WERs for various ASR systems. The systems trained on the HUB4 data set provide the lowest WERs while those trained on the ICSI data set give the highest WER. M3 models perform the best out of the three models trained for each data set except WSJ for speaker JL.

	HUB4			WSJ			ICSI			WSJ + HUB4 + ICSI		
	MR	BO	JL	MR	BO	JL	MR	BO	JL	MR	BO	JL
10-best Unweighted ROVER WERs												
M1	31.7	33.9	45.7	39.2	47.7	55.3	39.9	44.1	53.8	33	36.9	48
M2	29	30.4	43.9	37	44.3	58.6	35.3	41.1	54.6	30.3	34.1	48.1
M3	28	29.7	43.4	35.8	42.8	58.1	33.9	39.4	52.8	29.2	32.9	46.3
M1 + M2 + M3	28	30	43.6	35.9	41.7	55.1	34.3	39.3	50.7	29.5	32.2	44.6
10-best Oracle Confidence-Weighted ROVER WERs (α = 0.65, picked using cross-validation)												
M1	28.7	31.6	43.2	36.2	45.1	52.6	37	41.8	50.9	26.3	30.3	41.5
M2	25.9	28.2	41.4	34.4	41.7	56.5	32.5	38.7	51.3	23.9	27.5	40.8
M3	24.9	27. 1	40.8	33.2	40.7	56.0	30.6	36.9	50.9	22.8	25.9	39.7
M1 + M2 + M3	24.8	26.9	39.9	30.7	37.6	49.5	29.6	35.5	47.1	**22.6**	**25.2**	**38.2**

Table 3: This table summarizes the testing set WERs for various ASR systems after 10-best ROVER under various conditions. '+' denotes fusion of N-best list across training data sets and/or systems. E.g. the M1 + M2 + M3 row for the WSJ data set indicates that we fused the top-3 hypotheses from the M1, M2, and M3 models before performing ROVER. Confidence-weighted ROVER performs better than both unweighted ROVER and the lowest WER (oracle) hypothesis in the N-best list. We have not included the WERs for the latter in this table due to lack of space.

Setting $\alpha = 1$ gives the unweighted ROVER which does not use any word confidence scores.

$$\textit{Unweighted ROVER:} \quad w_i^* = \arg\max_{w_{ij}} f_{ij} \ . \tag{4}$$

Most works on ASR system ensembles use this unweighted variant. We next present our experiments and analysis.

3. Experiments and Analysis

Table 3 shows the testing set WERs for various system combinations using 10-best ROVER.We created the merged 10-best list by picking equal number of top hypothesis from the individual 10-best lists. For example, the fused N-best list for the M1 + M2 + M3 results using the WSJ data set was created by taking the top 3 hypotheses from each individual list. We also show results when the top hypothesis from each of M1, M2, and M3 models from the three training sets WSJ, HUB4, and ICSI were merged to create a single 10-best list.

We present two sets of results using unweighted ROVER and confidence-weighted ROVER. We make the following observations from Table 3:

1. Both unweighted and confidence-weighted 10-best ROVER result in lower WER than the 1-best WERs in Table 2. The former gives only a marginal improvement over the 1-best WER while the latter gives a significant improvement.
2. Confidence-weighted 10-best ROVER gives significantly lower WER than 10-best unweighted ROVER.
3. Fusion of top 3 hypotheses from the M1, M2, and M3 (in the M1 + M2 + M3 rows of Table 3) systems results in a significant WER reduction over individual systems for confidence-weighted ROVER. However, it does not always reduce the WER for unweighted ROVER.
4. Fusion of top 3 hypotheses from the WSJ, HUB4, and ICSI data sets for each system (in the WSJ + HUB4 + ICSI columns of Table 3) results in a significant WER reduction over individual systems for confidence-weighted ROVER. It is also better than individual data set-dependent systems when using confidence-weighted ROVER. But this is not necessarily the case for unweighted ROVER.
5. Fusion of the top hypothesis from WSJ, HUB4, and ICSI across the three systems (M1, M2, and M3) gives the lowest WER (bottom-right corner of Table 3) when using confidence-weighted ROVER.

The above observations from Table 3 indicate a link between the diversity of the N-best list being fused and the choice of fusion scheme. We discuss it in the next subsection.

3.1. Diversity-Fusion Performance Link

We first define the diversity of an N-best list of word hypotheses before exploring its link with the decision rule. Let $\mathcal{H} = \{h_1, \ldots, h_N\}$ be a set of hypothesis sentences in an N-best list. The average pairwise WER of the list defined as

$$D(\mathcal{H}) = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{m>n} \mathcal{E}(h_n, h_m) \tag{5}$$

is an intuitive measure of the diversity of the N-best list where $\mathcal{E}$ computes the WER between two hypotheses. $\mathcal{D}$ is closely re-

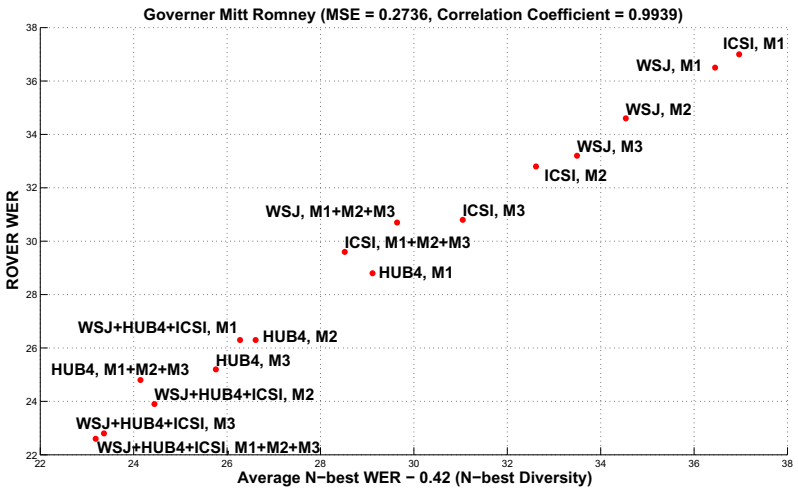

Figure 2: This figure shows the plot between the WER after confidence-weighted 10-best ROVER fusion ($\alpha = 0.65$) and an optimal linear combination of the average 10-best WER and the diversity for various ASR systems. We found the estimated $\gamma = 0.42$ to be consistent in cross-validation. Average 10-best WER and diversity individually give much smaller correlation coefficients of 0.6779 and -0.6147 respectively with the ROVER WER. We observed similar trends for the other speakers and 10-best unweighted ROVER.

lated to the definition of diversity used for an ensemble of maximum entropy models in [42] and the later complementary phone error objective function in [43]. [42] replaces $\mathcal{E}$ by the negative cross-correlation between the class label posteriors from the individual models.

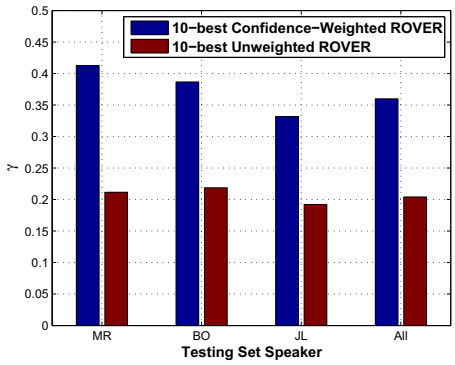

Figure 3: This figure shows the coefficient γ of the diversity term in (1) using an optimal linear combination of the average N-best WER and the diversity for various ASR systems. The estimated γ was consistent across cross-validation folds.

We performed a simple least squares linear regression to estimate the ROVER WER from the average N-best WER and the N-best diversity $D(\mathcal{H})$ for each speaker using the two ROVER schemes. This enabled us to arrive at the following easily-interpretable relation:

$$\text{ROVER WER} \approx \text{Average N-best WER} - \gamma(\text{N-best Diversity})$$

where $\gamma > 0$ is the weight of the diversity term. We found that the above approximation is able to predict the true ROVER WER with a high correlation coefficient (≈ 1) and an extremely low mean squared error. This is apparent from Figure 2 which shows the scatter plot between the true confidence-weighted 10-best ROVER WER (y-axis) and its optimal estimate (x-axis) for various ASR systems using speech from speaker MR. We obtained similar perfect correlation for other speakers using both ROVER schemes. However, the average N-best WER and N-best diversity individually don't predict the ROVER WER well.

Figure 3 shows the estimated diversity term weight γ for different cases. First, γ is always positive, which shows that the ROVER WER reduces with increasing diversity. Second, γ is significantly higher for 10-best confidence-weighted ROVER than unweighted ROVER. This shows that the former exploits N-best diversity much more than the latter.

3.2. The Importance of Accurate ASR Confidence Scores

Our experiments show that the N-best diversity is best exploited by confidence-weighted ROVER. For example, an oracle confidence estimator gives a 0.3-0.5% absolute drop in WER for every 1% increase in N-best diversity. This makes accurate automatic ASR confidence estimation all the more important. We conducted initial experiments with a simple maximum entropy model-based confidence estimator. This estimator uses standard confidence features from the word lattice, such as word posterior probability and normalized AM and LM scores. It gave a 5-10% absolute improvement in error classification accuracy above chance. However, the estimated confidence scores gave an insignificant improvement in WER after 10-best confidence-weighted ROVER. This underscores the importance of accurate confidence score estimation.

4. Conclusion and Future Work

We presented an empirical link between diversity of an N-best list and the WER after ROVER fusion. ROVER WER is highly correlated with a linear function of average N-best WER and N-best diversity. This relation explains many of our empirical observations, such as consistently better performance of confidence-weighted ROVER and reduction in ROVER WER with increasing N-best diversity. Our work shows that there is great incentive for designing diverse ASR systems and accurate confidence estimators. Future work should perform theoretical analysis of the presented link and informed diversity-promoting design of ASR systems.

5. References

[1] H. Soltau et al., "Advances in Arabic speech transcription at IBM under the DARPA GALE program," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 884–894, 2009.

[2] D. Stallard et al., "The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System," in *Proc. Interspeech*, 2007.

[3] S. F. Chen et al., "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.

[4] G. Tur et al., "The CALO meeting speech recognition and understanding system," in *Proc. SLT*. IEEE, 2008, pp. 69–72.

[5] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley-Interscience, 2004.

[6] T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, pp. 1–15, 2000.

[7] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Advances in neural information processing systems*, pp. 231–238, 1995.

[8] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," in *IEEE International Conference on Neural Networks*, 1996, vol. 1, pp. 90–95.

[9] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition*, vol. 29, no. 2, pp. 341–348, 1996.

[10] H. Schwenk, "Using boosting to improve a hybrid HMM/neural network speech recognizer," in *Proc. ICASSP*. IEEE, 1999, vol. 2, pp. 1009–1012.

[11] G. Cook and T. Robinson, "Boosting the performance of connectionist large vocabulary speech recognition," in *Proc. ICSLP*. IEEE, 1996, vol. 3, pp. 1305–1308.

[12] C. Dimitrakakis and S. Bengio, "Boosting HMMs with an application to speech recognition," in *Proc. ICASSP*. IEEE, 2004, vol. 5, pp. 618–621.

[13] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Contructing ensembles of ASR systems using randomized decision trees," in *Proc. ICASSP*. IEEE, 2005, vol. 1, pp. 197–200.

[14] G. Saon and H. Soltau, "Boosting systems for large vocabulary continuous speech recognition," *Speech Communication*, vol. 54, no. 2, pp. 212–218, 2012.

[15] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*. Springer, 1995, pp. 23–37.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] C. Breslin, *Generation and combination of complementary systems for automatic speech recognition*, Ph.D. thesis, Cambridge University Engineering Department and Darwin College, 2008.

[20] K. Audhkhasi, P. G. Georgiou, and S.S. Narayanan, "Analyzing quality of crowd-sourced speech transcriptions of noisy audio for acoustic model adaptation," in *Proc. ICASSP*, 2012.

[21] K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan, "Reliability-weighted acoustic model adaptation using crowd-sourced transcriptions," in *Proc. Interspeech*, 2011.

[22] K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan, "Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics," in *Proc. ICASSP*, 2011.

[23] G. Parent and M. Eskenazi, "Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges," in *Proc. Interspeech*, 2011.

[24] B. C. Roy, S. Vasoughi, and D. Roy, "Automatic estimation of transcription accuracy and difficulty," in *Proc. Interspeech*. ISCA, 2010, pp. 1902–1905.

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

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

[27] D. Povey et al., "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*. Dec. 2011, IEEE.

[28] J. Fiscus, "A post processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. ASRU*. IEEE, 1997, pp. 347–354.

[29] S. Young et al., "The HTK book," *Cambridge University Engineering Department*, 2002.

[30] K-F. Lee, H-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990.

[31] C. Allauzen et al., "OpenFst: A general and efficient weighted finite-state transducer library," *Implementation and Application of Automata*, pp. 11–23, 2007.

[32] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[33] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett, "English broadcast news speech (HUB4)," *Linguistic Data Consortium, Philadelphia*, 1997.

[34] A. Janin et al., "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–364.

[35] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 2003.

[36] W. Wang, P. Lv, and Y. Yan, "An improved hierarchical speaker clustering," *Acta Acoustica*, 2006.

[37] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[38] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[39] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[40] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," in *Proc. NAACL*, 2010.

[41] M. Siu and H. Gish, "Evaluation of word confidence for speech recognition systems," *Computer Speech and Language*, vol. 13, no. 4, pp. 299–319, 1999.

[42] K. Audhkhasi, A. Sethy, B. Ramabhadran, and S. S. Narayanan, "Creating ensemble of diverse maximum entropy models," in *Proc. ICASSP*. IEEE, 2012, pp. 4845–4848.

[43] F. Diehl and P. C. Woodland, "Complementary phone error training," in *Interspeech*, 2012.