Accepted, Jan 2014

Theoretical Analysis of Diversity in an Ensemble of Automatic Speech Recognition Systems

Kartik Audhkhasi, *Student Member, IEEE,* Andreas M. Zavou, *Student Member, IEEE*
Panayiotis G. Georgiou, *Senior Member, IEEE* and Shrikanth S. Narayanan, *Fellow, IEEE*

***Abstract*—Diversity or complementarity of automatic speech recognition (ASR) systems is crucial for achieving a reduction in word error rate (WER) upon fusion using the ROVER algorithm. We present a theoretical proof explaining this often-observed link between ASR system diversity and ROVER performance. This is in contrast to many previous works that have only presented empirical evidence for this link or have focused on designing diverse ASR systems using intuitive algorithmic modifications. We prove that the WER of the ROVER output approximately decomposes into a difference of the average WER of the individual ASR systems and the average WER of the ASR systems with respect to the ROVER output. We refer to the latter quantity as the diversity of the ASR system ensemble because it measures the spread of the ASR hypotheses about the ROVER hypothesis. This result explains the trade-off between the WER of the individual systems and the diversity of the ensemble. We support this result through ROVER experiments using multiple ASR systems trained on standard data sets with the Kaldi toolkit. We use the proposed theorem to explain the lower WERs obtained by ASR confidence-weighted ROVER as compared to word frequency-based ROVER. We also quantify the reduction in ROVER WER with increasing diversity of the N-best list. We finally present a simple discriminative framework for jointly training multiple diverse acoustic models (AMs) based on the proposed theorem. Our framework generalizes and provides a theoretical basis for some recent intuitive modifications to well-known discriminative training criterion for training diverse AMs.**

***Index Terms*—Automatic speech recognition, diversity, ambiguity decomposition, system combination, ROVER, ensemble methods, discriminative training.**

I. Introduction

AUTOMATIC speech recognition (ASR) is a challenging task due to several factors ranging from variability in speaker and environmental acoustic characteristics to mismatch in topics or domains of the conversation [2]. Most ASR systems adopt statistical models such as hidden Markov models (HMMs) and N-gram language models (LMs) to account for this variability. However, many large-scale ASR tasks such as exemplified by DARPA GALE [3], TRANSTAC [4], EARS [5], CALO [6], and BABEL [7] projects have shown that the fusion of hypotheses from multiple ASR systems is essential to achieve state-of-the-art word error rates (WERs). The ROVER algorithm [8] typically performs this fusion. The observed reduction in WER is primarily due to complementary errors made by the different ASR systems.

Researchers hence have also focused on designing diverse ASR systems that make complementary errors for fusion. The simplest approach is to use intuitively diverse data sets and diverse acoustic features such as Mel frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) features for training the ASR systems. Another common approach is to combine structurally diverse AMs such as based on Gaussian mixture model (GMM) HMMs and deep neural networks (DNNs) [7]. Other works [9]–[14] have used 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. A recent work by Cui, Huang, and Chien [20] presents a multi-view and multi-objective algorithm for semi-supervised training of HMM acoustic models. The authors generate multiple ASR systems on different views of the training data by using different front-ends and randomized decision trees. The diversity of these ASR systems is a necessary condition for their mutual information-based optimization algorithm because they use the ROVER output on unlabeled data as reference. Chen and Zhao [21] train diverse AMs using cross-validation and speaker clustering. They also explore several intuitive measures of AM diversity such as the standard deviation of per-frame acoustic scores computed by the AMs. They empirically show that the increased diversity compensates for a reduction in the quality of the AMs being combined, which results in an improvement in WER after fusion. No prior work has however theoretically analyzed the impact of ASR diversity on fusion performance to the best of our knowledge despite widespread interest in training and fusing diverse ASR systems.

On the other hand, system complementarity or diversity is a well-studied problem in machine learning and statistical signal processing. Many researchers have explored formally the benefits of ensemble diversity in machine learning [22]–[26]. We present a theoretical link between ASR system diversity and ROVER performance in this paper motivated by these previous works. We specifically use the *ambiguity decomposition* proposed in [24] for squared-error regression as the framework. The ambiguity decomposition states that the squared-error of a convex sum of regressors decomposes into the difference between the weighted squared-error of the individual regressors and the diversity of the ensemble. This diversity is defined as the weighted squared-error of

This work supported in part by the NSF, DARPA, and IARPA.

K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan are with the Signal Analysis and Interpretation Lab (SAIL) of the Electrical Engineering Department at the University of Southern California, Los Angeles, CA, 90089-2564, USA. E-mail: audhkhas@usc.edu.

A. Zavou is with the Electrical Engineering Department at the Cyprus University of Technology, Limassol, Cyprus.

Some content from this paper appears in Interspeech, 2013 [1].

Variable	Description
M	Number of ASR hypotheses being fused by ROVER
K	Vocabulary size
N	Number of cohort sets in a given confusion network generated by ROVER
$\mathcal{C}_i$	i-th cohort set
$\mathbf{w}_i^m$	1-in-K encoding of system m's word hypotheses in $\mathcal{C}_i$
$\mathbf{s}_i^m$	1-in-K encoding of system m's confidence score in $\mathcal{C}_i$
α	word frequency weight in ROVER
$\mathbf{h}_i^m$	$\alpha\mathbf{w}_i^m + (1-\alpha)\mathbf{s}_i^m$
$\mathbf{h}_i^{\text{avg}}$	$\frac{1}{M}\sum_{m=1}^{M}\mathbf{h}_i^m$
$\mathbf{r}_i$	1-in-K encoding of the reference word in $\mathcal{C}_i$
$\mathbf{h}_i^*$	1-in-K encoding of the ROVER word hypothesis $\mathcal{C}_i$
$E(.)$	true WER
E_{approx}	approximate WER
p	probability of a correct word
p_{ML}	ML estimate of p

TABLE I
A LIST OF KEY VARIABLES AND THEIR DESCRIPTION USED IN LEMMAS 1-2 AND THEOREMS 1-3.

each regressor from the ensemble's prediction. This result is equivalent to the bias-variance-covariance decomposition derived in [25] and forms the basis of negative correlation learning for neural networks [27]. It is also intuitively similar to the bias-variance decomposition [28]. Appendix A further explores this link.

We first propose a simple vector space model for ROVER WER in Section II to establish the link with ambiguity decomposition. We use this model to approximate the ROVER WER by a simpler expression involving the squared error. This enables us to directly apply the ambiguity decomposition in Section III and decompose the WER between the reference transcription and the hypothesis transcription generated by ROVER. We show that our proposed decomposition applies both at the per-utterance level and also on average over a data set. Section IV describes our experimental setup using the Kaldi ASR toolkit [29] and ASR confidence estimation using a variety of lattice-based and prosodic features within a conditional random field (CRF) [30] model. We also describe the different fusion strategies we tested in ROVER and our test data set. We empirically validate our proposed theorems on the test set in Section V among other experiments and analysis. We then use our theoretical results to give a unified discriminative training framework using the minimum Bayes risk (MBR) criterion for training diverse ASR systems in Section VI. We conclude the paper in Section VII with some directions for future work.

II. A Vector Space Model for ROVER WER

This section presents a vector space model for the WER of the ROVER output of an ASR system ensemble. We will use this model for proving the link between hypotheses diversity and ROVER WER. ROVER first aligns the multiple word sequence hypotheses using dynamic programming (DP). This alignment also allows for insertions and deletions in the word sequence hypotheses and minimizes the total cost of word insertions, deletions, and substitutions[1]. This total cost is also known as Levenshtein distance metric over the set of strings. The aligned output is called a word confusion network (WCN). It consists of a temporal sequence of sets of competing word hypotheses. We refer to each such set as a *cohort set* [31].

Consider M ASR systems decoding a single given audio file. Readers must note that this includes the case of an M-best list generated from any single ASR system. We first perform the DP alignment of the M decoded 1-best sentence hypotheses. Let N be the total number of cohort sets in the resulting confusion network. We note that our theoretical analysis applies to ROVER fusion of any type of ASR output (1-best, M-best, confusion network, and word lattice) because ROVER always performs fusion on the confusion network output by the DP alignment. Let K be the number of words in the decoding vocabulary of the ASR systems. We assume that this vocabulary also includes a special symbol to account for word insertions and deletions during the DP alignment. Consider the i-th cohort set $\mathcal{C}_i$. We encode each word in this cohort set using a 1-in-K or *one-hot* encoding scheme[2]. Let $\mathbf{w}_i^m$ be the 1-in-K encoding of the word hypothesis from ASR system m in cohort set $\mathcal{C}_i$. Each ASR system can additionally provide a confidence score in $[0,1]$ for every word hypothesis. We also embed this confidence score in a K-dimensional vector $\mathbf{s}_i^m$ which contains the score in the location of the word hypothesis and zeros everywhere else.

ROVER computes the following convex combination of the word bit vector $\mathbf{w}_i^m$ and the confidence score vector $\mathbf{s}_i^m$:

$$\mathbf{h}_i^m = \alpha\mathbf{w}_i^m + (1-\alpha)\mathbf{s}_i^m \tag{1}$$

where $\alpha \in [0,1]$ is a user-defined parameter. ROVER then averages $\mathbf{h}_i^m$ across the M ASR systems to obtain

$$\begin{aligned} \mathbf{h}_i^{\text{avg}} &= \frac{1}{M}\sum_{m=1}^{M}\mathbf{h}_i^m && (2)\\ &= \alpha\frac{1}{M}\sum_{m=1}^{M}\mathbf{w}_i^m + (1-\alpha)\frac{1}{M}\sum_{m=1}^{M}\mathbf{s}_i^m && (3)\\ &= \alpha\mathbf{w}_i^{\text{avg}} + (1-\alpha)\mathbf{s}_i^{\text{avg}}\ . && (4) \end{aligned}$$

The first term on the right-hand side of the above equation (4) is a vector containing the frequency of each vocabulary word in the cohort set $\mathcal{C}_i$. The second term contains the average confidence score of each word in the vocabulary. Elements of $\mathbf{h}_i^{\text{avg}}$ lie in $[0,1]$ but they do not sum to 1 ($\mathbf{1}^T\mathbf{h}_i^{\text{avg}} \neq 1$) because $\mathbf{1}^T\mathbf{s}_i^{\text{avg}} \neq 1$.

ROVER next thresholds $\mathbf{h}_i^{\text{avg}}$ such that its maximum element is set to 1 and all others are set to 0. We denote the resulting 1-in-K bit vector by $\mathbf{h}_i^*$. Let $\mathbf{r}_i$ be the 1-in-K encoding of the DP-aligned reference word for $\mathcal{C}_i$. Thus the WER for $\mathcal{C}_i$ is the following $0/1$ loss function:

$$E(\mathbf{r}_i, \mathbf{h}_i^*) = \left\{ \begin{array}{ll} 0 & ;\text{ if } \mathbf{r}_i = \mathbf{h}_i^* \\ 1 & ;\ \mathbf{r}_i \neq \mathbf{h}_i^* \end{array} \right. . \tag{5}$$

[1] We will assume an equal cost of 1 for insertions, deletions, and substitutions. Our theoretical analysis easily extends to unequal costs.

[2] $\mathbf{w}_i^m$ is a K-dimensional bit vector with 1 in the position of the occuring word and 0 everywhere else.

Both $\mathbf{r}_i$ and $\mathbf{h}_i^*$ are 1-in-K bit vectors which allows us to rewrite the WER using the L_2 norm as

$$E(\mathbf{r}_i, \mathbf{h}_i^*) = \frac{1}{2}\|\mathbf{r}_i - \mathbf{h}_i^*\|_2^2 \; . \tag{6}$$

The total number of word errors in the given audio file is just the sum of $E(\mathbf{r}_i, \mathbf{h}_i^*)$ over all N cohort sets $\mathcal{C}_i$. The next lemma proves the relation between $E(\mathbf{r}_i, \mathbf{h}_i^*)$ and the probability of a word error under the simplistic assumption of independent and identically distributed (IID) Bernoulli errors.

Lemma 1. *Define the probability of a correct word*

$$p = P(\mathbf{r}^T\mathbf{h}^* = 1) \tag{7}$$

where both $\mathbf{r}$ *and* $\mathbf{h}^*$ *are random vectors. We assume that all word errors are IID Bernoulli random variables with parameter* $1-p$. *Then*

$$p_{\mathit{ML}} = 1 - \frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) \tag{8}$$

is the maximum likelihood (ML) estimate of p.

The proofs for the above lemma and other results in this paper appear in Appendix B. We note that $\mathbf{h}_i^*$ is a non-linear (threshold) function of $\mathbf{h}_i^{\text{avg}}$ which makes the analysis of diversity in $E(\mathbf{r}_i, \mathbf{h}_i^*)$ from (6) difficult. The ambiguity decomposition considers the average prediction from the individual regressors. Thus we instead propose to use the following approximation to $E(\mathbf{r}_i, \mathbf{h}_i^*)$:

$$E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) = \frac{1}{2}\|\mathbf{r}_i - \mathbf{h}_i^{\text{avg}}\|_2^2 \; . \tag{9}$$

$E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}})$ is easier to analyse because it directly uses $\mathbf{h}_i^{\text{avg}}$ from (4) in place of its non-linear transformation $\mathbf{h}_i^*$. The next lemma relates $E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}})$ to the ML estimate p_{ML} of the probability of a correct word, p. Each $\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}}$ takes a value between 0 and 1. $\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}} = 0$ occurs when the true word does not appear in the cohort set $\mathcal{C}_i$ leading to a word error. $\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}} = 1$ occurs when all the M ASR systems predict the correct word in the cohort set. $\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}}$ thus equals the empirical estimate of the total probability of the correct word in $\mathcal{C}_i$. We thus make the natural assumption that $\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}}$ are IID random variables supported on $[0,1]$ with mean p.

Lemma 2. *We assume that* $\mathbf{r}_i^T\mathbf{h}_i^{\mathit{avg}}$ *are IID random variables supported on* $[0,1]$ *with mean* p. *Then*

$$p_{\mathit{ML}} \leq 1 - \frac{1}{N}\sum_{i=1}^{N} E_{\mathit{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\mathit{avg}}) \tag{10}$$

$$\text{and } p_{\mathit{ML}} \geq \frac{1}{2}\left(1 + \frac{\alpha^2}{M}\right) - \frac{1}{N}\sum_{i=1}^{N} E_{\mathit{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\mathit{avg}}) \tag{11}$$

where M *is the number of ASR systems and* $\alpha \in [0,1]$ *is the ROVER parameter.*

The assumptions in Lemma 2 allows us to relate the sample mean of $\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}}$ over the N cohort sets to the sample mean of the IID Bernoulli random variables $\mathbf{r}_i^T\mathbf{h}_i^*$ from Lemma 1. Lemmas 1 and 2 thus give us the following bounds which relate the averages of the true $E(\mathbf{r}_i, \mathbf{h}_i^*)$ and its approximation $E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}})$ over all N cohort sets for a given audio file.

Theorem 1. *Assume that all word errors are IID Bernoulli random variables with parameter* $1-p$ *and* $\mathbf{r}_i^T\mathbf{h}_i^{\mathit{avg}}$ *are IID random variables with mean* p *and support* $[0,1]$. *Then*

$$\frac{1}{N}\sum_{i=1}^{N} E_{\mathit{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\mathit{avg}}) \leq \frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) \text{ and} \tag{12}$$

$$\frac{1}{N}\sum_{i=1}^{N} E_{\mathit{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\mathit{avg}}) \geq \frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) - \frac{1}{2}\left(1 - \frac{\alpha^2}{M}\right) . \tag{13}$$

We note that the lower-bound in (13) becomes independent of α as the number of ASR systems M increases. In the limit of the number of systems $M \to \infty$ or when $\alpha = 0$, the approximate ROVER WER lies in the interval:

$$\frac{1}{N}\sum_{i=1}^{N} E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) \in \left[\frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) - \frac{1}{2}, \frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*)\right] . \tag{14}$$

Both the true and the approximate ROVER WERs are random variables. Hence, it is useful of think of the above bounds as a statistical *confidence interval* in which the approximate ROVER WER lies. The assumption of IID word errors in Theorem 1 will be violated for a few decoded utterances. But the approximate ROVER WER is highly likely to lie in this confidence interval for a large majority of the test cases. We illustrate this through experiments in Section V.

Theorem 1 provides a link between the averages of the true and the approximate ROVER WERs. This justifies our use of the simpler approximate ROVER WER in place of the true WER for analyzing diversity using the ambiguity decomposition in the next section. We finally use Theorem 1 to derive a link between diversity and the true ROVER WER.

III. Ambiguity Decomposition for ROVER WER

The previous section presented a vector space model for the true ROVER WER, its tractable approximation, and some bounds that relate the true and approximate WER. This framework allows us to define diversity of hypotheses from the ASR systems being combined and its impact on the ROVER WER in this section.

There are several definitions of diversity for an ensemble of machine classifiers [32], [33]. We use the ambiguity decomposition [24] because it readily applies to a convex combination of predictions from M regressors using the squared-error loss function. This fits our approximate WER proposed in the previous section and also provides an easily interpretable result. The bias-variance-covariance decomposition [25] is an equivalent result which was derived in the context of neural networks. The ambiguity decomposition uses a scalar target variable but it is easy to extend to the vector case which we use for our approximate ROVER WER. The next theorem presents the ambiguity decomposition for ROVER WER.

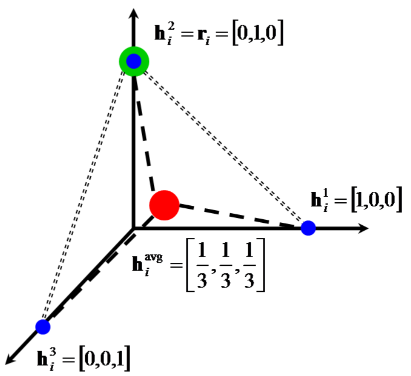

$\mathbf{h}_i^2 = \mathbf{r}_i = [0,1,0]$
$\mathbf{h}_i^1 = [1,0,0]$
$\mathbf{h}_i^{\text{avg}} = \left[\frac{1}{3},\frac{1}{3},\frac{1}{3}\right]$
$\mathbf{h}_i^3 = [0,0,1]$

Fig. 1. This figure illustrates the ambiguity decomposition for ROVER WER presented in Theorem 2. We consider recognition of a single word out of a vocabulary of $K = 3$ words and $M = 3$ ASR systems. Each of the ASR systems predicts a different word in the given cohort set. The three axes constitute the Euclidean vector space arising due to the 1-in-3 encoding of words. The average ROVER prediction $\mathbf{h}^{\text{avg}}$ is $[1/3, 1/3, 1/3]$ and the approximate WER is $1/3$. Theorem 2 decomposes this into a difference of the average WER of the 3 systems computed from the average squared-length of the finely-dotted lines (2/3) and the diversity of the ensemble computed from the average squared-length of the thick-dotted lines (1/3).

Theorem 2. *The approximate ROVER WER for any cohort set $\mathcal{C}_i$ decomposes as*

$$E_{approx}(\mathbf{r}_i, \mathbf{h}_i^{avg}) = \frac{1}{M}\sum_{m=1}^{M} E_{approx}(\mathbf{r}_i, \mathbf{h}_i^m) - \frac{1}{M}\sum_{m=1}^{M} E_{approx}(\mathbf{h}_i^{avg}, \mathbf{h}_i^m)\ . \tag{15}$$

Theorem 2 says that for any cohort set $\mathcal{C}_i$, the approximate ROVER WER equals the average approximate WER of the individual systems minus the diversity of the ASR system ensemble defined as

$$\text{Diversity: } \frac{1}{M}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{h}_i^{\text{avg}}, \mathbf{h}_i^m)\ . \tag{16}$$

This diversity equals the average of the approximate WERs of the individual systems from the ROVER's prediction. Figure 1 illustrates the nature of the ambiguity decomposition by means of an example in a 3-dimensional Euclidean space for a given cohort set with $M = 3$ ASR systems. Diversity measures the average spread of the individual ASR predictions around the average prediction $\mathbf{h}^{\text{avg}}$.

Because (15) applies point-wise, i.e. for each cohort set $\mathcal{C}_i$, it also applies on average across the N cohort sets for the given audio file. This gives

$$\frac{1}{N}\sum_{i=1}^{N} E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^m) - \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{h}_i^{\text{avg}}, \mathbf{h}_i^m)\ . \tag{17}$$

The above equation in conjunction with Theorem 1 enables us to derive ambiguity decomposition bounds for the average true ROVER WER. We present this in the next theorem.

Theorem 3. *The average true ROVER WER over N cohort sets $\{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$ decomposes into the following bounds:*

$$\frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) \leq \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} E(\mathbf{r}_i, \mathbf{h}_i^m) - \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} E(\mathbf{h}_i^m, \mathbf{h}_i^*) - \left(\frac{\alpha^2}{M} - 1\right) \textit{ and} \tag{18}$$

$$\frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) \geq \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} E(\mathbf{r}_i, \mathbf{h}_i^m) - \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M} E(\mathbf{h}_i^m, \mathbf{h}_i^*)\ . \tag{19}$$

Theorem 3 gives insight into the relation between ROVER WER and ASR ensemble diversity in terms of the true WER. We have made two key statistical assumptions to prove this result - the word errors are IID Bernoulli random variables with probability of error $1 - p$ and $\mathbf{r}_i^T \mathbf{h}_i^{\text{avg}}$ are IID random variables with mean p. These assumptions are not realistic because word errors are not IID random variables. ASR word errors often occur in clusters and tend to be bursty due to the context used in the AM/LM and the sequential nature of decoding. Furthermore, Theorem 3 only provides bounds on the true ROVER WER in terms of ensemble diversity. However, the results of this section do explain the impact of diversity on ROVER WER and motivate ASR system fusion experiments in the next section. We now present our experiments with multiple ASR systems to check the applicability of the proposed decomposition.

IV. Experimental Setup

We first describe the training of various ASR systems using the Kaldi toolkit in the next section. This is followed by a description of our ASR confidence estimation module in Section IV-B and our test data set in Section IV-C.

A. ASR System Training in Kaldi

The Kaldi toolkit [29] provides state-of-the-art open-source tools for training ASR systems. It offers many advantages over other publicly available ASR toolkits (e.g. HTK [34] and Sphinx [35]) such as tight integration with finite state transducers (FSTs) using the OpenFST toolkit [36], generic and extensible design, Apache 2.0 license, and recipes for various standard data sets. We used data from the Wall Street Journal (WSJ) [37], the HUB4 English broadcast news [38], and the ICSI meeting [39] data sets for our experiments. These data sets are popular with researchers in automatic speech recognition and speech processing.

We used the default Kaldi training recipe for WSJ to train all the 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. Kaldi's AM training tools do not use the Baum-Welch or the exact EM algorithm because the computationally cheaper Viterbi-EM algorithm gives similar word recognition performance. The training recipe next aligns the training data using these models and uses the resulting alignments to train triphone models (our baseline system M1). We used 2000 leaves for decision tree clustering and 10000 total Gaussians for the triphone models.

We then obtained the second system (system M2) by using the alignments from the M1 models to perform Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transformation (MLLT). Both these steps increase the discrimination between the various phones and thus lead to better recognition accuracy. We then discriminatively adapted the M2 acoustic models using the Maximum Mutual Information (MMI) [40] criterion which resulted in the final system M3. The MMI optimization maximizes the mutual information between the true state sequence and the acoustic feature vectors. Prior work has shown that it improves WER beyond a generatively-trained AM. Table II summarizes the three systems and their training steps.

ASR System	Training Steps
M1	Triphone Viterbi-EM training
M2	M1 → LDA → MLLT
M3	M2 → MMI

TABLE II

TABLE II SUMMARIZES THE TRAINING STEPS FOR THE THREE ASR SYSTEMS USED IN THIS PAPER FOR EACH OF THE WSJ, HUB4, AND ICSI DATA SETS.

We now discuss our ASR confidence estimation module which enables us to test the applicability of the ambiguity decomposition for confidence-weighted ROVER in addition to the conventional word frequency-based ROVER.

B. ASR Confidence Estimation

ROVER allows each word to possess a confidence score between 0 and 1 during fusion. A confidence score of 1 indicates that the word is correct and thus should be assigned more weight in its corresponding cohort set. ASR confidence scores can also provide useful information to the user of the ASR system and other modules which use the ASR output, such as spoken dialog systems. Many researchers have thus focused on designing algorithms for ASR confidence estimation. [41] provides a detailed survey of these algorithms. Most of these techniques use rich representations of ASR hypotheses such as word lattices, confusion networks, and N-best lists.

The most common approach for ASR confidence estimation is to compute the posterior probability of the word hypotheses on each arc using the forward-backward algorithm [42] on the word lattice [43], [44]. This directly gives a word confidence score without any additional processing. However, many researchers have found out that training a classifier which uses additional features derived from the word lattice and AM/LM scores is better able to predict ASR word errors. Examples of such classifiers include maximum entropy models [45], conditional random fields [46], boosted weak learners such as decision stumps [47], and neural networks [48]. These classifier-based techniques typically out-perform the word lattice posterior probability in terms of standard metrics such as normalized cross-entropy (NCE) and equal error rate (EER) because they optimally combine additional features to minimize a loss function on a labeled data set.

We used conditional random fields (CRFs) [49] for performing confidence estimation in this paper. A CRF is a discriminative model which estimates the conditional probability $p(\mathbf{y}|\mathbf{x})$ of a label sequence $\mathbf{y}$ given input data $\mathbf{x}$. This conditional probability is written as the normalized log-linear function

$$p_{\Lambda}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\Lambda}(\mathbf{x})} \exp\Big(\sum_k \lambda_k f_k(\mathbf{x},\mathbf{y})\Big) \tag{20}$$

where $f_k(\mathbf{x},\mathbf{y})$ is the k-th feature, λ_k is its weight and $Z_{\Lambda}(\mathbf{x})$ is the normalization constant also known as the partition function. The features are defined on fully-connected subgraphs or *cliques* of a Markov random field (MRF) composed of elements of $\mathbf{x}$ and $\mathbf{y}$. Quasi-Newton optimization algorithms such as L-BFGS [50] typically perform parameter estimation of a CRF by maximizing the conditional likelihood of a labeled training data set. Negative L1 and/or L2 norm of the parameter vector Λ added to the log-likelihood objective function performs regularization and gives lower weight to redundant features.

Elements of our extracted data vector $\mathbf{x}$ fall into two categories – ASR system lattice-based cues and speech signal-based prosodic cues. We use the time boundaries for each word in the given hypothesis sentence to extract the following cues:

1) **ASR lattice-based cues:** The ASR lattice provides valuable cues for confidence estimation because it retains many competing word hypotheses and their associated AM/LM scores within the time boundaries of a given word in the 1-best hypothesis. We extracted the following cues from the ASR lattice:
   a) *Word hypothesis posterior probability* - A high posterior probability of the 1-best word in the current time segment indicates high confidence.
   b) *Entropy of word posterior PMF* - We computed the posterior probability mass function (PMF) over all words in the lattice within the given time boundaries. A peaky posterior distribution characterized by low entropy indicates that the ASR is confident about its word hypothesis.
   c) *Frequency of word hypothesis* - The frequency of the 1-best word hypothesis in the given time segment provides a proxy for its posterior probability.
   d) *Entropy of word frequency distribution* - This feature was motivated by the entropy of the word posterior pdf. A low entropy again indicates high confidence.
   e) *Number of unique word hypotheses* - A high number of unique word hypotheses indicates that the ASR system is confused between several competing word hypotheses and the 1-best hypothesis is likely to be of lower confidence.
   f) *Number of unique context-dependent states* - Kaldi's decoder generates lattices which contain the context-dependent state (transition ID) sequence on each arc. A high number of unique context dependent states indicates large number of state transitions and thus

	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 III

Table III summarizes the testing set 1-best WERs for various ASR systems. The systems trained on the HUB4 data set provide the lowest WERs. M3 models perform the best out of the three models trained for each data set except WSJ for speaker JL.

acoustic instability. This is likely to happen in case of a word error.

g) *AM and LM scores* - We used the duration-normalized acoustic and language model log-likelihoods as additional cues.

h) *Duration of the hypothesis word* - Including the word duration helps us model any duration-dependent behavior of word errors.

i) *Word identity* - We also included word identity to model the observation that the ASR systems committed more frequent errors on certain words than others.

2) **Prosody-based cues:** We also extracted mean, standard deviation, minimum, and maximum of the loudness, voicing probability, and pitch using the Opensmile toolkit [51]. Prior work has shown the usefulness of prosodic features for predicting ASR errors [52]–[54] and speech disfluencies [55].

Most CRF softwares including the one we used for this paper (CRF++[3]) allow only discrete inputs. We thus quantized each of the above cues into 10 levels using the K-means algorithm. Using a higher number of quantization levels led to an exponential increase in the number of CRF features and resulted in over-fitting. Let $y_t = 1$ if the t-th word in a given hypotheses sequence is correct and 0 otherwise. Let $\mathbf{x}_t$ represent the corresponding vector of the above confidence cues after quantization. We computed two broad categories of binary features for the CRF model - state (unigram) features which depend only on the current label y_t, and transition (bigram) features which depend on both the current label and previous label y_{t-1}. Examples of these unigram and bigram features are

$$f_k(\mathbf{x}_t, y_t) = I(x_{tl} = b \text{ and } y_t = 1) \quad \text{and} \tag{21}$$

$$f_k(\mathbf{x}_t, y_t, y_{t-1}) = I(x_{tl} = b \text{ and } y_t = 1 \text{ and } y_{t-1} = 0) \tag{22}$$

where x_{tl} is the l-th component of $\mathbf{x}_t$, b is the discrete bin index, and I is the indicator function. We also included features which utilized contextual information for the input $\mathbf{x}$ - $f_k(\mathbf{x}_{t-1}, y_t)$, $f_k(\mathbf{x}_{t+1}, y_t)$, $f_k(\mathbf{x}_{t-1}, y_t, y_{t-1})$, $f_k(\mathbf{x}_{t+1}, y_t, y_{t-1})$. This was motivated by the observed benefits of using temporal context in improving OOV detection [56]. We mitigated over-fitting the training data by including an L2 penalty in the training objective function. Using an L1 penalty instead reduced the ASR confidence estimation performance slightly. We also omitted any features which occurred less than 10 times in the training data.

The next section describes the test data set used in this paper and its associated in-domain language model.

C. The 2012 US Presidential Debate Data Set

We downloaded the audio and human-generated 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[4]. This debate was held on October 3, 2012 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 did not compose the test set using held-out audio from the WSJ, HUB4, ICSI or other well-known ASR data sets because we wanted to investigate the benefits of diversity in the ASR ensemble on a totally new domain.

We performed speaker diarization on this audio using the system from [57] 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 within memory constraints. The number of utterances for JL, BO, and MR were 107, 282, and 253, respectively. We cleaned the reference 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 to minimize any data set-specific processing.

Text data from the three standard ASR data sets used here is inappropriate for training the language models due to target domain mismatch. We thus obtained transcripts of all US Presidential and Vice-Presidential debates from 1960-2012[5] excluding the first Presidential debate in 2012. These transcripts contain approximately 0.5 million words and were used to train a 4-gram LM with back-off using SRILM [58]. Table III 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 broadcast news data set perform the best. The M3 systems give the lowest WER in most cases. We are making the debate audio, transcripts, and LM available here[6].

The next section presents various experimental results using this presidential debate data set.

[3] http://crfpp.googlecode.com/svn/trunk/doc/index.html

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

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

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

V. Results and Analysis

We start this section by presenting our experimental results for ASR system confidence estimation and WER after N-best fusion using different variants of ROVER. We then provide experimental validation for our proposed diversity-ROVER WER link.

A. ASR System Confidence Estimation

We generated word lattices with the Kaldi toolkit for all audio files in the US presidential debate data set using 9 different ASR systems - systems M1, M2, and M3 each trained on the WSJ, HUB4, and ICSI data sets. We then extracted the 10-best lists from the lattices along with time boundaries and AM/LM score for each word hypothesis. We did not pick a larger N-best list size because the lattices for a few short audio files contained less than 10 unique sentence hypotheses. We then computed the prosodic and lattice-based cues for word confidence estimation as described in section IV-B. We also labeled each word as correct or incorrect using the available human-generated transcriptions. We finally trained 3 separate CRFs for confidence estimation using the labeled 10-best lists for each ASR systems trained on the 3 different data sets (HUB4, WSJ, and ICSI). Each CRF used data for all 3 ASR systems (M1, M2, and M3) for a given training data set (e.g. HUB4). We performed 4-fold cross-validation during training by leaving out 25% of each test speaker's utterances for testing and used the remaining data for training. We used the default value of 1 for the L2 penalty weight in CRF++.

We used two popular performance metrics for evaluating the performance of the soft confidence score estimates given the true word error labels. The first one is the equal error rate (EER) which is the false alarm rate or the miss rate at the confidence score threshold where the false alarm and miss rates are equal. The EER can be graphically computed as the false alarm rate or the miss rate at the point where the 45° line intersects the detection error tradeoff (DET) curve.

We used normalized cross-entropy (NCE) as the second performance metric. We computed it as

$$\mathrm{NCE} = 1 - \frac{\mathrm{H}_{\mathrm{cond}}}{\mathrm{H}_{\mathrm{base}}} \tag{23}$$

where

$$\mathrm{H}_{\mathrm{cond}} = -\sum_{t=1}^{T} \Big(y_t \log p(y=1|\mathbf{x}_t) + (1-y_t) \log p(y=0|\mathbf{x}_t) \Big) \tag{24}$$

$$\mathrm{H}_{\mathrm{base}} = -T_1 \log\left(\frac{T_1}{T}\right) - (T - T_1) \log\left(\frac{T - T_1}{T}\right). \tag{25}$$

T is the total number of word hypotheses, T_1 is the number of correct words, and y_t is the label of the t-th word (1 for correct and 0 for incorrect). A small value of $\mathrm{H}_{\mathrm{cond}}$ indicates that the pdfs of the true error labels and estimated confidence scores are close. Division by $\mathrm{H}_{\mathrm{base}}$ normalizes for the chance case when all the word confidence scores equal the prior probability of the word being correct. NCE thus increases with better confidence score estimates.

Table IV shows the NCE values and Table V shows the EER values for the various ASR systems and test set speakers. We obtain an average NCE of 0.41 and an average EER of 0.17 over all the test cases. We next evaluate the utility of the estimated ASR confidence scores by using them in N-best ROVER.

B. ROVER WER

We used the ROVER algorithm to perform fusion of 10-best lists obtained from different ASR systems described in Section IV-A. We adopted three strategies for this fusion which allowed us to test the diversity-ROVER WER link across different variants of the fusion rule.

1) Word Frequency-based ROVER: The traditional word frequency-based ROVER was the first fusion scheme. It results by giving all the weight to the average word frequency vector $\mathbf{w}_i^{\mathrm{avg}}$ for each cohort set $\mathcal{C}_i$ in (4) by setting $\alpha = 1$. The first block of rows in Table VI shows the WERs obtained after 10-best fusion for different systems. We observe a reduction in WER with respect to the 1-best WER in Table III for 19 out of the 27 cases for 10-best fusion within each model (M1, M2, and M3) and within each training set (HUB4, WSJ, and ICSI). We next compare the WERs for the M1+M2+M3 row with the WERs of the individual systems and observe that the ROVER WER is lower than the best component system's WER in 5 out of 9 cases. We observe that 10-best ROVER out-performs the WER of the best component system in 2 out of 9 cases when fusing across training sets in the last column block under WSJ+HUB4+ICSI. The WERs for fusion across models and training sets shows a similar trend. This leads us to conclude that the word frequency-based ROVER algorithm may not always reduce the WER beyond the WER of the best component ASR system. Our experiments in Section V-C give a possible reason for this observation by showing that this variant of ROVER is unable to effectively utilize the inherent diversity in the N-best list.

2) Oracle Confidence-based ROVER: We next evaluated the WER of the oracle confidence-weighted ROVER algorithm to find out the lower bound on the WER after N-best fusion. Each hypothesis word in the N-best list is assigned a confidence score of 1 if it is correct with respect to the human-generated transcriptions and 0 otherwise. We used this oracle confidence score in (4) and tuned the trade-off parameter α using cross-validation. The second block of rows in Table VI shows the WER after oracle confidence-weighted ROVER for various cases. We observe a significant reduction in WER with respect to the unweighted ROVER WER for all cases. The best performance is obtained when the 10-best list is composed of decoded sentences from all three models and ASR systems trained on all three training data sets.[7] This is intuitive because such a 10-best list is very diverse and each sentence hypothesis makes complementary errors. The oracle confidence-weighted ROVER fusion rule is expected to utilize this diversity much better than simple word frequency-based ROVER. We illustrate this point through further experiments using the ambiguity decomposition in section V-C.

[7] We have highlighted the corresponding WERs in bold font in Table VI.

Normalized Cross-Entropy (NCE)									
	HUB4			WSJ			ICSI		
	MR	BO	JL	MR	BO	JL	MR	BO	JL
M1	0.3735	0.3884	0.3776	0.3387	0.316	0.3181	0.3105	0.3398	0.276
M2	0.4682	0.5344	0.4306	0.4502	0.4791	0.4191	0.4032	0.4573	0.3928
M3	0.4628	0.5301	0.4321	0.4508	0.473	0.4194	0.3994	0.4565	0.3692

TABLE IV

Table IV summarizes the testing set normalized cross-entropy (NCE) for ASR system confidence estimation. Higher values of NCE indicate better ASR confidence estimates. Perfect ASR confidence estimates give an NCE of 1.

Equal Error Rate (EER)									
	HUB4			WSJ			ICSI		
	MR	BO	JL	MR	BO	JL	MR	BO	JL
M1	0.1807	0.1814	0.1786	0.2	0.1946	0.2034	0.2101	0.2006	0.21
M2	0.15	0.1334	0.1721	0.1621	0.1574	0.1747	0.1759	0.1672	0.1801
M3	0.155	0.1334	0.1676	0.1563	0.1597	0.1789	0.1764	0.1624	0.1872

TABLE V

Table V summarizes the testing set equal error rate (EER) for ASR system confidence estimation. Lower values of EER indicate better ASR confidence estimates. Perfect ASR confidence estimates give an EER of 0.

	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	45.6	39.2	47.7	55.3	39.8	44.1	53.6	31.7	34.3	47.3
M2	29	30.4	44	37	44.2	58.7	35.3	41.1	54.6	29	31.6	44.8
M3	28.2	29.7	43.4	35.8	42.8	58.2	33.9	39.4	52.8	27.9	30.4	43
M1 + M2 + M3	28	30	43.5	35.8	41.6	55	34.3	39.2	50.7	29.5	33.8	45.2
10-best Oracle Confidence-Weighted ROVER WERs (α = 0.65 after 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**
10-best CRF Confidence-Weighted ROVER WERs (α = 0.8 for $^{\#}$, 0.9 for *, and 0.85 for all other cases, after cross-validation)												
M1	31.6	33.9	45.8	39.1	47.6	55	39.7	44.1	53.2	31.6	34.1	46.9
M2	29	30.3	43.9	36.9	44	58.5	35.1	41	54.6	28.6	31.3	45.3
M3	28	29.6	43.7	35.7	42.7	58	33.7	39.2	53	27.3	29.6	42.4
M1 + M2 + M3	27.9*	29.8*	43.2*	35.5*	41.3*	55*	33.9*	39.1*	50.6*	28.4$^{\#}$	32.3$^{\#}$	45.2$^{\#}$

TABLE VI

Table VI 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. Oracle confidence-based ROVER performs appreciably better than unweighted ROVER while the CRF confidence-based ROVER gives a minor improvement.

3) CRF Confidence-based ROVER: Our final fusion rule uses the confidence scores generated by the CRF-based system described in Section IV-B for ROVER fusion. We again tuned the trade-off parameter α using cross-validation. The third (last) block of rows in Table VI shows the WER after using the CRF confidence scores for ROVER fusion. We observe that the CRF confidence weighted ROVER reduces WER beyond the unweighted ROVER in 40 out of all 48 cases. This reduction in WER ranges from 0.1% to 1.5% (absolute). This indicates that while using the CRF confidence scores is better than word frequency-based ROVER, there is still a big gap in WER with respect to the lowest achievable WER after ROVER fusion using the oracle confidence scores. This is despite the acceptable values of EER and NCE for the ASR confidence estimation as shown in Tables V and IV respectively. We thus expect the CRF confidence-weighted ROVER to utilize N-best diversity slightly better than the traditional variant but worse than the oracle ROVER. Our experiments in the next section confirm this assertion.

C. Analysis of Diversity-ROVER WER Link

This section presents our experiments to validate the theoretical link between N-best diversity and ROVER WER presented in Section III. We use Theorems 1 and 2 to derive our proposed link between diversity and ROVER WER in Theorem 3. The proof of Theorem 3 simply requires substituting Theorem 1 in Theorem 2. Hence, we first empirically validate Theorems 1 and 2. We generated confusion networks for the combined N-best list and reference sentence hypothesis over all system combination variations on the test set in Table VI.

Figures 2(a) and 2(b) show the PDFs of the error between the left (approximate ROVER WER) and right hand sides of the two bounds in Theorem 1. The PDF of the error for the upper bound should ideally have all its density supported on the negative real axis, while the PDF for the lower-bound error should be supported on the positive real axis. Figures 2(a) and 2(b) show that this is true for an overwhelming majority of the test instances ($\geq 93\%$). However, Theorem 1 assumes that the word errors are IID, which is not the case in practice. Hence, the bounds are violated for a rare minority of the test instances, as indicated by the area under the two PDFs on their respective wrong sides of 0.

Figures 3(a) and 3(b) show the scatter plot between the approximate ROVER WER in Theorem 2 and the difference of the average N-best WER and the N-best diversity. Theorem 2 says that both these quantities are equal. This is evident from Figures 3(a) and 3(b) for all test instances across all system combination scenarios in Table VI.

To gain further insight into the nature of our proposed diversity-ROVER WER link, we performed more experiments on Theorem 3. This theorem relates the ROVER WER to average N-best WER and N-best diversity through upper and lower bounds. In order to compensate for the error in these bounds and provide interpretability to the experiments, we instead consider the following version of the decomposition for a given utterance with N cohort sets:

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i,\mathbf{h}_i^*)}_{\text{ROVER WER}} \approx \underbrace{\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{r}_i,\mathbf{h}_i^m)}_{\text{Avg. N-best WER}} - \gamma \underbrace{\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{h}_i^*,\mathbf{h}_i^m)}_{\text{N-best Diversity}} \tag{26}$$

where γ is a parameter that we learn through least squares regression. γ also provides us additional interpretation. It equals the decrease in ROVER WER with a unit increase in N-best diversity keeping the average N-best WER constant. It thus denotes the degree to which a given ROVER fusion rule utilizes diversity in the N-best list. Higher values of γ indicate a more diversity-sensitive fusion rule.

Figures 4 and 5 show the scatter plots of the ROVER WER with the average N-best WER and the N-best diversity, respectively, using the oracle confidence scores. As expected, the ROVER WER reduces with decreasing average WER of the N-best list in Figure 4. It is interesting to note from Figure 5 that the ROVER WER shows a decreasing trend with increasing diversity of the N-best list. Both Figures 4 and 5 however show only a moderately linear correlation.

We next performed least squares linear regression on the ROVER WER using both the average N-best WER and N-best diversity to check the proposed approximate decomposition in (26). Figure 6 shows the resulting scatter plot. The correlation coefficient is significantly higher than in Figures 4 and 5. We also obtain an appreciably small RMSE of 2.89. These results highlight the accuracy of the proposed approximate decomposition for the ROVER WER in (26). The estimated γ of 0.56 shows that the ROVER WER decreases by 0.56% (absolute) with a unit increase in the N-best diversity keeping the average N-best WER constant.

Figures 4 and 5 also highlight the trade-off between average N-best WER and N-best diversity predicted by the ambiguity decomposition. We consider the N-best list produced by the M3 model trained on the HUB4 data set, and a combination of N-best lists from the M1, M2, and M3 models trained on the same data set. Both have identical ROVER WERs close to 30%. Figure 4 shows that the M3 N-best list has an significantly lower average WER than the combined M1+M2+M3 N-best list. This is intuitive because the sentence hypotheses from M1 and M2 models have a higher WER than the hypotheses from the M3 models. However, Figure 5 shows that M1+M2+M3 N-best list also has a much higher N-best diversity than the M1 N-best list. The approximate ambiguity decomposition in (26) says that this higher diversity compensates for the higher N-best WER for the M1+M2+M3 N-best list and the fused hypotheses have similar WER for the two cases, as apparent from Figure 6.

Word frequency-based ROVER			
α	**MR**	**BO**	**JL**
1.00	0.898 (0.539,0.968)	0.902 (0.524,0.971)	0.869 (0.285,0.966)
Oracle Confidence ROVER			
α	**MR**	**BO**	**JL**
0.65	0.916 (0.607,0.980)	0.912 (0.604,0.976)	0.868 (0.297,0.970)
CRF Confidence ROVER			
α	**MR**	**BO**	**JL**
0.85	0.889 (0.501,0.966)	0.897 (0.556,0.965)	0.866 (0.251,0.967)

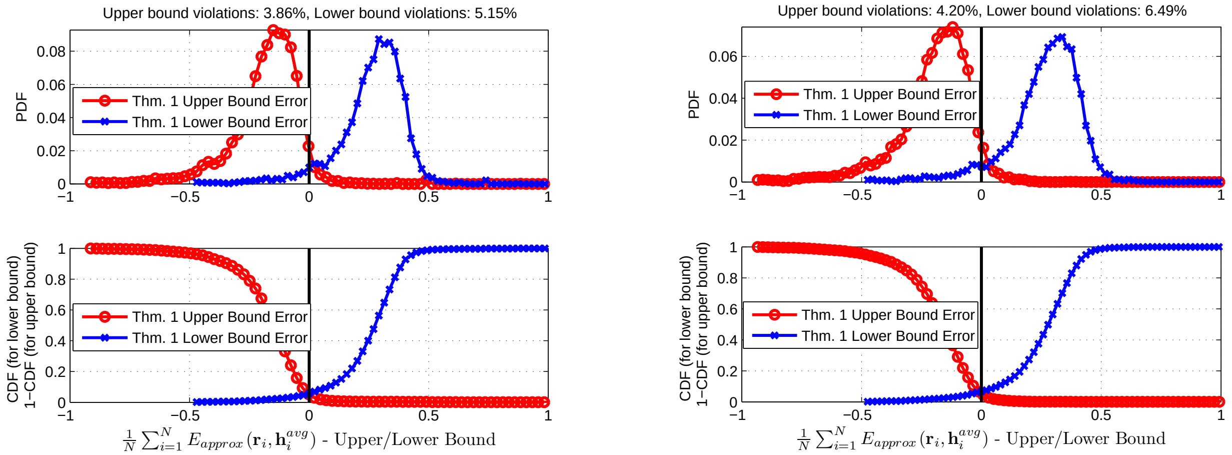
TABLE VII

TABLE VII SHOWS THE MEDIAN PER-UTTERANCE CORRELATION COEFFICIENTS BETWEEN THE ROVER WER AND ITS OPTIMAL APPROXIMATION IN (26). THIS TABLE ALSO CONTAINS THE 90% BOOTSTRAP CONFIDENCE INTERVALS FOR THE CORRELATION COEFFICIENTS. WE OBSERVE THAT ALL CORRELATION COEFFICIENTS ARE CLOSE TO 0.9 AND SIGNIFICANT AT THE 10% LEVEL.

We next present statistical significance tests over the entire test set for the three speakers. Table VII shows the median per-utterance correlation coefficient for each speaker between the ROVER WER and the right-hand side of (26) with optimal γ found using least squares regression. We present results for the three ROVER fusion rules - word frequency-based ($\alpha = 1$), oracle confidence-based, and CRF confidence-based. The α for the latter two rules were found using cross-validation and are the same as in section V-B. We observe that all correlation coefficients are close to 0.9 and significant at the 10% level using a bootstrap confidence interval. This indicates that the proposed approximate decomposition accurately predicts the true ROVER WER.

We next compare the three ROVER fusion rules with respect to their sensitivity to N-best diversity. The WER results in section V-B showed that ROVER with oracle confidence scores gave the lowest WER after fusion. This was followed by the CRF confidence-weighted ROVER. The word frequency-based ROVER gave the least improvement over the 1-best WER. We now provide an explanation for this observation based on the proposed decomposition. Table VIII shows the median per-utterance γ for the three ROVER fusion rules across speakers in the test set. (26) says that γ is the sensitivity of the fusion rule to diversity in the N-best list. More sensitive fusion rules are expected to have a higher γ and thus utilize the N-best diversity better, leading to lower WER upon ROVER fusion.

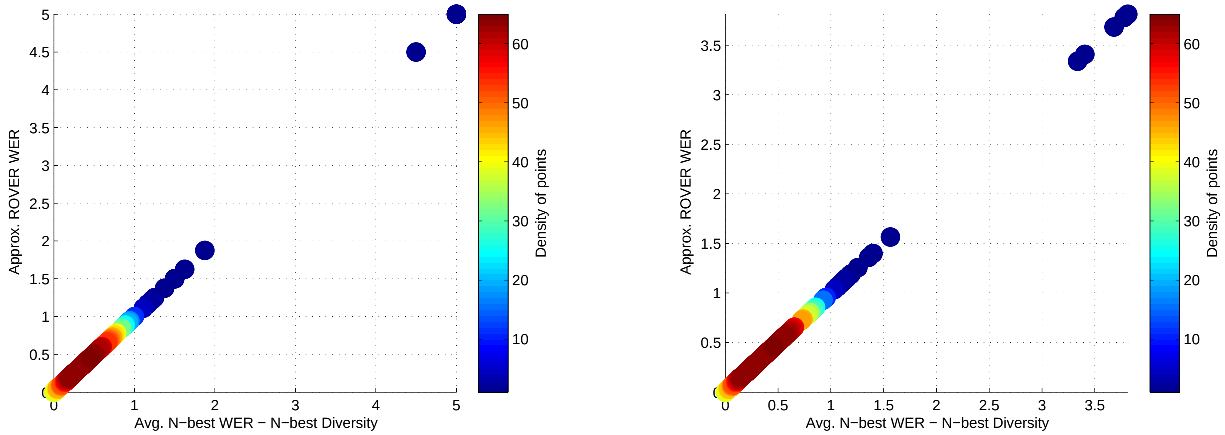
Table VIII shows that the oracle confidence ROVER has the highest median γ for all test speakers out of the three fusion strategies. This is intuitive because it uses the true word error label as a confidence score and is thus able to give low emphasis to erroneous words in each confusion bin during fusion.



(a) $\alpha = 1.00$

(b) Optimal α tuned on development set

Fig. 2. This figure shows the probability density function (pdf) and cumulative distribution function (cdf) for the error in the bounds in Theorem 1. We observe that the bounds hold for an appreciably high fraction ($\geq$ 93%) of the decoded test files over all system combination variations in Table VI. The fraction of the few bound violations is denoted by the height of the curves in the two bottom figures at the point of intersection with the black 0 line. These violations occur because Theorem 1 assumes word errors to be IID, which is not the case in practice.



(a) $\alpha = 1.00$

(b) Optimal α tuned on development set

Fig. 3. This figure shows the scatter plot between the approximate ROVER WER and the difference of the average N-best WER and N-best diversity for all decoded test files over all system combination variations in Table VI. Theorem 2 says that the approximate ROVER WER equals the difference of the average N-best WER and the N-best diversity. The above plots illustrate this because all the points lie on the 45^o line.

This leads to better utilization of diversity or complementarity in the N-best list. The CRF confidence-weighted ROVER gives the next highest γ because the confidence scores generated by the CRF model are not perfect as indicated by the EER and NCE results in section V-A. Hence some words with high confidence might actually be incorrect and may thus degrade ROVER fusion performance. These errors in ASR confidence estimation thus prevent ROVER from taking advantage of the inherent diversity in the N-best list. The word frequency-based ROVER weighs each system equally during fusion and is thus totally oblivious to word errors. Hence it gives the highest WERs out of the three schemes.

We have established and investigated into our proposed theoretical link between ASR diversity and ROVER WER in this section through several experiments. The next section gives a general discriminative framework for jointly training diverse ASR systems which utilizes the decomposition presented in this paper. We also show that some recent approaches for training diverse ASR systems are special cases of our framework.

VI. A Unified Discriminative Approach for Jointly Training Diverse ASR Systems

Prior work has used intuitive algorithmic modifications to train diverse ASR systems as discussed in Section I. However, some recent works have also focused on explicit discriminative training algorithms for this purpose [59]–[61]. Our proposed theoretical link between diversity and ROVER WER unifies these approaches and motivates a principled approach for jointly training diverse ASR systems in a discriminative fashion.

Let the M ASR systems have parameters $\mathbf{\Theta} = \{\Theta_1, \ldots, \Theta_M\}$. Consider a given training data set of T

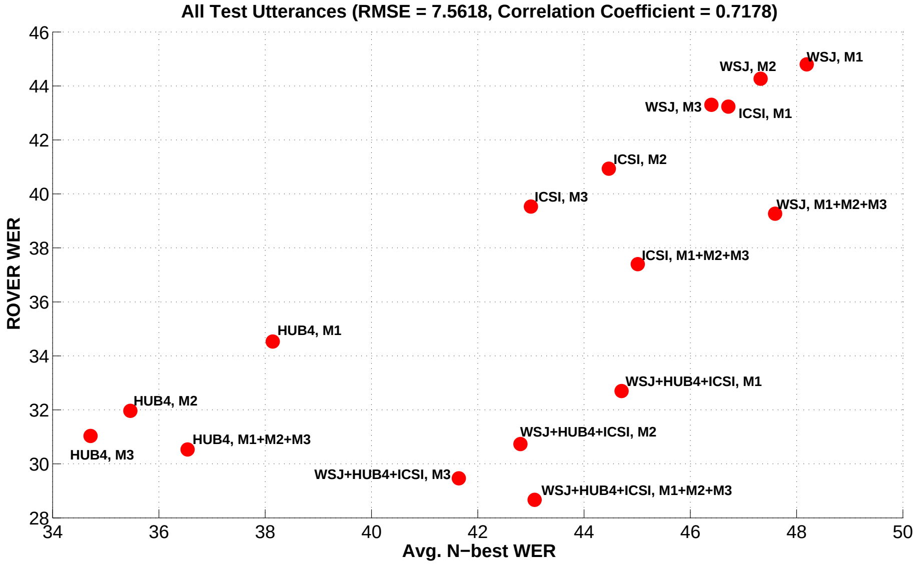


Fig. 4. This figure shows the scatter plot between ROVER WER and average N-best WER of different 10-best lists combined using the oracle confidence-weighted ROVER algorithm. We have averaged the WERs across all test utterances and ignored any impact of N-best diversity in this plot. We observe a correlation coefficient of 0.718 and root mean-squared error (RMSE) of 7.562 between average N-best WER and the ROVER WER. Hence in general, ROVER WER decreases as the ASR systems being combined become more accurate.

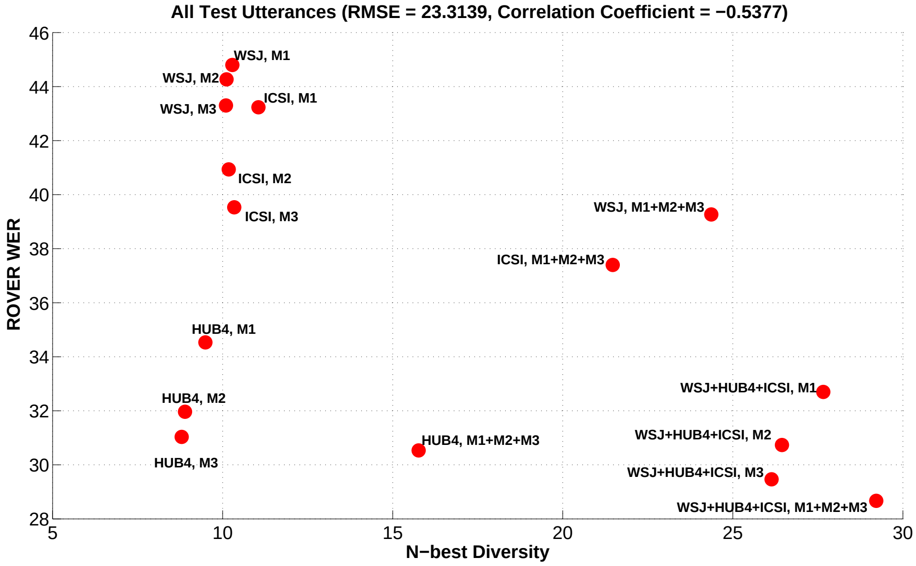


Fig. 5. This figure shows the scatter plot between ROVER WER and diversity of different 10-best lists combined using the oracle confidence-weighted ROVER algorithm. We have averaged the WERs across all test utterances and ignored any average N-best WER in this plot. We observe a correlation coefficient of -0.538 between N-best diversity and the ROVER WER. Hence in general, ROVER WER decreases as the ASR systems being combined become more diverse.

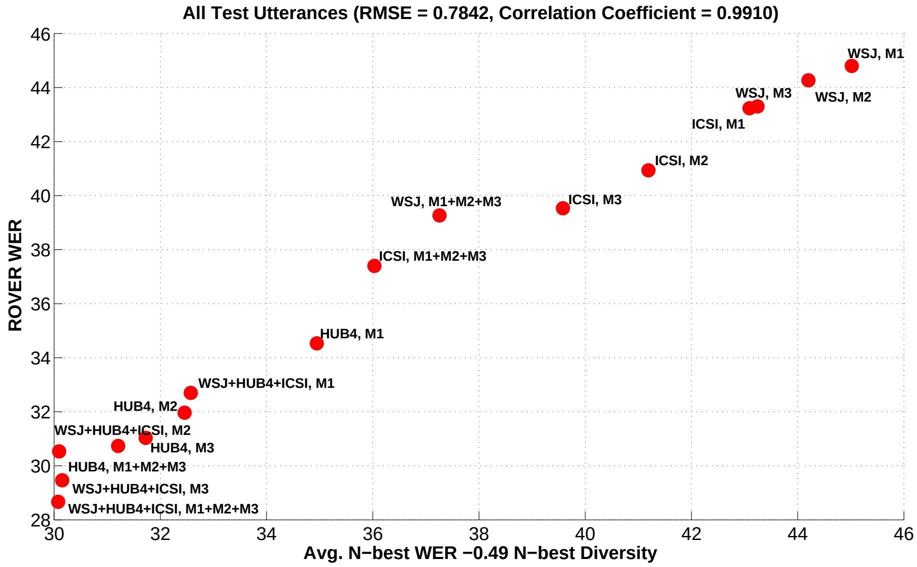


Fig. 6. This figure shows the scatter plot between ROVER WER and an optimal linear combination of average N-best WER and diversity of different 10-best lists combined using the oracle confidence-weighted ROVER algorithm. We have averaged the WERs across all test utterances and computed the coefficient $\gamma = 0.49$ using least-squares linear regression. We observe a correlation coefficient of 0.991 and RMSE of 0.784 between the optimal linear combination and the ROVER WER. Thus this optimal linear combination predicts the ROVER WER better than the average N-best WER and diversity considered individually.

Word frequency-based ROVER			
α	**MR**	**BO**	**JL**
1.00	0.145	0.120	0.134

Oracle Confidence ROVER			
α	**MR**	**BO**	**JL**
0.65	$0.508^{*\#}$	$0.468^{*\#}$	$0.404^{*\#}$

CRF Confidence ROVER			
α	**MR**	**BO**	**JL**
0.85	0.186^{*}	0.240^{*}	0.208^{*}

TABLE VIII

TABLE VIII SHOWS THE MEDIAN PER-UTTERANCE γ IN (26) ESTIMATED USING LEAST SQUARED REGRESSION BETWEEN THE ROVER WER AND ITS OPTIMAL APPROXIMATION. * INDICATES THAT THE γ IS SIGNIFICANTLY HIGHER THAN THE CORRESPONDING γ FOR $\alpha = 1$ USING WILCOXON'S SIGNED RANK TEST AT THE 10% SIGNIFICANCE LEVEL. $\#$ INDICATES THAT THE γ IS SIGNIFICANTLY HIGHER THAN THE CORRESPONDING γ FOR $\alpha = 0.85$ USING CRF CONFIDENCE-WEIGHTED ROVER. WE OBSERVE THAT THE ORACLE CONFIDENCE ROVER IS MOST SENSITIVE TO DIVERSITY IN THE N-BEST LIST DUE TO ITS SIGNIFICANTLY HIGHER γ, FOLLOWED BY THE CRF CONFIDENCE ROVER AND THE WORD FREQUENCY-BASED ROVER.

audio files with observed acoustic feature vector sequences $\boldsymbol{\mathcal{O}} = \{\mathcal{O}_1, \ldots, \mathcal{O}_T\}$ and reference words sequences $\boldsymbol{\mathcal{W}} = \{\mathcal{W}_1, \ldots, \mathcal{W}_T\}$. Let the M ASR systems produce word hypothesis sequences $\{\mathcal{H}_t^1, \ldots, \mathcal{H}_t^M\}$ for $\mathcal{O}_t$. We consider a multi-system version of the minimum Bayes risk (MBR) training objective function [62] because it generalizes other popular discriminative training objective functions[8] such as ones used in the maximum mutual information (MMI) [40], minimum word error (MWE) [64], and minimum word error (MPE) training. Our joint MBR optimization problem minimizes the expected WER of the ROVER word hypothesis sequence $\mathcal{H}_t^*(\alpha)$ obtained by combining the M hypotheses for $\mathcal{O}_t$. This optimization problem is

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} \sum_{t=1}^{T} \sum_{\mathcal{H}_t^1} \ldots \sum_{\mathcal{H}_t^M} \mathcal{E}\Big(\mathcal{W}_t, \mathcal{H}_t^*(\alpha); \boldsymbol{\Theta}\Big) \times P\Big(\mathcal{H}_t^1, \ldots, \mathcal{H}_t^M | \mathcal{O}_t; \boldsymbol{\Theta}\Big) \qquad (27)$$

where $\mathcal{E}(.)$ computes the WER between two sentence hypotheses and $P(\mathcal{H}_t^1, \ldots, \mathcal{H}_t^M | \mathcal{O}_t; \boldsymbol{\Theta})$ is the joint probability density function (pdf) of the sentence hypotheses from the M ASR systems. Thus $\mathcal{E}(.)$ equals the average of the confusion bin WERs $E(.)$ over each audio file which we used in the prior sections. The above optimization problem is difficult to solve because of the expectation with respect to the joint pdf of hypotheses from all M ASR systems and the use of the ROVER hypothesis $\mathcal{H}_t^*(\alpha)$. We now show how the proposed decomposition in this paper considerably simplifies the MBR objective function in (27).

We first upper-bound the WER of the ROVER sentence hypothesis $\mathcal{H}_t^*(\alpha)$ by the average WER of the sentence hypotheses and the diversity using Theorem 3:

$$\mathcal{E}\Big(\mathcal{W}_t, \mathcal{H}_t^*(\alpha); \boldsymbol{\Theta}\Big) \leq \frac{1}{M} \sum_{m=1}^{M} \mathcal{E}\Big(\mathcal{W}_t, \mathcal{H}_t^m; \Theta_m\Big) - \frac{\gamma}{M} \sum_{m=1}^{M} \mathcal{E}\Big(\mathcal{H}_t^*(\alpha), \mathcal{H}_t^m; \Theta_m\Big) + \text{constants} . \qquad (28)$$

The constants in the above equation depend on the ROVER parameter α and M as discussed in Section III, and are independent of $\boldsymbol{\Theta}$. We hence ignore them. The trade-off parameter γ is non-negative. We note that the diversity term still depends on the ROVER hypothesis $\mathcal{H}_t^*(\alpha)$ which makes joint training difficult because we need to sum over all possible hypotheses sequences from all M ASR systems in (27). However, $\mathcal{E}$ is just

[8] Heigold et. al [63] provide an excellent overview of various discriminative training algorithms for ASR.

the Levenshtein string metric and we use the triangle inequality for any two pairs of hypotheses sentences $\mathcal{H}_t^m$ and $\mathcal{H}_t^n$:

$$\mathcal{E}\Big(\mathcal{H}_t^n, \mathcal{H}_t^m; \Theta_m, \Theta_n\Big) \leq \mathcal{E}\Big(\mathcal{H}_t^*(\alpha), \mathcal{H}_t^m; \Theta_m\Big) + \mathcal{E}\Big(\mathcal{H}_t^*(\alpha), \mathcal{H}_t^n; \Theta_n\Big) \quad \forall m, n \in \{1, \ldots, M\} \,. \tag{29}$$

Adding the above inequalities over all possible unique pairs of hypotheses gives the following pairwise lower-bound on the diversity term:

$$\sum_{m=1}^{M} \mathcal{E}\Big(\mathcal{H}_t^*(\alpha), \mathcal{H}_t^m; \Theta_m\Big) \geq \frac{1}{(M-1)} \times \sum_{m=1}^{M} \sum_{n=m+1}^{M} \mathcal{E}\Big(\mathcal{H}_t^m, \mathcal{H}_t^n; \Theta_m, \Theta_n\Big) \,. \tag{30}$$

Hence the upper-bound on the WER of the ROVER hypothesis in (28) can be relaxed to the following upper-bound that is independent of the ROVER hypothesis $\mathcal{H}_t^*(\alpha)$:

$$\mathcal{E}\Big(\mathcal{W}_t, \mathcal{H}_t^*(\alpha); \boldsymbol{\Theta}\Big) \leq \frac{1}{M} \sum_{m=1}^{M} \mathcal{E}\Big(\mathcal{W}_t, \mathcal{H}_t^m; \Theta_m\Big) - \frac{\gamma}{M(M-1)} \sum_{m=1}^{M} \sum_{n=m+1}^{M} \mathcal{E}\Big(\mathcal{H}_t^m, \mathcal{H}_t^n; \Theta_m, \Theta_n\Big) \,. \tag{31}$$

The above bound enables us to marginalize the joint pdf in the diversity term and gives the following relaxation to the original joint MBR problem in (27):

$$\boldsymbol{\Theta^*} = \arg\min_{\boldsymbol{\Theta}} \sum_{m=1}^{M} \Big[\sum_{t=1}^{T} \sum_{\mathcal{H}_t^m} \mathcal{E}\Big(\mathcal{W}_t, \mathcal{H}_t^m; \Theta_m\Big) P(\mathcal{H}^m | \mathcal{O}_t, \Theta_m) \Big] - \frac{\gamma}{(M-1)} \sum_{m=1}^{M} \sum_{n=m+1}^{M} \Big[\sum_{t=1}^{T} \sum_{\mathcal{H}_t^m} \sum_{\mathcal{H}_t^n} \mathcal{E}\Big(\mathcal{H}_t^m, \mathcal{H}_t^n; \Theta_m, \Theta_n\Big) \times P\Big(\mathcal{H}_t^m, \mathcal{H}_t^n \Big| \mathcal{O}_t, \Theta^m, \Theta^n\Big) \Big] \,. \tag{32}$$

The first term in the above objective function is proportional to the average Bayes risk for all M ASR systems, and the second term is proportional to the average pairwise Bayes risk of the systems. This latter diversity term is significantly easier to compute than the joint diversity term in (27) because it involves a sum over all hypothesis from pairs of ASR systems.

Setting $\gamma = 0$ in (32) leads to disjoint MBR training of the M ASR systems without any explicit diversity criterion, though the systems can be made diverse through implicit techniques such as randomized decision trees [12]. For $\gamma > 0$, the optimization problem in (32) can be solved iteratively over each ASR system. Thus the parameters for system m are estimated such that they have low Bayes risk with respect to the reference transcriptions and high average Bayes risk or diversity from all other $(M-1)$ ASR systems.

Our joint optimization approach is different from the MBR leveraging (MBRL) algorithm by Breslin and Gales [60]. MBRL is a sequential training algorithm where the m-th ASR system is trained with respect to the previously trained $(m-1)$ systems. The confusion networks from the previous $(m-1)$ systems are aligned with the reference to find incorrectly predicted words, which are then assigned higher loss $\mathcal{E}$ while training the m-th ASR system. Our objective function in (32) is an upper-bound on the Bayes risk of the ROVER hypothesis. Jointly training the M ASR systems via (32) thus directly impacts the average fusion performance.

Tachioka and Watanabe [61] propose an MMI-based discriminative training criterion for training diverse ASR systems. The m-th system is trained by maximizing the mutual information with the correct word sequence and minimizing the average mutual information to the hypotheses word sequences to the $(m-1)$ base systems. This is a special case of our ambiguity decomposition-based formulation in (32) because the MBR objective with a $0/1$ error function $\mathcal{E}$ reduces to the MMI objective. Our approach also similarly reduces to the complementary phone error (CPE) formulation by Diehl and Woodland [59] when the ASR systems are trained sequentially and only the 1-best hypotheses transcriptions from the previous $(m-1)$ systems are used for computing the diversity instead of taking an expectation over all possible hypotheses sequences as in (32).

VII. Conclusion and Future Work

This paper presented a theoretical basis of the link between the WER of the fused hypothesis generated by the ROVER algorithm and the diversity of the constituent ASR systems. We draw upon the ensemble methods literature in machine learning for our purpose by first presenting a vector space model for ROVER fusion. This enables us to approximately decompose the WER of the ROVER output in terms of the average WER of the sentence hypotheses being combined and the diversity of the N-best list. This decomposition gives a natural definition of N-best diversity - the spread of the individual sentence hypotheses about the ROVER output in the proposed vector space model. It also highlights the trade-off between average N-best WER and diversity. Sentence hypotheses with high WER can lower the ROVER WER provided they are diverse and the fusion rule is able to take advantage of them. We also refine our proposed approximate decomposition through upper and lower bounds on the error rate.

We next present experimental evidence for the accuracy of the proposed decomposition using multiple ASRs trained with the Kaldi toolkit and multiple ROVER fusion schemes. Our experiments also provide insights into different ROVER WERs for different fusion schemes based on the decomposition. Using the true (oracle) word error label as confidence score leads to lowest WER upon fusion because it is more sensitive to diversity in the N-best list as indicated by higher γ in (26). This is intuitively followed by ROVER which uses CRF-generated confidence scores and no confidence scores (only word frequencies) in order of increasing ROVER WER. We also present a unified minimum Bayes risk (MBR) approach for systematically training diverse AMs using the proposed decomposition. This discriminative training framework generalizes several recent attempts at incorporating diversity in the AM training objective function.

There are several interesting implications of this work. First and foremost, it provides theoretical insight into the often-observed empirical benefit of fusing diverse ASR systems using ROVER. This is especially important given the prevalence of real-world systems involving multiple ASR systems. Second, we believe that this work can motivate more principled approaches for training diverse ASR systems. Recent work on minimum Bayes risk leveraging [60], complementary phone error training [59], and diverse MMI training [61] for ASR systems are steps in the right direction[9]. We also believe that the proposed results can be useful in non-ASR contexts, such as ROVER fusion of transcriptions from multiple human annotators in a crowd-sourcing setting.

Appendix

A. The Link Between Bias-Variance and Ambiguity Decompositions

Consider a training data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N \subseteq \mathbb{R}^D \times \mathbb{R}$ for learning a regressor $f : \mathbb{R}^D \to \mathbb{R}$ by minimizing the squared error loss function. We know that the optimal regressor which minimizes squared error is the conditional expectation $E_{p(y|\mathbf{x})}\{y|\mathbf{x}\}$ or $\mathbb{E}\{y|\mathbf{x}\}$ to simplify the notation. We denote the learned regressor as $f^*_{\mathcal{D}}$ because it is a function of the training data set $\mathcal{D}$. Consider a given $\mathbf{x}$. The random variable $f^*_{\mathcal{D}}(\mathbf{x})$ might be an excellent approximation to $\mathbb{E}\{y|\mathbf{x}\}$ only for a specific choice of training data set $\mathcal{D}$. Hence, a better measure of the approximation error is the expected squared error between the optimal and learned regressors, where the expectation is taken over $p(\mathcal{D})$, i.e. a probability distribution over the space of training data sets.

The bias-variance decomposition [28] states for an input $\mathbf{x}$, this expected squared error decomposes as:

$$\mathbb{E}_{p(\mathcal{D})}\Big\{\Big(f^*_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}\{y|\mathbf{x}\}\Big)^2\Big\} = \Big(\mathbb{E}_{p(\mathcal{D})}\{f^*_{\mathcal{D}}(\mathbf{x})\} - \mathbb{E}\{y|\mathbf{x}\}\Big)^2 + \mathbb{E}_{p(\mathcal{D})}\Big\{\Big(f^*_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{p(\mathcal{D})}\{f^*_{\mathcal{D}}(\mathbf{x})\}\Big)^2\Big\} . \tag{33}$$

The first term on the right-hand side is the squared-bias of the learned regressor $f^*_{\mathcal{D}}(\mathbf{x})$ and computes the squared error between the mean prediction and the optimal estimate $\mathbb{E}\{y|\mathbf{x}\}$. The second term is the variance of the random variable $F^*_{\mathcal{D}}(\mathbf{x})$ over $p(\mathcal{D})$ and measures the spread of the predictions about the mean prediction.

We now show the relation between (33) and the ambiguity decomposition [24]. Let $p(\mathcal{D})$ be the following mixture model over M data sets:

$$p(\mathcal{D}) = \sum_{m=1}^M w_m \delta(\mathcal{D} - \mathcal{D}_m) . \tag{34}$$

Hence the expectation of any function $g(\mathcal{D})$ over $p(\mathcal{D})$ becomes the convex sum

$$\mathbb{E}_{p(\mathcal{D})}\{g(\mathcal{D})\} = \sum_{m=1}^M w_m g(\mathcal{D}_m) . \tag{35}$$

Substituting the above expression in (33) and re-arranging the terms gives the ambiguity decomposition

$$\Big(\sum_{m=1}^M w_m\{f^*_{\mathcal{D}_m}(\mathbf{x})\} - \mathbb{E}\{y|\mathbf{x}\}\Big)^2 = \sum_{m=1}^M w_m \Big(f^*_{\mathcal{D}_m}(\mathbf{x}) - \mathbb{E}\{y|\mathbf{x}\}\Big)^2 - \sum_{m=1}^M w_m \Big(f^*_{\mathcal{D}_m}(\mathbf{x}) - \sum_{n=1}^M w_n f^*_{\mathcal{D}_n}(\mathbf{x})\Big)^2 . \tag{36}$$

We conclude that the bias-variance decomposition equals the ambiguity decomposition only when all regressors have the same functional form f and are estimating the conditional expectation $\mathbb{E}\{y|\mathbf{x}\}$ as the target. The ambiguity decomposition does not impose such constraints.

B. Proofs

Proof of Lemma 1. Expand $E(\mathbf{r}_i, \mathbf{h}_i^*)$ in (6) as

$$E(\mathbf{r}_i, \mathbf{h}_i^*) = \frac{1}{2}\Big(\|\mathbf{r}_i\|_2^2 + \|\mathbf{h}_i^*\|_2^2 - 2\mathbf{r}_i^T\mathbf{h}_i^*\Big) . \tag{37}$$

Both $\|\mathbf{r}_i\|_2^2$ and $\|\mathbf{h}_i^*\|_2^2$ are 1 because $\mathbf{r}_i$ and $\mathbf{h}_i^*$ are 1-in-K bit vectors. Hence

$$1 - E(\mathbf{r}_i, \mathbf{h}_i^*) = \mathbf{r}_i^T\mathbf{h}_i^* . \tag{38}$$

Now $\{\mathbf{r}_i^T\mathbf{h}_i^*\}_{i=1}^N$ are IID samples from a Bernoulli random variable with parameter p. Hence the ML estimate of p is

$$p_{\text{ML}} = \frac{1}{N}\sum_{i=1}^N \mathbf{r}_i^T\mathbf{h}_i^* \tag{39}$$

$$= 1 - \frac{1}{N}\sum_{i=1}^N E(\mathbf{r}_i, \mathbf{h}_i^*) . \tag{40}$$

□

Proof of Lemma 2. We first prove the upper-bound on p_{ML}. Expand $E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}})$ as

$$E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) = \frac{1}{2}\Big(\|\mathbf{r}_i\|_2^2 + \|\mathbf{h}_i^{\text{avg}}\|_2^2 - 2\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}}\Big) \tag{41}$$

where $\|\mathbf{r}_i\|_2^2 = 1$ because $\mathbf{r}_i$ is a 1-in-K bit vector. Also

$$\|\mathbf{h}_i^{\text{avg}}\|_2^2 = \|\alpha\mathbf{w}_i^{\text{avg}} + (1-\alpha)\mathbf{s}_i^{\text{avg}}\|_2^2 \tag{42}$$

$$\leq \alpha\|\mathbf{w}_i^{\text{avg}}\|_2^2 + (1-\alpha)\|\mathbf{s}_i^{\text{avg}}\|_2^2 \tag{43}$$

due to (4) and Jensen's inequality for the convex squared-L_2 norm. We next use the fact that $\|\mathbf{w}_i^{\text{avg}}\|_2 \leq \|\mathbf{w}_i^{\text{avg}}\|_1$. Since $\|\mathbf{w}_i^{\text{avg}}\|_1 = \mathbf{1}^T\mathbf{w}_i^{\text{avg}} = 1$ because all entries of $\mathbf{w}_i^{\text{avg}}$ are non-negative and sum to 1, this gives us $\|\mathbf{w}_i^{\text{avg}}\|_2 \leq 1$. Similarly, $\|\mathbf{s}_i^{\text{avg}}\|_2 \leq \|\mathbf{s}_i^{\text{avg}}\|_1 = \mathbf{1}^T\mathbf{s}_i^{\text{avg}} \leq 1$ because the sum of entries of the average confidence score vector $\mathbf{s}_i^{\text{avg}}$ is less than or equal to 1. Hence we can write the upper-bound

$$\frac{1}{N}\sum_{i=1}^N E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) \leq 1 - \frac{1}{N}\sum_{i=1}^N \mathbf{r}_i^T\mathbf{h}_i^{\text{avg}} . \tag{44}$$

We now use the fact that $\{\mathbf{r}_i^T\mathbf{h}_i^{\text{avg}}\}_{i=1}^N$ are IID random variables with support $[0, 1]$ and mean p. Hence

$$p_{\text{ML}} = \frac{1}{N}\sum_{i=1}^N \mathbf{r}_i^T\mathbf{h}_i^{\text{avg}} . \tag{45}$$

[9] [65] presents a similar approach for training diverse maximum entropy models.

is the ML estimate of p. Thus (44) becomes

$$\frac{1}{N}\sum_{i=1}^{N} E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) \leq 1 - p_{\text{ML}} \ . \tag{46}$$

Re-arranging this gives us the desired upper-bound on p_{ML}. We use the following inequality for $\|\mathbf{h}_i^{\text{avg}}\|_2^2$ in place of the upper-bound in (43) to prove the lower-bound on p_{ML}:

$$\|\mathbf{h}_i^{\text{avg}}\|_2^2 \geq \alpha^2 \|\mathbf{w}_i^{\text{avg}}\|_2^2 + (1-\alpha)^2 \|\mathbf{s}_i^{\text{avg}}\|_2^2 \tag{47}$$

$$\geq \frac{\alpha^2}{M} \tag{48}$$

because $\|\mathbf{w}_i^{\text{avg}}\|_1 = \mathbf{1}^T\mathbf{w}_i^{\text{avg}} = 1 \leq \sqrt{M}\|\mathbf{w}_i^{\text{avg}}\|_2$ and $\|\mathbf{s}_i^{\text{avg}}\|_2^2 \geq 0$. We have utilized the fact that the *effective dimension* of $\mathbf{w}_i^{\text{avg}}$ is M and not K in the upper-bound on $\|\mathbf{w}_i^{\text{avg}}\|_1$. Rest of the steps of the proof for the lower-bound on p_{ML} remain the same as for the upper-bound. □

Proof of Theorem 1. We substitute the equality for p_{ML} from Lemma 1 into the inequalities for p_{ML} proved in Lemma 2 to prove the theorem. □

Proof of Theorem 2. We use the definition of the approximate ROVER WER from the last section and expand

$$2E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^m) = \|\mathbf{r}_i - \mathbf{h}_i^m\|_2^2 \tag{49}$$

$$= \|\mathbf{r}_i - \mathbf{h}_i^{\text{avg}} + \mathbf{h}_i^{\text{avg}} - \mathbf{h}_i^m\|_2^2 \tag{50}$$

$$= \|\mathbf{r}_i - \mathbf{h}_i^{\text{avg}}\|_2^2 + \|\mathbf{h}_i^{\text{avg}} - \mathbf{h}_i^m\|_2^2 + 2(\mathbf{r}_i - \mathbf{h}_i^{\text{avg}})^T(\mathbf{h}_i^{\text{avg}} - \mathbf{h}_i^m) \tag{51}$$

$$= 2E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) + 2E_{\text{approx}}(\mathbf{h}_i^{\text{avg}}, \mathbf{h}_i^m) + 2(\mathbf{r}_i - \mathbf{h}_i^{\text{avg}})^T(\mathbf{h}_i^{\text{avg}} - \mathbf{h}_i^m) \ . \tag{52}$$

Taking average of both sides over m and dividing by 2 gives

$$\frac{1}{M}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^m) = E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}}) + \frac{1}{M}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{h}_i^{\text{avg}}, \mathbf{h}_i^m) + (\mathbf{r}_i - \mathbf{h}_i^{\text{avg}})^T \left(\mathbf{h}_i^{\text{avg}} - \frac{1}{M}\sum_{m=1}^{M}\mathbf{h}_i^m\right) \ . \tag{53}$$

The final term on the right-hand side is zero because of the definition of $\mathbf{h}_i^{\text{avg}}$. Re-arranging the resulting equation gives us the ambiguity decomposition for ROVER WER. □

Proof of Theorem 3. We start with the result of Theorem 2 and substitute the lower-bound in (13) on the average of $E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^{\text{avg}})$. This gives

$$\frac{1}{N}\sum_{i=1}^{N} E(\mathbf{r}_i, \mathbf{h}_i^*) \leq \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^m) - \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{m=1}^{M} E_{\text{approx}}(\mathbf{h}_i^{\text{avg}}, \mathbf{h}_i^m) - \frac{1}{2}\left(\frac{\alpha^2}{M} - 1\right) \ . \tag{54}$$

We then use the fact that $E_{\text{approx}}(\mathbf{r}_i, \mathbf{h}_i^m) = E(\mathbf{r}_i, \mathbf{h}_i^m)$ because $\mathbf{h}_i^m$ is a 1-in-K bit vector. We finally again use the lower-bound in (13) on the average of $E_{\text{approx}}(\mathbf{h}_i^{\text{avg}}, \mathbf{h}_i^m)$ to give the desired upper-bound.

The proof of the lower-bound proceeds similarly as above except that we use the upper-bound in (12) instead of (13). □

References

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

[2] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[3] H. Soltau, G. Saon, B. Kingsbury, H. K. J. Kuo, L. Mangu, D. Povey, and A. Emami, "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.

[4] D. Stallard, F. Choi, C. L. Kao, K. Krstovski, P. Natarajan, R. Prasad, S. Saleem, and K. Subramanian, "The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System," in *Proc. Interspeech*, 2007.

[5] S. F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "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.

[6] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, and M. Graciarena, "The CALO meeting speech recognition and understanding system," in *Proc. SLT*. IEEE, 2008, pp. 69–72.

[7] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy, "Developing speech recognition systems for corpus indexing under the IARPA BABEL program," in *Proc. ICASSP*. 2013, IEEE.

[8] 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.

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

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

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

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

[13] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 519–528, 2008.

[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] X. Cui, J. Huang, and J.-T. Chien, "Multi-view and multi-objective semi-supervised learning for hmm-based automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1923–1935, 2012.

[21] X. Chen and Y. Zhao, "Building acoustic model ensembles by data sampling with enhanced trainings and features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3-4, pp. 498–507, 2013.

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

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

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

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

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

[27] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[28] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*. Dec. 2011, IEEE.

[30] J. Laferty, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001, pp. 282–289, Morgan Kaufmann.

[31] P. Xu, D. Karakos, and S. Khudanpur, "Self-supervised discriminative training of statistical language models," in *Proc. ASRU*. IEEE, 2009, pp. 317–322.

[32] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, pp. 181–207, 2003.

[33] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, no. 2, pp. 135–148, 2002.

[34] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University Engineering Department*, 2002.

[35] 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.

[36] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFST: A general and efficient weighted finite-state transducer library," *Implementation and Application of Automata*, pp. 11–23, 2007.

[37] 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.

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

[39] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, and A. Stolcke, "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–364.

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

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

[42] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[43] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proc. Eurospeech*. Rhodes, Greece: ESCA, 1997, vol. 2, pp. 827–830.

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

[45] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–809.

[46] M. S. Seigel and P. C. Woodland, "Combining information sources for confidence estimation with CRF models," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011, pp. 905–908.

[47] P. J. Moreno, B. Logan, and B. Raj, "A boosting approach for confidence scoring," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2001.

[48] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. IEEE, 1997, vol. 2, pp. 887–890.

[49] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML-01*, 2001, pp. 282–289.

[50] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[51] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[52] D. J. Litman, J. B. Hirschberg, and M. Swerts, "Predicting automatic speech recognition performance using prosodic cues," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 2000, pp. 218–225.

[53] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[54] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1, pp. 155–175, 2004.

[55] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge sources," in *Proc. Eurospeech*. Geneva, Switzerland, 2003, vol. 1, pp. 957–960.

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

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

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

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

[60] C. Breslin and M. J. F. Gales, "Generating complementary systems for speech recognition," in *Interspeech*, 2006.

[61] Y. Tachioka and S. Watanabe, "Discriminative training of acoustic models for system combination," in *Interspeech*, 2013.

[62] V. Doumpiotis and W. Byrne, "Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition," *Speech Communication*, vol. 48, no. 2, pp. 142–160, 2006.

[63] G. Heigold, H. Ney, R. Schluter, and S. Wiesler, "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 58–69, Nov. 2012.

[64] D. Povey, P. C. Woodland, and M. J. F. Gales, "Discriminative MAP for acoustic model adaptation," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–312.

[65] 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.