ESTIMATING CONFIDENCE SCORES ON ASR RESULTS USING RECURRENT NEURAL NETWORKS

Kaustubh Kalgaonkar, Chaojun Liu, Yifan Gong

Microsoft Corporation
{kakalgao, chaojunl, ygong}@microsoft.com

Kaisheng Yao

Microsoft Research
kaisheny@microsoft.com

ABSTRACT

In this paper we present a confidence estimation system using recurrent neural networks (RNN) and compare it to a traditional multi-layered perception (MLP) based system. The ability of RNN to capture sequence information and improve decisions using processed history was main motivation to explore RNN's for confidence estimation. In this paper we also explore two subtle variations of confidence estimator: one that uses objective extracted over the entire sequence for training, and other that uses dynamic programming to decode and estimate confidence on all the words of the sequence jointly.

In our experiments, we observed that for a constant false positive (FP) rate of 3% we can secure a relative reduction of 10% in false negative (FN) rate when we replaced a MLP in confidence estimator with a RNN. We also observed that relative gains achieved by a RNN based confidence estimator are directly proportional to the number of word in the utterances.

Index Terms— Confidence Measures, Word Identity, Recurrent Neural Network

1. INTRODUCTION

Automatic speech recognition (ASR) systems used today are able to produce high quality transcriptions, due to a combination of newer modeling techniques (e.g. deep neural networks), ability to consume and process large training and test datasets. As a result ASR systems are getting widely deployed, with voice search on smart-phones to voice controlled devices like Xbox-One. For these types of applications of ASR systems it has become imperative to provide some measure of confidence on the accuracy of the output of a recognizer. Depending on the requirement, a recognizer could produce either word or sentence/phrase confidence estimates.

Over the years various techniques have been proposed to estimate word confidences, Jiang in his survey paper [1] provides a comprehensive overview of techniques developed for estimating confidences. These techniques can be broadly classified into three categories. In the first, posterior probability of the word given the acoustic signal is regarded as the confidence measure [2, 3, 4, 5, 6]. The second set of techniques treat confidence estimation as a statistical hypothesis testing problem [7, 8, 9]. The third set of methods build a two-class classifier using features (also referred to as predictors) generated from an ASR engine during the decoding process [7, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. The parametric confidence estimation techniques (hypothesis-testing and classifier based) system usually tend to outperform a word posterior based systems, this is due to the fact that parametric techniques can use posteriors as one of the input parameters and improve the performance of a system using additional features from the engine.

Classifiers based estimators can be improved by either improving the input predictors as demonstrated by Huang et. al. in [19], or by improving the classification methodology. Over the years various types of classifier have been explored in literature e.g. linear discriminant functions [7, 10], generalized linear models [11, 12], Gaussian mixture models [13], neural networks [14, 15, 19, 20] and conditional random fields [21, 22] to name a few

The confidence estimator presented in this paper is based on third category of estimators. We present the use of recurrent neural network (RNN) as a classifier that also estimates confidences on the two classes (positive/negative). In recent years RNN have been successfully applied to tasks like language modeling [23] and language understanding [24]. RNN's have achieved great success at both these tasks mainly due to the presence of a recurrent layer which models context and history with relative ease. This leads us to believe that RNN's will be especially helpful in improving confidences estimates for longer utterance which have multiple words (e.g. short message dictations). To our knowledge, our work is the first to apply RNN's to the task of confidence estimation.

The remainder of this paper is organized as follows. Section 2 describes the topology of RNN's we use for confidence estimation. Section 3 outlines the experimental setup and discusses the results. Section 4 presents concluding remarks on the performance of RNN's.

2. RECURRENT NEURAL NETWORKS FOR CONFIDENCE ESTIMATION

The RNN's presented in this paper use a modified form of the classical Elman RNN architecture [25]; all the modifications were performed to make the classical RNN more suitable for the task of confidence estimation. This architecture is illustrated in Figure 1, where the RNN is '*unrolled*' across time to cover feature inputs for three consecutive words. The topology of RNN used for confidence estimation is simple; it has three layers, an input layer at the bottom, a hidden layer in the middle with recurrent connections (shown as dashed lines), and an output layer at top. The layers in the network are connected with weights denoted by the matrices $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{V}$. The input layer matrix $\mathbf{U}$ is an augmented matrix that also includes bias.

The RNN's used for either language understanding or language modeling have a discrete input layer $\mathbf{x}(t)$ with dimensionality equal to the size of the vocabulary, but for confidence estimation, however, input is a set of features in continuous space (described in Section 3.1), these features are usually derived from a word lattice. Due to the continuous nature of the feature space we do not employ the traditional 1-of-N coding for the input layer. The output layer $\mathbf{y}(t)$ has two nodes that emit confidence scores indicating if the word is

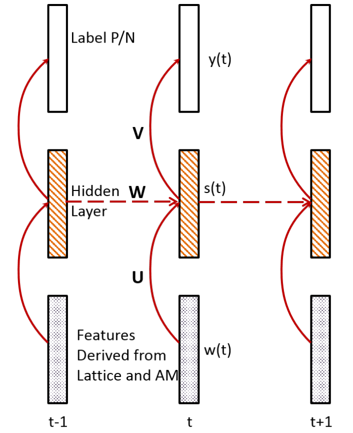


Fig. 1. Recurrent Neural Network for Confidence Estimation

correct or incorrectly identified. The recurrent hidden layer $\mathbf{s}(t)$ by design is responsible for maintaining a representation of utterance history and therefore, provides all the necessary context. The values in the hidden and output layers are computed as shown in equations (1a) and (1b)

$$\mathbf{s}(t) = f\left(\mathbf{U}[\mathbf{x}(t); 1] + \mathbf{W}\mathbf{s}(t-1)\right) \tag{1a}$$

$$\mathbf{y}(t) = g\left(\mathbf{V}\mathbf{s}(t)\right), \tag{1b}$$

where

$$f(l) = \frac{1}{1+e^{-l}}, \quad g(l_m) = \frac{e^{l_m}}{\sum\limits_k e^{l_k}}. \tag{2}$$

The RNN is trained using standard back-propagation through time algorithm to maximize the data conditional likelihood:

$$\prod_t P(\mathbf{y}(t)|\mathbf{x}(1)\ldots\mathbf{x}(t)) \tag{3}$$

Note that this model has no direct temporal interdependence at the output layer; this is due to the fact that the output probability distribution is strictly a function of the hidden layer activations, which in turn depend only on the input features derived over a words segment (and its own past values). Thus, the most likely sequence of labels (correct/incorrect) can be output with a series of online decisions:

$$\mathbf{y}^{\star}(t) = \arg\max P(\mathbf{y}(t)|\mathbf{x}(1)\ldots\mathbf{x}(t)) \tag{4}$$

This RNN topology is efficient yet simple. It does not require a dynamic programming search over labels to find the optimal sequence of output labels. However, we have observed that it is beneficial in some case to find the optimal correct/incorrect label sequence using dynamic programming. More details about training and sequence decoding using RNNs can be found in [24].

3. EXPERIMENTS AND RESULTS

3.1. Features for Confidence Estimation

All input features were derived from the lattice. These features were aggregated and averaged within word boundaries avoiding silence segments. Some of the features that we used as inputs to the confidence estimator are: normalized acoustic score, normalized background model score, normalized noise score, normalized LM score, normalized duration, LM perplexity, active channels, LM fanout and active senones. We used 16 features derived from the lattice as an input to the confidence estimator more details about these features can be found in [19].

3.2. Data, Models and Topologies

We conducted experiments using Microsoft's proprietary short message dictation and voice search data. The data used in our experiments was divided into three sets: 1) Train, 2) Validation, and 3) Test sets. The training set consists of 82K utterances and test set consists of 13K utterances (67K words), 2% of the training data was set aside for validation. All confidence estimation algorithms presented in this paper used same acoustic and language for all experiments. The RNN's used in our experiments have been trained using either cross-entropy or sequential criterion [26]. The weight matrices of both type of networks were initialized randomly at the start of training.

While experimenting on the validation set, we observed that the performance of the MLP confidence estimator peaked with 10 neurons in the hidden layer. Any further addition of hidden units did not result in a significant benefit, and therefore all MLP's presented in this paper have a single hidden layer with 10 neurons. This observation in-line with previous use of MLP for confidence estimation e.g. [27] uses an MLP with 6 or 8 nodes in the hidden layer.

The RNN based confidence estimator also followed trend similar to a MLP with respect to the nodes in the hidden layer. We did not observe any improvement by increasing the size of the hidden layer beyond 10 units. We also noticed that increasing the window of error back propagation through time (BPTT) beyond 4 did not yield any further improvements on the validation set.

3.3. Evaluation Metric

The performance of all our confidence estimators was evaluated using two metrics: receiver operating characteristic (ROC) curve was the primary tool used to compare different systems. we also used area under a ROC curve as the second metric.

A ROC curve in our case is a plot of false positive rates (**FPR**) against the true positive rates (**TPR**) rates calculated at the same confidence threshold value. A TP event at threshold '**t**'is the aggregate of all the words in the test set that are correctly recognized according to transcripts and have confidence higher than the selected threshold '**t**'; equivalently words that are mis-recognized with a confidence higher than '**t**'are aggregated to report a FP event.

We also use area under the ROC curve (AUC) as an alternative metric to measure performance of a confidence estimator. The area under a ROC curve is a good indicator of the overall ability of the confidence estimator to discriminate between correct and incorrectly recognized samples. Theoretically the best estimator would have an AUC of 1.0, therefore, to compare the performance of a confidence estimation algorithms/modules one needs to only measuring the closeness of its AUC to 1.0.

3.4. Deep Neural Networks for Confidence Estimation

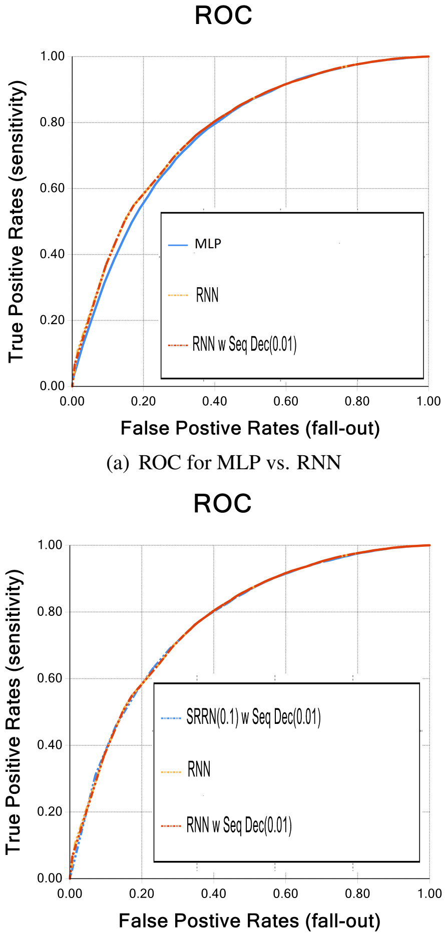
In recent years deep neural networks (DNN) have been successfully applied various classification tasks. In speech DNN have proven to be extremely successful as demonstrated in [28]. During experimentation we also performed tests where we replaced the simple MLP with a network that had multiple hidden layers.

We experimented with DNN that had upto 3 hidden layers and 10 to 100 hidden nodes. A DNN using the feature set described in

Section 3.1 did not demonstrate any significant improvement in AUC when compared to the MLP with 10 hidden nodes. This observation is similar to the results presented in [19, 29]. Our belief is that the features used in our task for confidence estimation are not very complex, and hence do not benefit from the ability of DNN to extract and distill discriminative information using multiple hidden deep layers.

Therefore, for the rest of the experiment and results section we do not include DNN as a comparative methodology for confidence estimation.

3.5. Results and Discussion

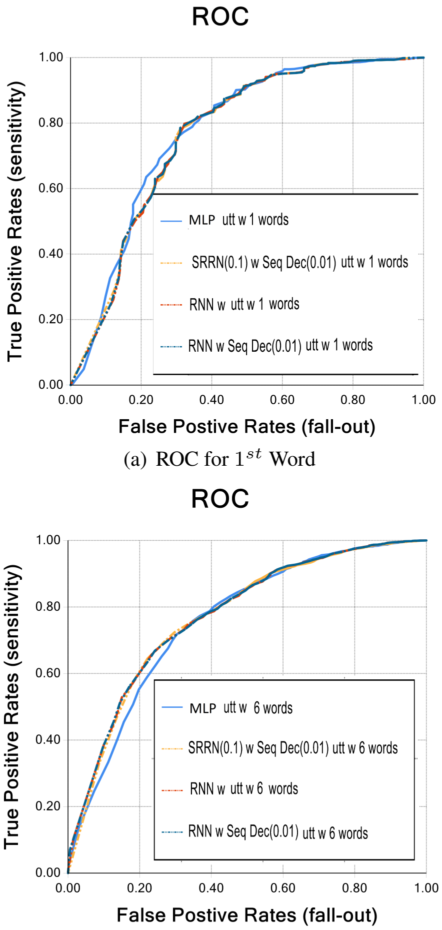


(a) ROC for MLP vs. RNN

(b) ROC for RNN vs. Seq-RNN

Fig. 2. ROC curves for different confidence estimators

The Figure 2(a) compares a RNN based system to a MLP based confidence estimator. We can observe from the ROC curves that a RNN is better than a MLP specifically in the regions of low FPR. The same chart also contains ROC curves for a confidence estimator where the output of a RNN was sequential decoded using bigram language model trained on the positive/negative symbols. The RNN and bigram language models was trained using the same data set. As we can observe from the figure, we obtain the biggest gain in AUC by replacing a MLP (AUC of 0.763) with a RNN (AUC of 0.78) (a relative improvement of 7.674% in AUC). Further by using sequence decoding, we observe a small incrementally improvement (relative AUC improvement of 7.676%) in the performance of the estimator. This result highlights the fact that the recurrent layer in the network is able to capture and extract enough information from the history so as to reduce any benefits that are traditionally achieved from sequential decoding. The Figure 2(b) compares a cross-entropy trained RNN system to a sequentially trained RNN system. Sequential training does not provide any additional benefit over a cross entropy trained RNNs.

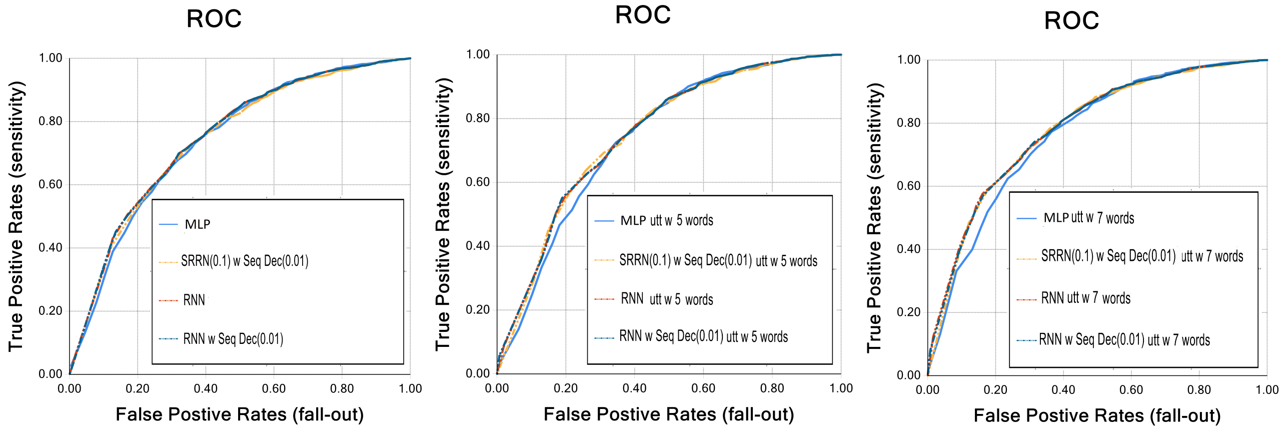


(a) ROC for 1^{st} Word

(b) ROC for 6^{th} Word

Fig. 3. ROC curves for individual words in an utterance

The presence of a recurrent layer in RNN's provides the model with additional insight in the past behavior of the recognizer. The estimator can then exploit this ability to provide a more refined estimate on the confidence. In contrast our current experimental MLP has no notion of context and so it treats every word in an utterance as an independent isolated event; this diminishes MLP's power to model and benefit from the context. It is however possible to add context to our current MLP setup, we can do so by combining the features from adjacent words. Unfortunately this method of modeling context make handling of utterances with different number of words difficult in our current infrastructure, we were therefore unable to cover those experiments in this paper. To measure and quantify impact of this context retention we generated ROC for a word at specific location in the utterance. These ROC curves are presented in Figure 3 and they highlight differences between the two confidence estimation systems. The ROC curves in the Figure 3(a) are gen-



(a) Three Word Utterance

(b) Five Word Utterance

(c) Seven Word Utterance

Fig. 4. ROC curves for different confidence estimators w.r.t. utterance length

erated by using only the labels and confidences of first word in an utterance. A MLP based confidence estimator is better at predicting the correct label for the first word of a utterance than any of the RNN based estimators as seen from the figures above. A recurrent layer in a RNN is the cornerstone of its performance but, for the first word in an utterance RNN based confidence estimator has not built sufficient history to make a sound decision and hence it suffers for short utterances. These observations are also reflected in the AUC's (measured for the first word ROC curve) for each system, AUC for RNN based estimator is 4% relatively worse than its MLP counterpart (Table 1). In the same vein Figure 3(b) shows the plot of ROC curves for the 6^{th} word in all the utterances from the test set. We can observe that the ROC for a RNN based estimator is significantly better than that of a MLP based estimator. In the same chart (Figure 3(a)) we can also observe that a sequentially trained RNN has a slight edge over a cross-entropy trained RNN. This observations from the figure are also reinforced when we look at the relative changes in the AUC presented in Table 1. The MLP base confidence estimator has an AUC of 0.776 and 0.794 for the first and sixth word respectively. From the observations and results presented in this section we can now state that for longer utterances, RNN based estimator is more accurate at assigning a word an appropriate confidence value than a MLP based estimator, and a sequentially trained RNN has the best in class AUC.

Table 1. *Relative changes in AUC of RNNs wrt MLP estimator.*

	Rel changes in AUC(%)	
Type	1^{st} word	6^{th} word
RNN	-4.12	8.37
RNN w Seq Decoding	-3.68	9.94
Seq-RNN	-3.39	10.61

In light of the above observations we expect that a RNN based estimator will be better at correctly estimating confidences for longer utterances, and the performance should be directly correlated with length of the utterances. The set of plots in Figure 4 show ROC curves of different estimators when input to the estimator is restricted to utterance with specific number of words. It is evident from Figure 3(a), that for a single word utterance traditional MLP based confidence estimators has the best in class performance, this behavior however start changing with increase in the length of the utterance. RNN based estimators start surpassing a MLP based systems for utterances with two or more words, and the gap widens with each additional word in the utterances. This trend in improvement of performance can be observed from figures 4(a) through 4(c).

The Table 2 lists relative improvement in true positive rate (TPR) for different confidence estimation systems at a constant 3% false positive rate. We can again observe that sequentially trained RNN do have slight edge over cross-entropy trained RNN's.

Table 2. *Relative TPR improvement for different confidence estimators w.r.t to MLP baseline at 3% FPR.*

Type of MLP	Rel TP improvement(%)
RNN	8.76
RNN w Seq Decoding	7.72
Seq-RNN	10.20

As we have observed throughout this section RNN's lag MLP when it comes to predicting correct confidence for the first word of an utterance, but the performance quickly and significantly improves with length of the utterance. This gain on longer utterances is significant enough to overcome the deficit caused over the first word and on an average RNN's have a much better performance than it MLP based counterpart (See Figure 2(b)).

4. CONCLUSIONS

In this paper we presented a RNN based confidence estimation system and compared it to a traditional MLP base system. We also presented two variations of the RNN system, one with sequential decoding and the second trained with sequential criterion. The ability of RNNs to model context is extremely useful for confidence estimation, and proves to be more effective for longer utterances. We also observed that sequential decoding and sequential training does yield small incremental improvement to the baseline RNN systems. For the proposed confidence estimator we also observed an approximately 10% reduction in false negative rates over the baseline MLP system. Since a MLP based estimator is better at short (one-word) word utterances than a RNN based estimator, we intend to investigate a topology that can combine the two systems to build a better confidence estimator.

5. REFERENCES

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

[2] Thomas Kemp and Thomas Schaaf, "Estimating confidence using word lattices," in *in Proceedings of EuroSpeech*, 1997, pp. 827–830.

[3] Frank Wessel, Klaus Macherey, and Ralf Schlter, "Schlter: using word probabilities as confidence measures," in *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1998, pp. 225–228.

[4] Frank Wessel, Klaus Macherey, and Hermann Ney, "A comparison of word graph and n-best list based confidence measures," in *in Proc. EUROSPEECH*, 1999, pp. 315–318.

[5] Frank Wessel, Ralf Schlter, Klaus Macherey, and Hermann Ney, "Ney: Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, pp. 288–298, 2001.

[6] Bernhard Rueber, "Obtaining confidence measures from sentence probabilities.," in *EUROSPEECH*, George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas, Eds. 1997, ISCA.

[7] Rafid A. Sukkar and Chin-Hui Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition.," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 6, pp. 420–429, 1996.

[8] R.C. Rose, B.-H. Juang, and C.H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, May 1995, vol. 1, pp. 281–284 vol.1.

[9] Mazin G. Rahim, Chin-Hui Lee, and Biing-Hwang Juang, "Discriminative utterance verification for connected digits recognition.," in *EUROSPEECH*. 1995, ISCA.

[10] R.A. Sukkar, "Rejection for connected digit recognition based on gpd segmental discrimination," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, Apr 1994, vol. i, pp. I/393–I/396 vol.1.

[11] L. Gillick, Y. Ito, and J. Young, "A probabilistic approach to confidence estimation and evaluation," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Apr 1997, vol. 2, pp. 879–882 vol.2.

[12] Man-Hung Siu and Herbert Gish, "Evaluation of word confidence for speech recognition systems.," *Computer Speech and Language*, vol. 13, no. 4, pp. 299–319, 1999.

[13] B. Chigier, "Rejection and keyword spotting algorithms for a directory assistance city name recognition application," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Mar 1992, vol. 2, pp. 93–96 vol.2.

[14] L. Mathan and L. Miclet, "Rejection of extraneous input in speech recognition applications, using multi-layer perceptrons and the trace of hmms," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference*, Washington, DC, USA, 1991, ICASSP '91, pp. 93–96, IEEE Computer Society.

[15] Mitch Weintraub, Franoise Beaufays, Ze'ev Rivlin, Yochai Konig, and Andreas Stolcke, "Neural-network based measures of confidence for word recognition," in *in Proc. ICASSP*, 1997, pp. 887–890.

[16] E. Eide, Herbert Gish, P. Jeanrenaud, and A. Mielke, "Understanding and improving speech recognition performance through the use of diagnostic tools," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, May 1995, vol. 1, pp. 221–224 vol.1.

[17] C.V. Neti, S. Roukos, and E. Eide, "Word-based confidence measures as a guide for stack search in speech recognition," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Apr 1997, vol. 2, pp. 883–886 vol.2.

[18] Pedro J. Moreno, Beth Logan, and Bhiksha Raj, "A boosting approach for confidence scoring.," in *INTERSPEECH*, Paul Dalsgaard, Brge Lindberg, Henrik Benner, and Zheng-Hua Tan, Eds. 2001, pp. 2109–2112, ISCA.

[19] Po-Sen Huang, Kshitiz Kumar, Chaojun Liu, Yifan Gong, and Li Deng, "Predicting speech recognition confidence using deep learning with word identity and score features.," in *ICASSP*. 2013, pp. 7413–7417, IEEE.

[20] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauch, C. Richey, E. Shriberg, K. Snmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *In Proceedings of the NIST Speech Transcription Workshop*, 2000.

[21] Matthew Stephen Seigel, Philip C Woodland, et al., "Combining information sources for confidence estimation with crf models.," 2011.

[22] Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, Patrick Gros, et al., "Crf-based combination of contextual features to improve a posteriori word-level confidence measures.," 2010.

[23] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, 2010, pp. 1045–1048.

[24] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu, "Recurrent neural networks for language understanding.," in *INTERSPEECH*, Frdric Bimbot, Christophe Cerisara, Ccile Fougeron, Guillaume Gravier, Lori Lamel, Franois Pellegrino, and Pascal Perrier, Eds. 2013, pp. 2524–2528, ISCA.

[25] Jeffrey L. Elman, "Finding structure in time," *COGNITIVE SCIENCE*, vol. 14, no. 2, pp. 179–211, 1990.

[26] Kaisheng Yao, Baolin Peng, Geoffrey Zweig, Dong Yu, Xiaolong Li, and Feng Gao, "Recurrent conditional random fields for language understanding," in *ICASSP*, 2014.

[27] Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang, "Asr error detection using recurrent neural network language model and complementary asr," in *Proceedings of ICASSP*, 2014.

[28] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[29] Dong Yu, Jinyu Li, and Li Deng, "Calibration of confidence measures in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, November 2011.