WORD CHARACTERS AND PHONE PRONUNCIATION EMBEDDING FOR ASR CONFIDENCE CLASSIFIER

Kshitiz Kumar, Tasos Anastasakos, Yifan Gong

Microsoft Corporation, Redmond, WA

{kshitiz.kumar, Tasos.Anastasakos, yifan.gong}@microsoft.com

ABSTRACT

Confidence classifier is an integral component of an automatic speech recognition (ASR) system. These classifiers predict the accuracy of an ASR hypothesis by associating a confidence score in [0,1] range, where larger score implies higher probability of the hypothesis being correct. Confidence scores have significant applications in ASR system design, training data selection, model adaptation, and other ASR applications. In this work we focus on word embedding features to improve confidence classifier, and introduce character and phone embeddings as confidence features. We motivate these features in the context of representing and factorizing acoustic scores along the proposed features. We evaluate our work on large scale ASR tasks, and demonstrate significant improvement in the confidence performance with the proposed features. At our typical operating point, we report 8% relative reduction in false alarm (FA) for limited vocabulary enUS Xbox task, and 9.9% relative reduction in FA for large vocabulary enUS server task. We also conducted server experiments for our proposed features in combination with natural language *Glove* embeddings, and improved the overall relative reduction in FA to 16%.

Index Terms— Confidence Classifier, Speech Recognition, Deep Learning, Word Embedding

1. INTRODUCTION

Confidence classifier is an integral component of an automatic speech recognition (ASR) system. Over past years, speech researchers have made significant advances in the ASR accuracy. That enabled the invention as well as large scale deployment of speech services catering many practical applications. The advances in deep learning [1, 2, 3, 4, 5] directly led to the creation of today's digital assistants in Cortana, Alexa, Google Home, and Siri. Although we made significant progress in ASR, we realize that the ASR hypotheses may have errors. In this context, confidence classifiers produce a confidence score in the range [0,1] for an ASR hypothesis, where higher score indicates larger probability of the hypothesis being correct.

ASR confidences have numerous applications. Confidences are key metrics that help speech applications better handle their responses to possibly incorrect ASR hypothesis. Confidence classifier is important for push-to-talk devices like cell phone but is especially critical for continuously listening devices like *Xbox*, where the speech engine is always running in background. Thus the ASR is listening to speech intended for it as well as unintended speech in side-speech, background noise, and other ambient sounds.

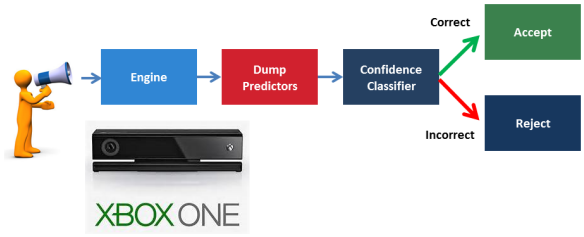

Fig. 1. *Confidence classifier for speech application*

There the ASR may produce in-grammar (IG) recognitions for unintended or out-of-grammar (OOG) utterances. ASR systems leverage confidence classifiers to detect incorrect recognitions and avoid a system response. Confidences have also have been applied to other speech applications [6, 7]. [8] applied confidences to guide the ASR decoding. Confidences have been applied for downstream ASR applications in arbitration and model adaptation in [9]. Calibration and normalization of confidence scores was done in [10, 11]. Confidences are also critical for data selection [12] and model combination with ROVER [13].

We typically use a multilayer perceptron (MLP) or a deep learning model [14, 15] to train the confidence scores from a defined set of features. Over the years a number of confidence features and training methods have been developed for confidences [16]. [17] conducts a broad survey of confidence techniques and applications. Confidences can be computed for words [18] as well as utterances [19]. Confidence features are computed from ASR lattices and N-best in respectively [20] and [21]. A maximum entropy method was proposed in [22], and a boosting method in [23].

Next we describe our confidence classifier framework with respect to Xbox application in Fig. 1. Xbox supports diverse applications like skype, games, command and query, switching menu etc., and provides high correct-accept (CA) at very low false-accepts (FA). There confidence classifier framework constitutes an ASR engine that decodes speech, and produces an hypothesis as well as a set of features for consumption by confidence classifier. Speech applications consume these confidence scores and make a decision on accepting the recognition events by comparing the score against a set threshold. The confidence scores help mitigate unwarranted Xbox response to background noise or TV sound etc.

Our current work focuses on developing new confidence features; we motivate and present word embedding features to improve confidence classifiers. Recently *Glove* [24] is a popular word embedding technique that has been applied to many natural language applications. Our work proposes word character and phone pronunciation embeddings to specially represent and factorize acous-

tic confidence features, and demonstrates significant improvements on large scale tasks. The rest of our work is organized as follows: we review our confidence classifier training and current confidence features in Sec. 2, we introduce word embedding features and related motivation in Sec. 3. We present our experiments and results in Sec. 4. Sec. 5 concludes this study.

2. REVIEW OF CONFIDENCE CLASSIFIER FEATURES AND TRAINING

We refer to [25] for broader introduction to our confidence classifier framework, features and training techniques. Confidence classification is essentially a binary classification problem [26] with the 2-classes in: (1) correct SR recognitions, (2) incorrect recognitions that includes misrecognitions over IG utterances as well recognitions from OOG utterances or background audio. The confidence features typically include:

1. acoustic-model scores
2. background-model scores
3. silence-model and noise-model scores
4. language-model scores
5. duration features

Our baseline confidence system consists of 21 features that are obtained from ASR lattices during decoding. We find acoustic scores to be the most important for confidence performance. We obtain confidence features from background, silence and noise model scores. We compute a set of language model (LM) features to represent LM score, perplexity and fanout. We also include duration-based features to model speaking rate and absolute speech duration. We normalize the features to be robust to speech with different duration and intensity.

3. CHARACTER AND PHONE EMBEDDING

In this work we develop new confidence features to improve the confidence performance. In current system, we obtain acoustic score for individual words in an ASR hypothesis as an aggregation of frame-level acoustic scores for the particular word. There stronger acoustic score indicates greater match of the constituent speech frames with the acoustic model, thus greater probability of the word being correct. ASR systems use context-dependent tied-triphones, *i.e.* senones, as states to represent the words. During decoding, we find the best path along the states under language model constraints, to predict the best hypothesis. Naturally the per-frame acoustic score represents a match between the speech frame and the particular acoustic state. Note that the baseline confidence features include duration that implicitly helps explain acoustic score from smaller vs. longer words. Additionally, we conduct a number of normalization of engine scores. Still the acoustic scores has a significant dependency upon the underlying acoustic states. Next, we motivate to better represent above dependency in terms of word embeddings.

3.1. Representing acoustic scores in terms of acoustic states

In Sec. 2 we noted that acoustic scores are typically the most important features for ASR confidence classifier. However, we also highlighted a dependency between the acoustic scores and underlying ASR states. We reasoned that a confidence classifier assigns higher confidence score to words with stronger acoustic scores but aforesaid dependency implies that the aggregated acoustic scores are insufficient at precisely representing the acoustic match without representing the underlying acoustic states. Assuming a large scale ASR task that consists of data across acoustic conditions, speakers, and audio pipeline, we will see considerable variation in acoustic scores for even correctly recognized words. We specifically represent the dependency between a few words and associated acoustic scores in Fig. 2. There we plot a distribution of the acoustic score for 3 words: "The", "Play", and, "Game". The distribution was obtained from words that were correctly recognized by ASR. Assuming rest of the confidence features to be identical, above differences in acoustic scores will impact confidence scores. We stated that the confidence scores indicate the probability of the word being correct, so different acoustic score distribution in Fig. 2 will lead to different interpretation for "The" and "Play" words for any given confidence score. Specifically the recognized word "The" at confidence, say 0.9, may have higher or lower probability of being correct than the word "Play" at confidence 0.9. From above understanding, we propose word embedding features to represent and rationalize acoustic score in acoustic states.

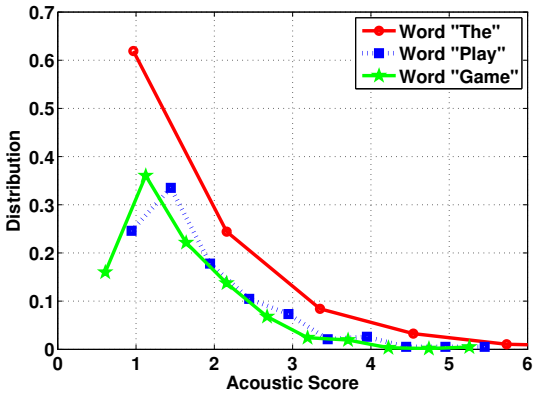

Fig. 2. *Acoustic score distribution for a few words. There lower score indicates stronger match.*

3.2. Word Character Embedding

Following the motivation in Sec. 3.1, we propose word character embedding to represent and factorize acoustic scores. The character embedding is simply a count of the alphabets in the language. For enUS, we build a 26-dimensional character embedding. Referring to Table. 1, the character embedding for "cortana" is a vector with $\{2, 1, 1, 1, 1, 1\}$ at respective locations for {'a', 'c', 'n', 'o', 'r', 't'}. The rest of the vector elements are 0.

Above character embedding offers several advantages: (a) these are smaller dimensional features, (b) they require almost no computing resource, (c) it's easily computed on the fly, and doesn't require any memory or storage. We show a flowchart for our proposed confidence work in Fig. 3. We build on the existing confidence mechanism, and extract baseline confidence features from ASR lattices [25]. The specific word is the only requirement for character embedding, so we embed a functionality in the lattice generation or lattice post-processing steps to compute character embedding for the words in ASR hypothesis. We realize that ASR systems essentially model phones, and the character embedding is at best a good approximation. Furthermore, "Cortana" pronounced in different ways will have identical character embeddings, despite different acoustic scores. Given that, we motivate and propose phone embedding in the next section.

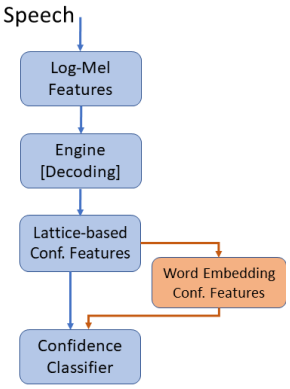

Fig. 3. *Lattice-based and Word embedding Confidence features.*

Table 1. *Word Characters and Phone Pronunciation embedding examples.*

Embedding Types	Embedding representation for the word "cortana"
Character embedding	c o r t a n a
Pronunciation embedding	k ao r t ae n ax k ao r t aa n aa

3.3. Phone Pronunciation Embedding

An ASR system is essentially a match between the speech frames and acoustic states under language model constraints. We use 9000 context-dependent triphones to represent acoustic states. We can choose to build a 9000-dimensional vector to represent a count of each of the triphones in a word but that's significantly larger than 21 in our baseline confidence features, and will likely overfit the task. It will also be difficult to train and maintain due to sparsity issues, as only a few of the states will be non-zero in a word. We therefore propose monophone units for word pronunciation embedding. We illustrate the phone embedding with the "cortana" example in Table. 1. Our enUS ASR model consists of 40 monophones, where we use a hand-crafted dictionary to represent the words in monophone units.

Phone embedding retains all the advantages of character embedding we noted in Sec. 3.2. There we also noted an issue with identical character embedding for different pronunciations of a word. The phone embeddings address that issues by allowing multiple pronunciations for words in a dictionary. The computation for phone embedding is similar to that for character embedding in Sec. 3, except that the embedding units are phones. We compute embedding for multiple pronunciations for a word as an average over the embedding from individual pronunciations. This computation simply requires the specific word and a monophone dictionary, that the ASR decoding already has access to.

4. EXPERIMENTS AND RESULTS

We evaluate our work on real recorded spontaneous speech. We present experiments on enUS limited vocabulary, *Xbox* task, as well as large vocabulary server task. The Xbox IG data consists of a variety of tasks in MarketPlace, Dashboard, and Take-Home. We prepare OOG utterances from movie or meeting tasks. We also simulate OOG data by decoding IG utterances against a mismatched grammar. The Server task consists of data from large scale applications in Mobile, Desktop, Bing search tasks. We measure the performance of confidence classifier in terms of mean squared error (MSE) on training and validation tasks, as well as in terms of $CA = \frac{\#AllCorrects\ beyond\ a\ threshold}{\#AllCorrects}$, and, $FA = \frac{\#AllIncorrects\ beyond\ a\ threshold}{\#AllIncorrects}$, there $\#$ indicates count. Our confidence training data consists of over 1000 hrs of speech for Xbox as well as Server task. We refer to [25] for details on our task and MLP-based confidence training.

Table 2. *Mean squared error (MSE) improvements from Character embeddings on Xbox task.*

Items	Confidence Features	Train MSE	Validation MSE
1	Character embedding	0.220	0.221
2	Acoustic Conf. features	0.218	0.216
3	(2) + (1)	0.199	0.199
4	All Conf. features	0.187	0.188
5	(4) + (1)	0.182	0.183

Table 3. *Mean squared error (MSE) improvements from Phone Pronunciation embeddings on Xbox task.*

Confidence Features	Train MSE	Validation MSE
Phone Pronunciation embedding (1)	0.211	0.213
Acoustic Conf. features + (1)	0.194	0.195
All Conf. features + (1)	0.174	0.175

We report MSE for our baseline and character embedding in Table. 2 for Xbox task. We noted that acoustic confidence features are typically the most significant for confidence performance. Table. 2 shows that the validation MSE for a confidence classifier trained from just the character embedding (0.221), is competitive with the classifier trained from purely the acoustic features (0.216). We also note that the combination of acoustic and character embedding improves the MSE to 0.199. Furthermore, integrating the embedding with all the baseline features improves the MSE from 0.188 to 0.183. We report similar results for phone embedding in Table. 3. The validation MSE for phone embedding is 0.213, and is stronger than 0.221 for character embedding in Table. 2. Furthermore, including phone embedding improves the baseline validation MSE from 0.188 to 0.175.

We also conducted an analysis on the overall confidence features, and report some higher ranking embedding features in Table. 4. Some of the baseline confidence features ranked higher than the embedding features but the ranking among the embedding features shows that the vowel sounds received higher importance than the consonant sounds. This result is very insightful and can lead to new embedding features for greater focus on vowels.

Next, we report confidence performance in CA and FA charts. Fig. 4 demonstrates the improvements from character and phone embedding for Xbox task. We see significant improvement in CA with the embedding features at all FA levels. At a typical operating point with CA=90%, we lower FA from 16.37% for baseline classifier to 16% with including character embedding, and 15.05% with phone

Table 4. *Higher ranked embedding features.*

Embeddings	Highly ranking features (in order)
Character embedding	u, o, i, e, a
Pronunciation embedding	eh, ey, iy, ay, ax

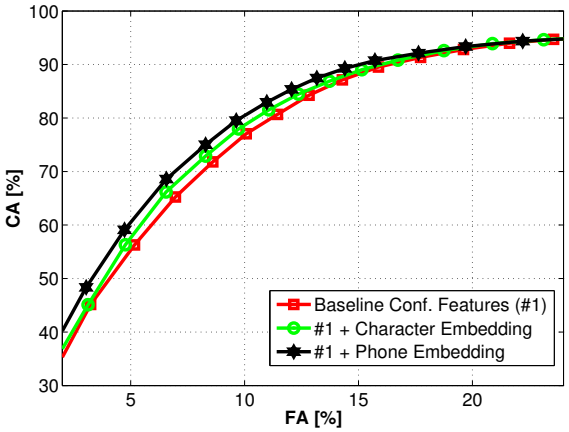

Fig. 4. *Confidence performance for enUS Xbox task.*

embedding. That leads to respectively 2.2% and 8% relative reduction in FA for character and phone embedding systems. Many of our systems use a variety of operating points, so the broad improvement from both the embedding techniques are very valuable.

We report confidence performance for enUS server tasks in Fig. 5 for the baseline and the embedding techniques we developed. As expected the FAs are even more challenging for large-vocabulary server tasks than Xbox tasks. At CA=75%, the character embedding lowers the FA for baseline from 8.25% to 7.53%. Phone embedding lowers FA to 7.43%, thus 9.9% relatively reduction in FA. At a broader level, the character and phone embeddings have similar confidence performance, with character emedding being better for lower FA targets.

We also conducted experiments with *Glove* [24] word embed-

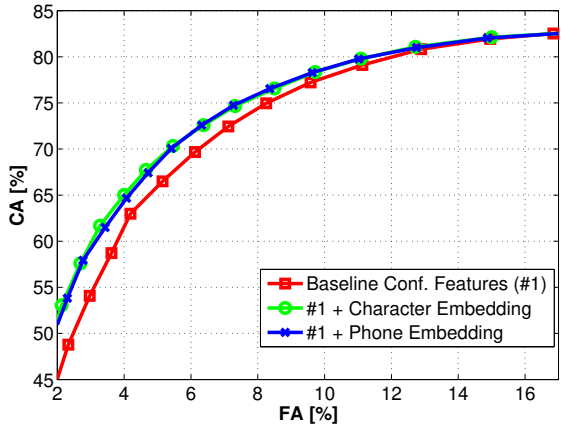

Fig. 5. *Confidence performance for enUS Server task.*

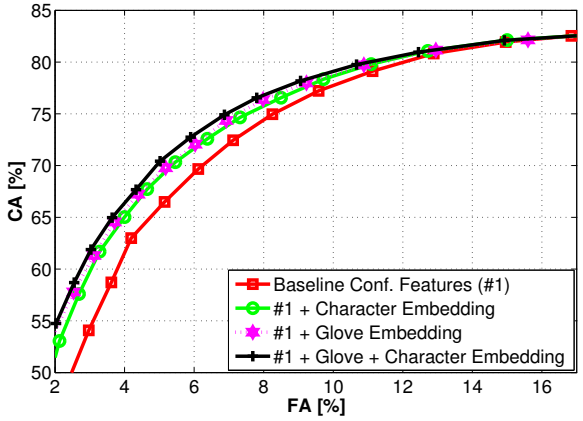

Fig. 6. *Confidence performance for enUS Server task in the context of Glove embedding.*

dings developed for natural language processing tasks. Glove embeddings encode contextual word information; that's distinct from the focus of the character and phone embeddings we developed, so we expect additional gains with combining these embeddings. Fig. 6 shows that the confidence performance for the Glove and character embeddings are competitive. We note that at CA=75%, the FA is respectively 8.25%, 7.53%, 7.26%, 6.92% for baseline, with character embedding, with Glove, and with character as well as Glove embedding. This leads to an overall 8.7% and 16.1% relative reduction in FA with just the character embedding, and character along with Glove embedding, respectively. Along with Glove, we also experimented with Facebook's *FastText* [27] embedding but found superior results with Glove, and used that in our study.

5. CONCLUSION

We develop new word embedding features and apply that to improve confidence classifier. We build our work from the observation that the acoustic scores are typically the most important features for confidences but they have a strong dependency upon the underlying acoustic states. We develop word embedding features to specifically factorize above dependency and provide basis for the confidence model to learn and improve the overall confidence performance. We propose word characters and phone pronunciations embeddings. Interestingly, we also found that the higher ranking embedding features corresponded to vowel sounds. We applied our work to limited vocabulary as well as large vocabulary tasks. At our confidence operating point, character embedding provided 2.2% and 8.7% relative reduction in FA for Xbox task and server task, respectively. On those tasks, the phone embedding showed 8% and 9.9% relative reduction. We also expanded our server experiments to use Glove embedding, and demonstrated an overall 16% relative reduction in FA with character embedding combined with Glove embedding.

6. REFERENCES

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

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoffrey Zweig, Xiaodong He, Jason D. Williams, Yifan Gong, and Alex Acero, "Recent advances in deep learning for speech research at microsoft," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8604–8608, 2013.

[4] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[5] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[6] 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.

[7] 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.

[8] 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.

[9] Kshitiz Kumar, Ziad Al Bawab, Yong Zhao, Chaojun Liu, Benoit Dumoulin, and Yifan Gong, "Confidence-features and confidence-scores for asr applications in arbitration and dnn speaker adaptation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] D. Yu, J. Li, and L. Deng, "Calibration of confidence measures in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2461–2473, Nov. 2011.

[11] Kshitiz Kumar, Chaojun Liu, and Yifan Gong, "Normalization of asr confidence classifier scores via confidence mapping," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] Yan Huang, Dong Yu, Yifan Gong, and Chaojun Liu, "Semi-supervised gmm and dnn acoustic model training with multi-system combination and confidence re-calibration.," in *Interspeech*, 2013, pp. 2360–2364.

[13] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[14] 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.

[15] 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.

[16] Gustavo Hernández-Abrego and José B Marino, "Contextual confidence measures for continuous speech recognition," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*. IEEE, 2000, vol. 3, pp. 1803–1806.

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

[18] 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.

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

[20] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proc. of EuroSpeech*, 1997, pp. 827–830.

[21] F. Wessel, K. Macherey, and H. Ney, "A comparison of word graph and n-best list based confidence measures," in *Proc. of EUROSPEECH*, 1999, pp. 315–318.

[22] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007, vol. 4, pp. 809–812.

[23] 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.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[25] 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," *IEEE ICASSP*, 2013.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., 2006.

[27] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.