CONFIDENCE MEASURES FOR CTC-BASED PHONE SYNCHRONOUS DECODING

Zhehuai Chen, Yimeng Zhuang and Kai Yu

Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Connectionist Temporal Classification (CTC) model has achieved state-of-the-art LVCSR performance. However, due to the introduction of the `blank` symbol, word-level confidence measures (CM) based on CTC model can not be easily calculated by directly using the traditional phone posterior normalization or confusion network (CN) approaches. Recently, a *phone synchronous decoding* (PSD) framework has been proposed for efficient decoding with CTC model. By automatically ignoring `blank` frames, PSD decoding not only achieves significant speed-up, but also yields highly compact and precise CTC phone lattices. In this work, two CM generation approaches on top of the PSD CTC lattice are proposed. Detailed investigation is also carried out to demonstrate the effectiveness of PSD CTC lattice. Experiments on an English switchboard LVCSR task showed that the performance of the proposed PSD CTC lattice based CM can significantly outperform the CM based on traditional frame synchronous decoding with CTC or HMM models.

Index Terms— CTC, Confidence measure, PSD, CTC lattice

1. INTRODUCTION

Automatic speech recognition (ASR) has achieved substantial successes in past few decades. However, when speech recognition systems are migrated from laboratory demonstrations to real-world applications, even the best ASR systems available today will inevitably make some mistakes during recognition [1], i.e., outputs from any ASR system are always fraught with a variety of errors. Thus, in any real-world application, it requires the ASR systems to automatically assess reliability or probability of correctness for every decision made by the system.

In speech recognition, *confidence measures* (CM) are used to evaluate reliability of recognition results [2]. Such kinds of methods can be divided into the following categories:

- *Predictor features based CM*. Feature from ASR decoding process can be called a predictor if its probabilistic distribution of correctly recognized words is clearly distinct from that of misrecognized words. CM can be derived from one or more of them, e.g., normalized acoustic score [3], duration [4], local entropy [5]. However, none of the above predictor features is ideal in distribution distinctness [2]. Therefore, models are proposed to combine these features together and predict an overall CM, e.g, CRF [6], Neural Network [7], etc. But these methods are still imperfect. Firstly it is because individual predictors are not statistically independent, and secondly it requires an additional training stage and assumes training data matches test data.
- *Posterior based CM*. Another kind of methods formulates ASR as the *maximum a posterior* (MAP) decision process. The posterior probability of ASR output given the whole utterance can be served as CM. Several methods are proposed to model the normalizing term of it, e.g., filler model [8], word lattice [9] and confusion network [10]. However, ASR decoders are commonly designed for finding the single best path, which results in imperfect word lattice and leads to unnormalized posteriors in CM [11].

Connectionist temporal classification (CTC) [12] has been proposed as a new type of acoustic model in ASR and achieved state-of-the-art performance[13][14][15][16]. Besides, context independent mono-phone CTC model also shows competitive performance compared with context dependent state clustered hybrid neural network HMM model[16][17][18][19]. However, due to the use of `blank` symbol, CM calculation becomes tricky. As shown in the experiments of the paper, simply treating `blank` symbol as a special phone and applying traditional CM approaches will yield very poor performance.

Recently, a *phone synchronous decoding* (PSD) framework has been proposed for fast decoding with CTC models. By automatically ignoring the `blank` frames during decoding and only carrying out Viterbi search at phone frames, PSD not only achieves significant decoding speed-up, but also generates extremely compact and high quality phone lattices[20][21]. In this paper, two CM calculation approaches, phone posterior average and confusion network, are proposed on top of the PSD CTC lattice. Compared to CM generated on top of the phone lattice from the traditional frame synchronous decoding (FSD) with both CTC and HMM models, the proposed approaches are more effective. The whole paper is arranged as follows. In section 2, PSD framework and CTC lattice are briefly reviewed. In section 3, two CMs are proposed as a pair of complements in CTC PSD framework. Section 4 describes experiments and analysis, followed by the conclusion in section 5.

2. PHONE SYNCHRONOUS DECODING

CTC model [12] predicts the conditional probability of the whole label sequence as (1)

$$P(\mathbf{l}|\mathbf{x}) = \sum_{\boldsymbol{\pi}\in\mathcal{B}^{-1}(\mathbf{l})} P(\boldsymbol{\pi}|\mathbf{x}) = \sum_{\pi:\pi\in L',\mathcal{B}(\boldsymbol{\pi}_{1:T})=\mathbf{l}} \prod_{t=1}^{T} y^t_{\pi_t} \quad (1)$$

where , $\mathbf{l}$ denotes a phone label sequence, $l \in L$ and L is the phone set for ASR. $\mathbf{x} = (x_1, \ldots, x_T)$ is the corresponding feature sequence, t is the index of frame and T is the total number of frames. $\boldsymbol{\pi}_{1:T} = (\pi_1, \ldots, \pi_T)$ is the frame-wise CTC output symbol *path* from frame 1 to T. Each output symbol $\pi \in L'$ and $L' = L \cup \{\texttt{blank}\}$. `blank` is a special symbol modelling the ambiguous variabilities outside the defined phone sets. y^t_k is the probability of output symbol of CTC network k at time t. A many-to-one

This work was supported by the Shanghai Sailing Program No. 16YF1405300, the China NSFC projects (No. 61573241 and No. 61603252) and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

mapping $\mathcal{B}$ is defined as $\mathcal{B} : L' \mapsto L$ to determine the correspondence between a set of *paths* and a phone label sequence.

In CTC model, the use of function $\mathcal{B}$ results in peaky and concentrated phone posteriors. In [21], by identifying each frame as `blank` or phone and ignoring all `blank` frames, search is carried out only at phone frames, leading to tremendous search redundancy removal. The special variable frame rate approach is referred to as *phone synchronous decoding* (PSD). In addition to decoding speed-up, PSD also benefits lattice generation. In traditional phone lattice generation framework [22][23], a postprocessing procedure is needed to combine phonemic arcs with similar time boundary. In contrast, due to the hard removal of `blank` frames, compared with the traditional *frame synchronous decoding* (FSD), less search errors and phone boundary disambiguity are made. This results in more compact, precise and boundary-clear phone lattice, referred to as *PSD CTC lattice*.

3. CONFIDENCE MEASURES FROM PSD CTC LATTICE

3.1. Phone Synchronous Phonemic Acoustic Confidence

CTC phone posteriors within the time span of a word can be accumulated to produce an estimation of the acoustic confidence measure (CM) of the word. In PSD framework, the word-level CM $\mathcal{C}(w)$ can be derived from the logarithmic posterior probability of the corresponding best decoding path.

$$\mathcal{C}(w) = \log \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{l}_w)} P(\boldsymbol{\pi}|\mathbf{x}) \triangleq \max_{\pi' : \pi' \in L, \mathcal{B}(\pi'_{\mathbf{j}_w}) = \mathbf{l}_w} \sum_{j : j \in \mathbf{j}_w} \log(y^{t_j}_{\pi'_j}) \quad (2)$$

$\mathbf{l}_w$ denotes the phone sequence corresponding to word w. j is the index of the *phone sequence* (i.e. non-`blank` CTC label sequence defined in [21]). As the word boundary is clear in PSD CTC, $\mathbf{j}_w$ can be defined as the set of phone index of word w in the best decoding path. Proposed phone synchronous phonemic acoustic confidence can be individually used as CM, or combined with other predictors to train model as in [7].

Empirically, the function of $\mathcal{B}$ in CTC is imperfect, which results in multiple continuous peaks in a single phonemic output. Therefore, it is reasonable to normalize on different numbers of continuous peaks. One method is to do arithmetic mean within each phonemic output (called *peak-mean*). However, because the multiple peaks come from imperfect modeling, a better method is to ignore the imperfect span and remain the best one. Therefore, choosing the maximum peak as phonemic output is another method (called *peak-max*). Besides, different words are of different lengths of phone sequence. To make CM between different words more competing and comparable, it is reasonable to normalize on the length of phone sequence (called *phone-mean*).

Another empirical issue is that there might be overlapped potion between `blank` and phonemic span (although the case is not common). It is then useful to introduce the confidence of non-`blank`. The phonemic frame confidence (called *phone-conf*) can be defined as the the probability of phone output in certain frame. Therefore, the combination of all the elaborate designment can be summarized as (3),

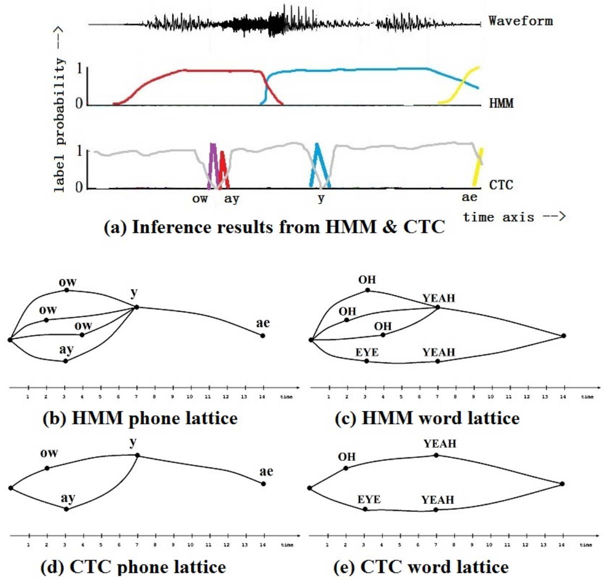
$$\mathcal{C}(w) \triangleq \max_{\pi' : \pi' \in L, \mathcal{B}(\pi'_{\mathbf{j}_w}) = \mathbf{l}_w} \frac{1}{|\mathbf{j}_w|} \sum_{j : j \in \mathbf{j}_w} \max_{t : t \in \mathbf{t}_j} \log(y^{t}_{\pi'_j}(1 - y^{t}_{\mathtt{blank}})^{\alpha}) \quad (3)$$

here *peak-max* is taken as an example (*peak-mean* and *peak-max* are two exchangeable setups). $\mathbf{t}_j$ is the set of frame index of phone j in the best decoding path. α is the weight of confidence interpolation.

3.2. Confusion Network from PSD CTC Lattice

Similar as [10], there are two steps to generate confusion network (CN): a) generate word lattice from phone level PSD CTC lattice as described in [21]; b) convert word lattice, in which word boundary time is included, to confusion network which is a pure word graph. In [24], the pivot clustering algorithm is proposed, which makes CN generation run in $O(n)$ time, n is the number of arcs in the lattice. In this work, the best path is used as pivot, and because of the compactness of CTC lattice, the CN generation is very efficient. During the construction of CN, word posterior is calculated and can be naturally used as word-level CM.

Figure 1 is a real example (an utterance ”OH YEAH”) showing the effectiveness of CN generated from PSD CTC lattice compared with CN from HMM-DNN based phone lattice. The inference results from HMM and CTC are plotted in Figure 1(a) and the resultant phone lattices and word lattices from HMM and CTC are plotted respectively in Figure 1(b∼e). It can be observed that both phone lattice and word lattice from CTC are more compact thanks to the use of function $\mathcal{B}$. In other word, lattice from HMM needs further heuristic many-to-one function to remove the redundancy, e.g., lattice pruning [23], which is not as efficient as the use of function $\mathcal{B}$ in CTC model. It is worth noting that when CTC model is used in the traditional *frame synchronous decoding* (FSD) framework, phone and word lattices can also be generated using similar approach[22] as in the HMM-DNN case by treating the `blank` symbol as a normal phone label. However, due to the search error from frame-level decoding and lattice pruning as well as ambiguous word boundaries, the resultant CN is of poorer quality. This will be investigated in detail in the experiment section.



(a) Inference results from HMM & CTC

(b) HMM phone lattice **(c) HMM word lattice**

(d) CTC phone lattice **(e) CTC word lattice**

Fig. 1. *Comparison of lattices from HMM & PSD with CTC*

4. EXPERIMENTS

A 300 hour English switchboard task was used to evaluate the proposed CM approaches. Both context dependent state level HMM (CD-state-HMM) and context independent mono-phone level CTC (CI-phone-CTC) models were trained. The training configuration and decoding setup were similar to [21]. All models were designed with around 2-2.5M parameters to get fair comparison. During testing, the switchboard subset from NIST Hub5e00 testset (1831 utterances) was used. With the CTC model, both frame synchronous decoding (FSD) and phone synchronous decoding (PSD) were both applied for comparison. Table 1 gives the performance of different models and decoding frameworks[1].

Table 1. *WER comparison*

Model Unit	AM	Decoding	WER
CD-state	DNN-HMM	FSD	16.7
CI-phone	LSTM-CTC	FSD	18.7
		PSD	18.8

4.1. Comparison of Confusion Network from FSD and PSD

This section demonstrates why confusion networks (CN) constructed from PSD CTC lattice is better than from FSD phone lattices. Phone lattice quality is analysed first, followed by discussions on word lattice and CN generation.

4.1.1. Phone Lattice Quality Analysis

As discussed in section 2, CTC model encodes the many-to-one function of $\mathcal{B}$ and results in peaky phonemic inference results. PSD phone lattice is generated simply by discarding significant amount of `blank` frames and constructing phone sausages on the remaining time-discontinuous frames. This procedure avoids search errors and ambiguous phone boundaries at `blank` frames and results in more compact and precise phone lattice compared to FSD decoding.

Oracle phone error rate (OPER) is used as the measurement to evaluate overall quality of phone lattice. It is calculated as the error rate of the best possible phone sequence existing in the lattice w.r.t. the reference phone sequence [25]. *Lattice density* (Arcs/Sec) is used as the metric of lattice compactness [26].

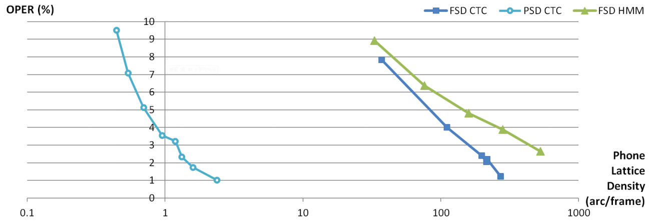


Fig. 2. *OPER v.s. lattice depth in FSD & PSD*

Figure 2 shows OPER versus lattice density from different decoding setup[2] of phone lattice generated from FSD and PSD. It can be observed that at similar OPER, the size of PSD lattice is more than 10 times smaller than FSD lattice generated from the same CTC model, and even more times smaller than that from HMM-DNN model. The reason of the distinct gap has two folds. Firstly, the peaky posterior property of CTC model naturally results in compact phone lattice. Secondly, PSD avoids search error at large amount of `blank` frames, while for FSD, lattice generation algorithms usually need to introduce heuristic approximation [27] and lattice pruning[22], which result in search errors. To sum up, with similar size, PSD CTC lattice contains more phone-level acoustic information. Since the focus of the section is to compare FSD with PSD, later experiments only compare CTC based FSD and PSD.

4.1.2. PSD CTC lattice Results in Better CN

To construct confusion network, word lattice needs to firstly be generated from phone lattice. Figure 3 shows lattice quality comparison between word lattices generated from the FSD and PSD CTC phone lattices. Here, *oracle word error rate* (OWER) is used as indicator of word lattice quality. In the figure, we plot the percentage of $1 - \frac{OWER}{WER}$ as the relative oracle word error rate reduction in the vertical axis to indicate the best possible performance improvement that the word lattice can bring about compared with the original 1-best WER. It can be clearly observed that with similar lattice density, PSD CTC word lattice mostly has better OWER. It can then be concluded that PSD based CTC phone lattice results in better word lattice quality, especially when lattice density is small.

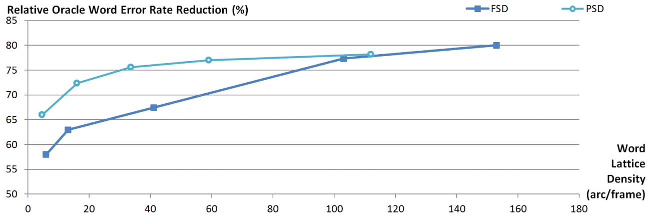


Fig. 3. *Lattice density v.s. relative oracle word error rate reduction*

Once word lattice is constructed, words with similar time boundaries are merged during CN generation using pivot based word clustering algorithm. Hence, time boundary accuracy of word lattice also affects the quality of the resultant CN. To analyse the time boundary quality of word lattice, *Nearest pivot boundary distance* is defined as $|b_{\texttt{arc}} - b_{\texttt{cluster}}| + |e_{\texttt{arc}} - e_{\texttt{cluster}}|$, where b_* and e_* are the begin and end of the word boundary and `arc` are the word arcs aligned to the best overlap pivot word `cluster`.

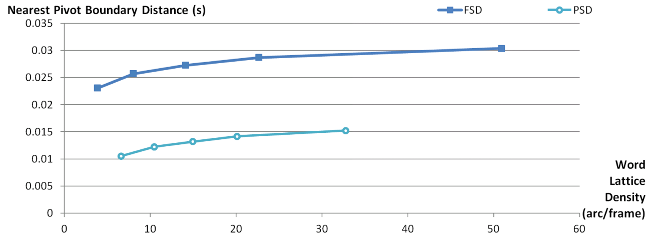


Fig. 4. *Boundary stability of PSD and FSD word lattice*

Figure 4 shows the average nearest pivot boundary distance of PSD and FSD word lattice with different sizes. It is revealed that nearest pivot boundary distance of PSD word lattice from PSD is distinctly smaller. In other word, the boundary of FSD based lattice is more unstable, which results in inefficient and inaccurate CN construction.

Finally the conversion efficiency from word lattice to CN was investigated. Figure 5 shows word lattice density versus confusion network depth (CN depth) after word clustering [24]. Result shows that with similar word lattice density, more competitors are generated from PSD lattice, which will bring about better OWER of CN

[1] Previous works [19][16][20] showed that with larger training dataset, CI-phone-CTC model can perform better than CD-state-HMM model. Since the focus of this paper is CM, confidence measures were still investigated using the switchboard task.

[2] The lattice density can be controlled by tuning CTC lattice threshold β in [21] and lattice beam pruning in [22].

and also more competing information of hypothesis. It is reasonable to assume that with more competing information of hypothesis, CN based confidence measure will get better performance.

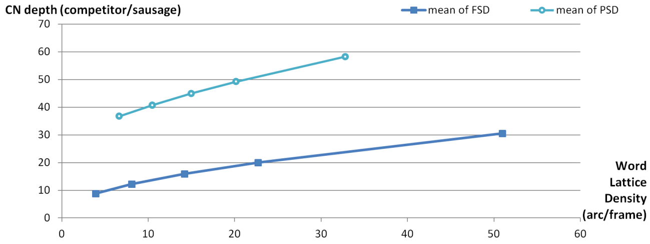


Fig. 5. *lattice density v.s. CN depth*

4.2. Confidence Measure Evaluation

In the section, word-level confidence measures (CM) are calculated using the two PSD CTC lattice based approaches described in section 3[3]. To evaluate the quality of word-level CM, *normalised cross entropy* (NCE) [28][29] is used here:

$$NCE = \frac{H(\mathbf{C}) - H(\mathbf{C}|\mathbf{x})}{H(\mathbf{C})} \tag{4}$$

where $H(\mathbf{C})$ corresponds to the entropy of the tag sequence, and $H(\mathbf{C}|\mathbf{x})$ is the entropy of the confidence score sequence. It's an information measure of how much additional information the tags can provide over the trivial baseline case of tagging all words with the average score. The higher NCE is the better.

4.2.1. PSD Phonemic Acoustic Confidence

Table 2 compares the proposed PSD phonemic acoustic confidence with the traditional phone posterior average approach. The traditional approach in [3] originally applied in HMM, was extended to CTC in FSD framework (denotated as *baseline*). The *peak-mean*, *peak-max*, *phone-mean*, *phone-conf* discussed in 3.1 are denotated as PN1, PN2, PN3 and PC respectively.

Table 2. *Comparison of phonemic acoustic confidence based CM*

AM	Decoding	CM	NCE
DNN-HMM	FSD	baseline	0.024
LSTM-CTC	FSD	baseline	0.058
	PSD	PN1 $\oplus$ PN3	0.105
		PN2 $\oplus$ PN3	0.135
		PN2 $\oplus$ PN3 $\oplus$ PC	0.141

Result shows that phonemic acoustic confidence proposed in [3] performs badly in word level. This is because the uncertainty of word boundary results in overlap of the acoustic score calculation, which consequently yields poor performance. When the approach is used with CTC model, the overlap problem can be alleviated because of peaky distribution characteristics of CTC inference result. But it is still problematic to decide whether allocate the probability mass of blank to the previous or the next phonemic label.

In PSD framework, the word boundary and blank allocation problem can both be solved, resulting in much better NCE. Besides, it is beneficial to use *peak-max* instead of *peak-mean*, which is parallel with analysis in 3.1. When the phonemic confidence is combined with *phone-conf* introduced previously, the performance becomes consistently better. Hence, in latter experiment, PN2 $\oplus$ PN3 $\oplus$ PC are used as the best phonemic acoustic confidence.

4.2.2. PSD Confusion Network based CM

Table 3 compares the confusion network based CM generated from the PSD CTC lattice and the FSD CTC or FSD HMM lattices.

Table 3. *Comparison of confusion network based CM*

AM	Decoding	CM	NCE
DNN-HMM	FSD	CN	0.172
LSTM-CTC	FSD	CN	0.019
	PSD	CN	0.224
		AC+CN	0.230

Result shows that although CN based confidence performs well in CD-state-HMM model (the result is consistent with previous works [10][9]), it can not be directly applied to CI-phone-CTC model. This is also due to the blank allocation problem. In contrast, CN based CM can be easily applied to PSD CTC phone lattice and achieve significantly better confidence score. Moreover, the NCE result is also significantly better than the CD-state-HMM system. We believe it is because the CTC lattice contains more competing information, as showed in Figure 5.

When the PSD phonemic acoustic confidence (PN2$\oplus$PN3$\oplus$PC in Table 2) and the CN based confidence are combined together (denotated as AC+CN), the performance can be further improved. It shows that the two types of CM are complementary as the previous one is a local CM while and the latter one is calculated using global utterance information.

5. CONCLUSION

In this paper, the potential of compact and precise PSD CTC lattice in preserving acoustic information was utilized to form better CMs. Phone synchronous phonemic acoustic confidence was proposed with elaborate phonemic normalization and blank information. Besides, the characteristics of lattice and confusion network generated from PSD framework were carefully investigated and confusion network based CTC lattice confidence was proposed. In experiments, both CMs achieve significantly better results compared to their competitors both in HMM and CTC. In addition, the two types of CMs can be combined together as a pair of complements. Future work includes applying proposed CMs as predictors in model training framework [6][7].

6. RELATION TO PRIOR WORK

Prior CM works all focus on ASR systems based on HMM within the frame synchronous decoding (FSD) framework. Approaches include predictor features based CMs [3] and confusion network based CMs [9][10]. However, for CTC model, due to the introduction of blank label and different acoustic model distribution[21], direct use of the previous approaches results in serious CM performance degradation. This work takes advantage of compact and precise CTC phone lattice generated from phone synchronous decoding (PSD) and proposes new predictor features based and confusion network based CMs suitable for CTC model. The proposed CMs achieve significantly better results compared to the traditional FSD CM approaches. Besides, the detailed analysis of word lattice and confusion network generated from PSD CTC lattice, will benefit further application based on it.

[3] To compensate for the effects of the lattice size and the resulting overestimation of the posteriors a decision tree was trained for each system to map the posterior probabilities to confidence scores [10]. Note that sentence level confidence scores can be calculated similarly and the conclusion does not change.

7. REFERENCES

[1] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James Landay, "Speech is 3x faster than typing for english and mandarin text entry on mobile devices," *arXiv preprint arXiv:1608.07323*, 2016.

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

[3] Wenping Hu, Yao Qian, and Frank K Soong, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call).," in *INTERSPEECH*, 2013, pp. 1886–1890.

[4] Zejun Ma, Xiaorui Wang, and Bo Xu, "Fusing multiple confidence measures for chinese spoken term detection," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] Daniele Falavigna, Roberto Gretter, and Giuseppe Riccardi, "Acoustic and word lattice based algorithms for confidence scores.," in *INTERSPEECH*, 2002.

[6] Mathew Stephen Seigel, *Confidence Estimation for Automatic Speech Recognition Hypotheses*, Ph.D. thesis, Ph. D. thesis, University of Cambridge, 2013.

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

[8] Sheryl R Young, "Detecting misrecognitions and out-of-vocabulary words," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. IEEE, 1994, vol. 2, pp. II–21.

[9] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 288–298, 2001.

[10] Gunnar Evermann and Philip C Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1655–1658.

[11] Peng Yu, Duo Zhang, and Frank Seide, "Maximum entropy based normalization of word posteriors for phonetic and lvcsr lattice search," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[13] Santiago Fernández, Alex Graves, and Jürgen Schmidhuber, "Phoneme recognition in timit with blstm-ctc," *arXiv preprint arXiv:0804.3269*, 2008.

[14] Tara Sainath, Kanishka Rao, et al., "Acoustic modelling with cd-ctc-smbr lstm rnns," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 604–609.

[15] Dario Amodei et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[16] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[17] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," *arXiv preprint arXiv:1507.08240*, 2015.

[18] Yajie Miao, Mohammad Gowayyed, et al., "An empirical exploration of ctc acoustic models," in *the 41th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016.

[19] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Hasim Sak, Alexander Gruenstein, Francoise Beaufays, et al., "Personalized speech recognition on mobile devices," *arXiv preprint arXiv:1603.03185*, 2016.

[20] Zhehuai Chen, Wei Deng, Tao Xu, and Kai Yu, "Phone synchronous decoding with ctc lattice," in *Interspeech 2016*, 2016, pp. 1923–1927.

[21] Z. Chen, Y. Zhuang, Y. Qian, and K. Yu, "Phone synchronous speech recognition with ctc lattices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 86–97, Jan 2017.

[22] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukáš Burget, Arnab Ghoshal, Miloš Janda, Martin Karafiát, Stefan Kombrink, Petr Motlíček, Yanmin Qian, et al., "Generating exact lattices in the wfst framework," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4213–4216.

[23] Sabato Marco Siniscalchi, Torbjorn Svendsen, and Chin-Hui Lee, "A bottom-up modular search approach to large vocabulary continuous speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 786–797, 2013.

[24] Dilek Hakkani-Tür, Frédéric Béchet, Giuseppe Riccardi, and Gokhan Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.

[25] Björn Hoffmeister, Tobias Klein, Ralf Schlüter, and Hermann Ney, "Frame based system combination and a comparison with weighted rover and cnc.," in *INTERSPEECH*. Citeseer, 2006.

[26] Philip C Woodland, Julian J Odell, Valtcho Valtchev, and Steve J Young, "Large vocabulary continuous speech recognition using htk," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*. Ieee, 1994, vol. 2, pp. II–125.

[27] Andrej Ljolje, Fernando Pereira, and Michael Riley, "Efficient general lattice generation and rescoring," in *Sixth European Conference on Speech Communication and Technology*, 1999.

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

[29] Jon Fiscus, "Sclite scoring package version 1.5," *US National Institute of Standard Technology (NIST), URL http://www. itl. nist. gov/iaui/894.01/tools*, 1998.