Sequence-level Confidence Classifier for ASR Utterance Accuracy and Application to Acoustic Models

Amber Afshan[1], *Kshitiz Kumar*[2], *Jian Wu*[2]

[1]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA
[2]Microsoft Corporation, Redmond, WA, USA

amberafshan@g.ucla.edu, {kshitiz.kumar, jianwu}@microsoft.com

Abstract

Scores from traditional confidence classifiers (CCs) in automatic speech recognition (ASR) systems lack universal interpretation and vary with updates to the underlying confidence or acoustic models (AMs). In this work, we build interpretable confidence scores with an objective to closely align with ASR accuracy. We propose a new sequence-level CC with a richer context providing CC scores highly correlated with ASR accuracy and scores stable across CC updates. Hence, expanding CC applications. Recently, AM customization has gained traction with the widespread use of unified models. Conventional adaptation strategies that customize AM expect well-matched data for the target domain with gold-standard transcriptions. We propose a cost-effective method of using CC scores to select an optimal adaptation data set, where we maximize ASR gains from minimal data. We study data in various confidence ranges and optimally choose data for AM adaptation with KL-Divergence regularization. On the Microsoft voice search task, data selection for supervised adaptation using the sequence-level confidence scores achieves word error rate reduction (WERR) of 8.5% for row-convolution LSTM (RC-LSTM) and 5.2% for latency-controlled bidirectional LSTM (LC-BLSTM). In the semi-supervised case, with ASR hypotheses as labels, our method provides WERR of 5.9% and 2.8% for RC-LSTM and LC-BLSTM, respectively.

Index Terms: Confidence Classifier, AM adaptation, AM customization, Data selection, LSTM

1. Introduction

Deep learning techniques have led to highly accurate ASR systems, resulting in their widespread use. It is therefore critical to ensure the reliability of ASR hypotheses. A confidence classifier is one such integral part of an ASR system, providing a measure of ASR reliability. But the current confidence scores lack a common meaning and potentially change with variation to confidence model or AM. Hence, we develop a sequence-level confidence model with an objective to closely tie the CC score for an utterance to its ASR accuracy. Thus, providing interpretation and uniformity in confidence scores across updates to confidence model or AM and expand CC applications. We study one such use, AM customization. We propose a cost-effective method using confidence scores for optimal data selection for AM customization in supervised and semi-supervised cases.

ASR confidence applications include (a) mitigating false alarms in ASR-enabled devices in always-listening mode, (b) data selection for AM training and adaptation, (c) arbitration to select the best hypotheses between client and server results, (d) key-word spotting tasks. CCs are commonly built with multi-layer perceptron (MLP) [1], recurrent neural networks (RNNs) [2], and bidirectional RNNs [3, 4]. These work at a word or a sub-word level [5]. Few methods use context [6], and alternative hypotheses to perform utterance verification [7]. Recent works focus on training confidence models at an utterance-level to classify correct/incorrect utterances with applications to distributed end-to-end speech recognition systems [8, 9].

Conventional CC scores lack absolute interpretation and reflect the underlying CC training data, AM, language model, task, etc. So the scores from two different CCs are not directly comparable. Thus, needing to shift the confidence thresholds across CC updates or confidence normalization [10] to recalibrate scores and retain those thresholds. While normalization preserves operating confidence thresholds across CC updates, it imparts no meaning to the scores. ASR developers or consumers expect CC scores to have a reasonable interpretation. We pursue that missing feature and introduce the ability to tie CC scores directly to ASR accuracy. In prior systems, a confidence score of say 0.9 lacked an exact meaning except for the expectation that corresponding ASR hypotheses could be more accurate than the ones with lower confidence, say 0.8. However, in the proposed CC, we expect a confidence score of 0.9 for an utterance to imply 90% accuracy.

There is an increase in popularity of ASR systems [11, 12] built with "unified data", consolidated from domains like dictation, voice search, command-and-control, call center, etc. The acoustic model trained on unified data–a "unified ASR"–simplifies AM training, where a single large-scale model includes multiple acoustic scenarios. It also facilitates better data sharing and delivers higher accuracy on unseen tasks. Given sufficient data for a specific domain, it is cost-effective to customize the unified model to the target domain rather than training a domain-dependent model for higher accuracy [13, 14].

Change in the target domain, speakers, tasks, accents, etc., is dealt with using AM adaptation [14]. Typically, it needs data well-matched to the target with supervised annotations but is impractical due to privacy and data constraints. Previously, a seed model provided target labels in a semi-supervised manner, filtering data using normalized frame-level entropy [15], and minimum Bayes risk [16]. Others [17, 18] used multiple potential transcriptions with a lattice or graph. We aim to design an efficient and scalable approach to customize AM. We propose to use confidence scores for supervised and semi-supervised AM adaptation. If supervised, we optimally select data for subsequent transcription and adaptation with a goal to minimize transcription costs and maximize ASR adaptation gains. In semi-supervised cases, we chose higher-quality ASR hypotheses.

In this work, we propose a sequence-level LSTM-based confidence model whose scores align to ASR accuracy and stable across updates, eliminating the need for confidence normalization. Such a CC enables us to perform data selection and has

scores tied to accuracy. We used CC scores to choose an optimal subset for AM adaptation developing a cost-effective method resulting in an improved WER. Besides, semi-supervised adaptation performance with optimal data is comparable to the supervised one with all the data. Section 2 presents the proposed sequence-level CC. Section 3 details our recommendations for optimal data selection to customize AM. The experimental setup are in Section 4 and the results and discussion in Section 5. We conclude with Section 6.

2. Sequence-Level Confidence Classifier to Predict ASR Accuracy

Confidence classifiers predict the correctness of the ASR hypotheses. ASR errors can include substitution, deletion, insertion, and potential hypotheses over non-speech sounds and background noise. CC training used a set of features described in [1], which typically includes acoustic model score, language model score, word duration, and count of phonemes.

2.1. MLP Confidence Classifier

Our baseline is a word-level MLP classifier. The high accuracy of ASR systems results in imbalanced correct and incorrect words with fewer negative samples, requiring data balancing across the two classes. We trained baseline with cross-entropy criterion providing word-level confidence scores. CC applications, including AM adaptation addressed in this work, require confidence scores at an utterance level. Since an unweighted aggregation has shown to be less effective than a weighted average of the word-level scores with corresponding word duration [5], we obtain utterance-level scores from a duration weighted average. Two main drawbacks of MLP CC are that every word is processed independently of the rest of the utterance, and the aggregation of the utterance-level information happens at the end, providing confidence scores with no interpretable meaning. A sequence model can explicitly process information at an utterance level and directly produce confidence scores per utterance.

2.2. LSTM Sequence Confidence Classifier

We noted that the traditional confidence scores lacked meaning, requiring operating threshold adjustments across CC or AM model updates. Confidence normalization can map the scores and retain thresholds, but it is still desirable to impart an interpretable meaning to scores. We seek a close parity between accuracy and CC scores. To better capture accuracy, we need to process data at an utterance level, motivating a sequence-level CC. It is well-known that ASR models (LSTM [19], TDNN [20]) benefit from sequence information. We expect such benefits to extend to the CC model. Hence, we propose a sequence-level confidence model using LSTM trained on word-level features, performing a regression task with the cumulative sum of errors as labels. At a given time instant, LSTM CC predicts a CC score on the hypothesis from the beginning of the utterance until that instant. It generalizes, and the final prediction of the utterance reflects the confidence for the entire utterance. The cumulative errors as target labels in our design resulted in a dependency on utterance length. To maintain uniformity, we apply length normalization to the cumulative errors. The training loss is,

$$L = \sum_{n=1}^{N} l_n, \quad l_n = (\hat{y}_n - y_n)^2 \qquad (1)$$

$$y_n = \frac{\text{\# correct words in hypothesis}}{n} \qquad (2)$$

where l_n, y_n, $\hat{y}_n$ are loss, accuracy (1-WER), and prediction, respectively at the nth word of N-word utterance. $\hat{y}_N$ is the utterance confidence score. Although the deleted words are not assigned CC scores, the accuracy calculations include deletion errors. We expect our optimization to provide utterance-level CC scores tightly correlated to accuracy, as LSTMs capture context and embed sequential information in the CC. While we chose LSTM as the confidence model, we expect this work to generalize to other sequence-level models like Transformers.

2.3. Context Information for Confidence Classifier

Context is another source of information for a better CC. It can learn and distill valuable knowledge from the inter-dependencies within an utterance (for example, co-articulation). We propose stacking the features across a window of a few past and future words to capture the context in CC models. This simple approach to including context could benefit all CC models.

3. Optimal Data Selection for Acoustic Model Customization

AM training at a commercial scale is an intensive process, and unified AM eliminates the need to train and deploy AM for individual scenarios. Customization is a cost-effective way to incorporate domain-dependent data in the unified model to obtain higher gains on specific domains. Model customization is especially valuable for large-scale customers who want to focus the unified model on particular application scenarios. Customization is more efficient than adding domain-specific data to the original training pipeline or developing a domain-dependent model from scratch. Customization can leverage complex adaptation approaches given the availability of an adequate amount of data. A scalable, lower-cost solution for AM customization is desirable that can potentially work with training data at different scales. We apply customization to two unified AM baselines: LC-BLSTM [21, 22] and RC-LSTM [23].

3.1. Acoustic Model Structures

The BLSTMs learn from both past and future information, providing gains over LSTMs but at the cost of increased complexity, model size, and latency. LC-BLSTM addresses some of these issues by restricting the forward and backward computation to batches of data with minimal loss in ASR accuracy. RC-LSTM uses a row convolution structure to distill and embed some future information into uni-directional LSTMs. It implicitly factorizes training in orthogonal time and "frequency" components leading to efficient training for large-scale models. RC models have smaller lookahead and are computationally efficient, but LC models access additional future information and are more complex with higher accuracy than RC-LSTM.

3.2. KL-Divergence Adaptation

The training data for model adaptation is often limited. We, therefore, build a conservative adaptation approach. We choose from regularized adaptation techniques: $L2$ regularized loss [24], KL-divergence (KLD) regularized loss [25], etc. We use KLD adaptation where KLD measures the distance between distributions of original and adapted models with parameters θ and θ', respectively. Given acoustic features $X = (x_1, x_2, \ldots, x_T)$ of length T the loss is,

$$\mathcal{L}_{KL} = \mathcal{D}_{KL}(f(x; \theta) || f(x; \theta')) \qquad (3)$$

Table 1: *Recommendation for selecting optimal data for AM adaptation based on the range of confidence scores. We select a percentage of data as shown in examples instead of setting an operational threshold.*

Confidence Range	Examples	Accuracy (Predicted)	AM Adaptation Recommendation		Comments
			Supervised	Semi-supervised	
High	Top 20%	High	-	✓[+]	Can be proxy for manual transcriptions.
Mid	Top 20-40%, Top 40-60%	Partially accurate	✓	✓	Can provide partial improvement.
Penultimate	Bottom 10-30%, Bottom 20-40%	Low	✓[+]	-	Can learn from manual transcriptions.
Very low	Bottom 0-10%	Lowest	-	-	Nuisances and unwanted recordings.

[+] The subset that provided the maximal gain with minimal data.

Table 2: *Statistics of the database used for the acoustic model.*

Database	Number of hours
en-US Training	64,000
Voice search Adaptation	35
Voice search Test	3

where $f()$ is the AM with parameters θ outputting labels $Y = (y_1, y_2, \ldots, y_T)$. We used the KL distance along with cross-entropy training criteria ($\mathcal{L}_{xent}$), resulting in an overall loss,

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{xent} + \lambda\mathcal{L}_{KL} \tag{4}$$

λ is the weight of the original model and $\hat{P}(Y|X)$ is the distribution of provided labels of the adaptation data y^{adapt}. The loss is equivalent to cross-entropy criteria with target distribution:

$$P(Y|X) = (1 - \lambda)\hat{P}(Y|X) + \lambda f(x; \theta) \tag{5}$$

3.3. Optimization of Data Used in AM Customization

Budget constraints and scalability needs bound any data collection and transcription for custom AMs. There are constant changes in data, systems, ASR engines, etc. Finally, not all data is equal in providing ASR gains. Hence, we develop a compact adaptation set by optimally selecting data for AM customization. We seek minimal data that yields maximum gain. In supervised adaptation, the primary cost is the transcription of the customization data. So in a supervised case, where transcriptions for the data are labels, we intend to optimally select a minimal set of data for subsequent transcription and adaptation. In semi-supervised adaptation, where ASR hypotheses are labels for the data, we choose ASR hypotheses with high accuracy to form the adaptation set. In both cases, confidence scores are crucial in selecting optimal data for cost-effective customization.

Our recommendations are in Table 1 for using confidence scores to select data in supervised and semi-supervised methods. We have a trade-off between the amount of data and expected data quality, so we consider confidences and bucket the data in percentile bins. We recommend high confidence data for semi-supervised adaptation as we expect unified ASR to be highly accurate in this case. However, we chose penultimate confidence range data for supervised adaptation as unified ASR is already accurate at high and mid confidence bins. The above follows that there is more scope to learn from penultimate confidence data. We exclude very-low confidence scores because that data likely has nuisances like highly overlapped speech, no speech, speech not targeted at the device, etc. The mid-range confidence scores can be appropriate for either approaches.

4. Experimental Setup

We used a large-vocabulary en-US data for unified AM training. We trained the confidence model on ∼60k utterances sampled from the en-US data, along with ∼7k utterances for the CC test set. A universal data set, the en-US data comprised of
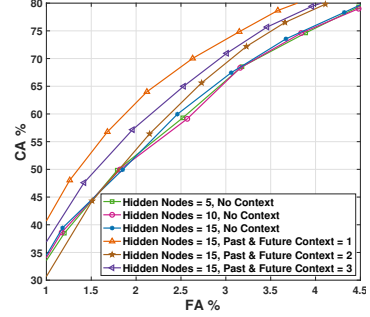

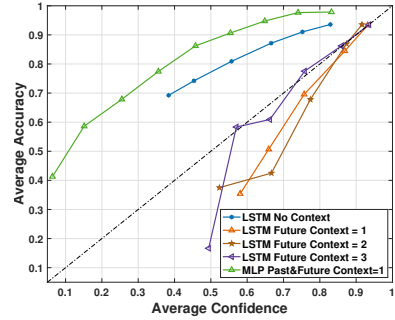Figure 1: *MLP CCs compared using CA-FA curves.*

Figure 2: *MLP compared with LSTM CC and context addition.*

data from tasks such as voice search, close-talk, natural conversation, command-and-control, dictation, call center, etc. We customized the unified model to a voice search task. This voice search task used for customization is from a different target application and is more recent, with queries from the latest trends than the en-US data. Table 2 has the statistics of the adaptation data. All the training and test data were anonymized, with all personal identifiable information removed. Our proposed CC uses a 4-layer unidirectional LSTM model with 512 cells with MSE training criteria. We removed one-word recordings from both training and testing lists of the LSTM. LSTM embeds and builds context, but the one-word utterances offer no context, and we found better convergence by excluding these utterances. We extracted 80-dim log-Mel features from 25 ms frames and a 10 ms frame-shift for ASR and used a 5-gram language model with a vocabulary of over 1M words. Adaptation was using KLD regularization with a KL weight, $\lambda = 0.5$.

5. Results and Discussion

5.1. Confidence Classifier

Figure 1 studies the number of hidden nodes and the inclusion of context to the baseline MLP CC. We measure performance as correct acceptance (CA) and false acceptance (FA) rates, where

we desire an operating range with high CA and low FA rates.

$$CA = \frac{\#\text{ Corrects beyond an threshold}}{\#\text{All corrects}}, FA = \frac{\#\text{Incorrects beyond a threshold}}{\#\text{All incorrects}}$$

In MLP CC, more neurons led to negligible performance gain, likely due to limited variability to encapsulate at word-level. But adding past and future contexts improves performance. We obtain the best performance with the context of one word.

CA-FA curve, commonly used with word-level CCs, is an inadequate measure of performance for a CC model trained at utterance-level in a regression setting. Besides, we aim to align confidences with ASR accuracy (1-WER), so we devise a new metric for the LSTM CCs in Figure 2. We preserve the core idea in CA-FA curves and divide the confidences into ten bins at 0.1 intervals. We produce average confidence and associated accuracy of utterances in those bins. Note that accuracy calculations include deleted words but they were not assigned confidence. We seek a high correlation between ASR accuracy and confidence, i.e., curves closer to the line of equality.

Figure 2 shows performance comparison for MLP and LSTM CCs. The confidence versus accuracy plot is closer to the line of equality for LSTM than MLP CC and concludes that LSTM CCs better predicts ASR accuracy. Confidence scores are better correlated to accuracy in higher confidence ranges demonstrating the effectiveness of sequential modeling of confidence scores. We show the benefit of adding future context in LSTM CCs with a few irregularities in the low confidence regions. We believe that is due to shorter utterances with limited future context that may dominate the low confidence regions. Additional future context may help LSTM CC for longer utterances but will have challenges due to limited training data and may regress on shorter utterances. We found a good trade-off with a future context of one word.

Earlier analysis used the en-US train and test data. We examine the performance of data mismatch by evaluating the voice search task in Table 3 for CCs trained with en-US. The results are in terms of correlation between the bin accuracies and confidence scores. As hypothesized, the proposed LSTM CC suffers little degradation compared to baseline MLP CC with data-mismatch between train and test. Thus, implying confidence normalization is unnecessary for the LSTM CC, making it more suitable for AM customization to different domains.

5.2. Customization of Acoustic Models

We customize unified AM to a voice search task and report WER and WER reduction (WERR) over baseline unified AM in Table 4. Model structures used are RC-LSTM and LC-BLSTM as described in [23]. The LC-BLSTM unified AM, as expected, performs better than RC-LSTM.

Supervised adaptation with all the adaptation data (35 hours) shows 6.39% WERR for RC and 3.02% WERR for LC model. A relatively higher gain for the RC model shows that simple model structures can be more effective for limited data adaptation. These results provide a reference for subse-

Table 3: *Comparison of the correlations (average bin accuracy and confidence scores) for tests on en-US and voice search data.*

Test data	CC (train: en-US)	Correlation
en-US	LSTM	0.99
(matched)	MLP	0.95
Voice search	LSTM	0.97
(mis-matched)	MLP	0.78

Table 4: *Word error rate (WER) and WER reduction (WERR) over the baseline unified model for customization to a voice search task using different optimal subsets (see Table 1).*

	Range	RC-LSTM		LC-BLSTM	
		WER%	WERR%	WER%	WERR%
Baseline		13.61	-	13.25	-
Sup	All	12.74	6.39	12.85	3.02
	Top 20%	12.72	6.54	12.88	2.79
	Bottom 10-30%	**12.46**	**8.45**	**12.56**	**5.21**
Semi-sup	All	13.28	2.42	13.02	1.74
	Top 20%	**12.81**	**5.88**	**12.88**	**2.79**
	Bottom 10-30%	12.99	4.56	13.03	1.66

quent AM customization following our recommendations in Table 1 of optimal data selection with confidence scores. As hypothesized, supervised adaptation with penultimate confidence range data (7 hours) performs the best with WERRs of 8.45% and 5.21% for RC and LC models, respectively. Thus, saving transcription costs significantly by transcribing data only from the penultimate range. Semi-supervised adaptation performed best using a higher confidence range data, the top 20% of the data (7 hours), resulting in WERR of 5.88% and 2.79% for RC and LC models, respectively. CC can hence effectively select data with reliable hypotheses for semi-supervised adaptation. Finally, our optimal data selection is better than all data in both cases. Our reported evaluations are on large-scale tests, and we expect any WER beyond 1% relative to be significant. Note that we do not report the mid and very-low range as they are not optimal and result in very low WERR. As an extension, we also combined supervised (ground-truth labels; bottom 10-30%) and semi-supervised (ASR hypothesis as labels; top 20%), a total of 14 hours of data to customize the LC model. We obtained a WERR of 4.08%, not outperforming the corresponding supervised-only adaptation. We also conducted similar experiments with confidences from MLP CC, resulting in some improvement over the baseline. For instance, the top 20% data selected based on MLP CC scores resulted in a WERR of 5% in a supervised adaptation.

6. Conclusion

We proposed a sequence-level LSTM-based confidence model and captured a rich context, resulting in confidence scores that closely followed accuracy–more prominently in high accuracy brackets. Traditionally, confidence scores lack universal interpretation that changes across CC updates. The proposed CC introduced scores that depict the accuracy and eliminate the need for confidence normalization across CC updates, making them suitable for many tasks. The CC, therefore, enabled the use of confidences for adaptation data selection by providing stable utterance-level scores that predict accuracy. Consequently, we developed an efficient, cost-effective, and scalable approach for AM customization using CC scores to select optimal adaptation data. This work is the first evidence that compact data set selected by a CC is better than using all data–both in supervised and semi-supervised AM customization. In the proposed approach, we chose an optimal subset of the data to perform transcription and reduce cost while obtaining maximal gain. Likewise, we chose a subset of optimal data with high-quality ASR hypotheses to act as labels. Semi-supervised adaptation with optimal data resulted in WERR comparable to the supervised one with all the data, proving the efficacy of the proposed method. Given the compact nature of the RC model, it had more benefits from customization when compared to the LC model.

7. References

[1] P. S. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng, "Predicting speech recognition confidence using deep learning with word identity and score features," in *ICASSP*, 2013, pp. 7413–7417.

[2] K. Kalgaonkar, C. Liu, Y. Gong, and K. Yao, "Estimating confidence scores on ASR results using recurrent neural networks," in *ICASSP*, 2015, pp. 4999–5003.

[3] M. A. Del-Agua, A. Gimenez, A. Sanchis, J. Civera, and A. Juan, "Speaker-adapted confidence measures for ASR using deep bidirectional recurrent neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 7, pp. 1194–1202, 7 2018.

[4] A. Ragni, Q. Li, M. J. Gales, and Y. Wang, "Confidence Estimation and Deletion Prediction Using Bidirectional Recurrent Neural Networks," *SLT*, pp. 204–211, 10 2019. [Online]. Available: http://arxiv.org/abs/1810.13025

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

[6] A. Sarma and D. D. Palmer, "Context-based speech recognition error detection and correction," in *Proceedings of HLT-NAACL: Short Papers*, 2004, pp. 85–88.

[7] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," pp. 602–605, 1996. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/607433

[8] A. Kumar, S. Singh, D. Gowda, A. Garg, S. Singh, and C. Kim, "Utterance confidence measure for end-to-end speech recognition with applications to distributed speech recognition scenarios," in *INTERSPEECH*, vol. 2020-Octob, 2020, pp. 4357–4361.

[9] Q. Li, Y. Zhang, B. Li, L. Cao, and P. C. Woodland, "Residual Energy-Based Models for End-to-End Speech Recognition," *arXiv preprint arXiv:2103.14152*, 2021. [Online]. Available: http://arxiv.org/abs/2103.14152

[10] K. Kumar, C. Liu, and Y. Gong, "Normalization of ASR confidence classifier scores via confidence mapping," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014.

[11] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified Architecture for Multichannel End-to-End Speech Recognition with Neural Beamforming," *IEEE Journal on Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017. [Online]. Available: http://www.merl.com

[12] Y. Fujita, R. Takashima, T. Homma, R. Ikeshita, Y. Kawaguchi, T. Sumiyoshi, T. Endo, and M. Togami, "Unified ASR system using LGM-based source separation, noise-robust feature extraction, and word hypothesis selection," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2 2016, pp. 416–422.

[13] J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, S. Parthasarathy, V. Mazalov, Z. Wang, L. He, S. Zhao, and Y. Gong, "Developing RNN-T Models Surpassing High-Performance Hybrid Models with Customization Capability," *Interspeech*, 2020. [Online]. Available: http://arxiv.org/abs/2007.15188

[14] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation Algorithms for Speech Recognition: An Overview," *arXiv preprint arXiv:2008.06580*, 8 2020. [Online]. Available: http://arxiv.org/abs/2008.06580

[15] S. H. Liu, F. H. Chu, S. H. Lin, and B. Chen, "Investigating data selection for minimum phone error training of acoustic models," in *Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, ICME 2007*. IEEE Computer Society, 2007, pp. 348–351.

[16] S. Walker, M. Pedersen, I. Orife, and J. Flaks, "Semi-Supervised Model Training for Unbounded Conversational Speech Recognition," *arXiv preprint arXiv:1705.09724*, 2017. [Online]. Available: http://arxiv.org/abs/1705.09724

[17] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-Supervised Training of Acoustic Models Using Lattice-Free MMI," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, 2018, pp. 4844–4848.

[18] O. Klejch, J. Fainberg, P. Bell, and S. Renals, "Lattice-Based Unsupervised Test-Time Adaptation of Neural Network Acoustic Models," *arXiv preprint arXiv:1906.11521*, 2019. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/material http://arxiv.org/abs/1906.11521

[19] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014.

[20] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-595

[21] S. Xue and Z. Yan, "Improving latency-controlled BLSTM acoustic models for online speech recognition," in *ICASSP*. Institute of Electrical and Electronics Engineers Inc., 6 2017, pp. 5340–5344.

[22] K. Chen and Q. Huo, "Training Deep Bidirectional LSTM Acoustic Model for LVCSR by a Context-Sensitive-Chunk BPTT Approach," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 7, pp. 1185–1193, 7 2016.

[23] K. Kumar, C. Liu, Y. Gong, and J. Wu, "1-D Row-Convolution Lstm: Fast Streaming Asr At Accuracy Parity With Lc-Blstm," in *Interspeech*, 2020.

[24] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *ICASSP*, 2013, pp. 7947–7951.

[25] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 7893–7897.