Combination of end-to-end and hybrid models for speech recognition

Jeremy H. M. Wong, Yashesh Gaur, Rui Zhao, Liang Lu, Eric Sun, Jinyu Li, and Yifan Gong

Microsoft Speech and Language Group

{jewong,yagaur,ruzhao,lial,ersun,jinyli,ygong}@microsoft.com

Abstract

Recent studies suggest that it may now be possible to construct end-to-end Neural Network (NN) models that perform on-par with, or even outperform, hybrid models in speech recognition. These models differ in their designs, and as such, may exhibit diverse and complementary error patterns. A combination between the predictions of these models may therefore yield significant gains. This paper studies the feasibility of performing hypothesis-level combination between hybrid and end-to-end NN models. The end-to-end NN models often exhibit a bias in their posteriors toward short hypotheses, and this may adversely affect Minimum Bayes' Risk (MBR) combination methods. MBR training and length normalisation can be used to reduce this bias. Models are trained on Microsoft's 75 thousand hours of anonymised data and evaluated on test sets with 1.8 million words. The results show that significant gains can be obtained by combining the hypotheses of hybrid and end-to-end NN models together.

Index Terms: Combination, end-to-end, hybrid, minimum Bayes' risk, speech recognition

1. Introduction

The hybrid Neural Network (NN)-Hidden Markov Model (HMM) architecture [1] has, up till recently, maintained its reputation as being the architecture of choice for state-of-the-art performance in Automatic Speech Recognition (ASR) [2, 3, 4, 5]. However, with the ever increasing quantity of diverse training data and the development of novel NN topologies, it may now be possible to construct ASR models based on end-to-end NN architectures [6, 7] that perform on-par with, or even outperform, hybrid models [8].

Two common end-to-end NN architectures are the Listen, Attend, and Spell (LAS) [6, 9] and the Recurrent Neural Network Transducer (RNN-T) [7, 10, 11]. There are several differences between hybrid and end-to-end NN models, such as in the conditional independence assumptions that they make, and in the availability of an externally defined pronunciation lexicon and language model. Between end-to-end NN models, one major difference between LAS and RNN-T is that RNN-T is frame-synchronous, as it produces a *blank* output for every input frame, while LAS is not frame-synchronous. These differences may result in significant diversity between the error patterns that the models make when performing recognition. A combination between the models may be able to take advantage of any complementary diversity.

Hypothesis-level combination methods, such as Recogniser Output Voting Error Reduction (ROVER) [12] and Minimum Bayes' Risk (MBR)-based combination methods [13, 14, 15] were originally proposed in the context of a combination between HMM-based models. However, many of these approaches are also usable with end-to-end NN models. This paper investigates the use of these hypothesis-level combination methods, to combine the predictions of hybrid and end-to-end NN models together. To the best knowledge of the authors, this is the first report on a combination between hybrid and end-to-end NN models. MBR combination relies on the accuracy of each model's hypothesis posteriors. However, the posteriors from end-to-end NN models tend to exhibit a bias toward shorter hypotheses [7, 6]. This paper investigates whether length normalisation and MBR training can alleviate this bias, and thereby improve the compatibility between end-to-end NN models and MBR combination.

There have been prior works investigating combination between multiple HMM-based models [12, 13, 14, 15], and also between multiple end-to-end NN models [16, 17, 18]. The authors in [16, 17] propose to combine the RNN-T and LAS together, by using a two-pass decoding strategy with a two-headed encoder-decoder style architecture. A shared encoder is used for both the RNN-T and LAS, while each of the RNN-T and LAS has its own separate decoder. The RNN-T decoder is used to perform a first-pass decoding in streaming mode, then the LAS decoder is used in offline mode to either re-score an N-best list of hypotheses from the RNN-T or to perform a fresh recognition pass, when given information about the RNN-T's hypotheses. The re-scoring or second-pass recognition can be interpreted as a combination between the RNN-T and LAS decoders. In [18], time alignments from a first pass connectionist temporal classification decoding run of the shared encoder are used to guide an LAS decoder. However, since these methods use a shared encoder, the diversity between the model behaviours may be limited. In this paper, hypothesis-level combination is performed between completely separate models, in the hope of taking full advantage of the diversity between their behaviours.

2. Models

2.1. Hybrid

The posterior probability of a word sequence, $\boldsymbol{\omega}_{1:L}$, for a hybrid model is computed as

$$P\left(\boldsymbol{\omega}_{1:L}|\mathbf{O}_{1:T}\right) \propto P^{\gamma}(\boldsymbol{\omega}_{1:L}) \sum_{\mathbf{s}_{1:T} \in \boldsymbol{\omega}_{1:L}} \prod_{t=1}^{T} \frac{P\left(s_t|\mathbf{o}_t\right)}{P\left(s_t\right)} P\left(s_t|s_{t-1}\right), \tag{1}$$

where L is the number of words, $\mathbf{O}_{1:T}$ are the input features, T is the number of frames, γ is the language scaling factor, and $\mathbf{s}_{1:T}$ are the HMM states. The HMM imposes two conditional independence assumptions; that the probability of the current state is independent of all observations and other states when given the previous state, and that the probability of the current observation is independent of all other observations and states when given the current state. The language model is often simplified to an n-gram model,

$$P\left(\boldsymbol{\omega}_{1:L}\right) = \prod_{l=1}^{L} P\left(\omega_l|\boldsymbol{\omega}_{l-n+1:l-1}\right), \tag{2}$$

which imposes the assumption that the probability of the current token is only dependent on a finite context of past tokens. It is also often trained separately from the NN acoustic model, $P(s_t|\mathbf{o}_t)$. The hybrid model incorporates a lexicon, which defines the mapping between token sequences, $\boldsymbol{\omega}_{1:L}$, and state sequences, $\mathbf{s}_{1:T}$. This lexicon is often manually defined, such that the states are related to either phonemes or graphemes.

2.2. LAS

The LAS is an encoder-decoder NN model, with an attention layer connecting the encoder with the decoder. The posterior probability of a token sequence, $\boldsymbol{\tau}_{1:J}$, for LAS is computed as

$$P(\boldsymbol{\tau}_{1:J}|\mathbf{O}_{1:T}) = \prod_{j=1}^{J} P(\tau_j|\boldsymbol{\tau}_{1:j-1}, \mathbf{O}_{1:T}), \quad (3)$$

where J is the number of tokens. Here, there are no conditional independence assumptions enforced upon the model. The LAS model does not utilise information from any manually-defined lexicon. The language and acoustic components of the LAS model are jointly trained. The LAS decoder produces one output for each token position, j.

2.3. RNN-T

Similarly to LAS, the RNN-T can also be viewed as an encoder-decoder NN model. The token sequence posterior probability for RNN-T is computed as

$$P(\boldsymbol{\tau}_{1:J}|\mathbf{O}_{1:T}) = \sum_{\mathbf{s}_{1:T+J} \in \mathbb{B}(\boldsymbol{\tau}_{1:J}, T)} \prod_{k=1}^{T+J} P(s_k|\mathbf{s}_{1:k-1}, \mathbf{O}_{1:T}). \quad (4)$$

The set of states, s, is the union of *blank* with the set of tokens, τ, and $\mathbb{B}(\boldsymbol{\tau}_{1:J}, T)$ represents the set of possible state sequences, $\mathbf{s}_{1:T+J}$, that have T *blank*s interpolated within $\boldsymbol{\tau}_{1:J}$. Similarly to LAS, the RNN-T also does not impose any conditional independence assumptions, and the language and acoustic components are jointly trained. However, unlike LAS, the RNN-T produces one output for each input frame and output token, and can therefore be considered as a frame-synchronous model. The input frame is only incremented each time a *blank* is produced at the output.

During recognition, the token sequence posteriors for both LAS and RNN-T are mapped to word sequence posteriors as

$$P(\boldsymbol{\omega}|\mathbf{O}_{1:T}) \approx \max_{\boldsymbol{\tau} \in \boldsymbol{\omega}} P(\boldsymbol{\tau}|\mathbf{O}_{1:T}), \quad (5)$$

The maximisation is only performed over token sequences explored within the beam search. A maximisation is used instead of a summation, because the sparsity of the token sequences explored within the beam search may result in an unintended bias toward word sequences that happen to have a larger number of token sequences being explored, if using summation. This maximisation is similar to the determinisation in a Viterbi semi-ring that is often applied to the recognition lattice of hybrid models.

3. Combination

The different model architectures may yield diverse error patterns. Through combination, the correct predictions of one model may be able to correct the wrong predictions of another model. One common combination method in ASR is ROVER [12]. This performs majority voting between the 1-best hypotheses from each model. Confidence scores can also be used.

It may be possible to use additional information by considering multiple hypotheses from each model, rather than only the 1-best, by using MBR-based combination methods [13, 14, 15]. Generalising ROVER to operate over N-best lists can be viewed as an approximation of MBR combination [14, 15]. In MBR combination, the combined hypothesis is computed as

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}'} \sum_{m=1}^{M} \lambda_m \sum_{\boldsymbol{\omega} \in \mathbb{N}} \mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\omega}') \frac{P_m^{\kappa_m}(\boldsymbol{\omega}|\mathbf{O}_{1:T})}{\sum_{\boldsymbol{\omega}' \in \mathbb{N}} P_m^{\kappa_m}(\boldsymbol{\omega}'|\mathbf{O}_{1:T})}, \quad (6)$$

where λ_m weighs the contributions between the models, κ_m is a scaling factor that balances the dynamic ranges between the models, and $\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\omega}')$ is the minimum edit distance between two word sequences, $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$. Here, $\mathbb{N}$ is the union of the set of hypotheses from the M models. If the set of hypotheses from each model is represented as an N-best list, then the support of the N-best lists from the separate models may be different. In this case, it is assumed that the posterior for hypothesis $\boldsymbol{\omega}$ from model m is $P_m(\boldsymbol{\omega}|\mathbf{O}_{1:T}) = 0$, if $\boldsymbol{\omega}$ is not contained within the N-best list that was generated by model m.

The minimum edit distance computation in $\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\omega}')$ is non-local, making it difficult to perform efficient computation using forward-backward operations over the set of hypotheses. Computing $\mathcal{L}(\boldsymbol{\omega}, \boldsymbol{\omega}')$ separately for each hypothesis can be computationally expensive. Furthermore, the set of hypotheses to minimise over is potentially infinite. As such, various approximations [13, 14, 15] have been proposed to perform MBR decoding. The works in [14, 15] try to first find an approximate alignment between the hypotheses, then selects the word with the highest score from each aligned confusion set. The work in [13] instead seeks to minimise an upper bound to (6), that can be computed efficiently using forward-backward operations.

4. Length normalisation

The effectiveness of MBR combination in (6) is reliant on the accuracy of the hypothesis posteriors. However, it has been empirically found that the LAS and RNN-T hypothesis posteriors tend to exhibit significant bias toward short token sequences [6, 7]. In fact, (2), (3), and (4) suggest that the hybrid, LAS, and RNN-T models may all innately exhibit biases toward shorter token sequences, since all hypothesis posteriors take the form of a product between posteriors of consecutive tokens or words. Each token posterior takes a value between zero and one. The posterior of a longer token sequence is composed of a product of more individual token posteriors, and may thus naturally tend to have a smaller value.

One common approach to correct for this innate bias in LAS and RNN-T is to re-rank the hypotheses in the final decoding beam with scores that are computed by scaling the hypothesis posteriors by a power, given by the inverse hypothesis length [6, 7]. This is referred to as length normalisation, and has been shown to yield gains when performing decoding by choosing the top re-ranked hypothesis [6, 7]. MBR combination, or decoding of a single model, can be used with these length-normalised scores, after re-normalising them to sum to one, to resemble posteriors. Although this preserves the rank order that length normalisation can yield, it may lose information about the relative probabilities between the competing hypotheses that a model originally produced. If a model was confident about its prediction, then it may have produced a low-entropy N-best list. Length normalisation does not preserve the relative magnitude of the entropies between different utterances. One possible method to preserve these entropies can be

to apply a per-utterance scaling factor to the length-normalised scores, such that the length-normalised N-best list has the same entropy as the original N-best list. However, initial MBR decoding experiments suggest that this entropy matching may not work well.

5. Minimum Bayes' risk training

Another approach to reduce the bias is to modify the training criterion. End-to-end NN models can be trained by minimising the negative log-probability of the reference [6, 7], referred to as the Conditional Maximum Likelihood (CML) criterion,

$$\mathcal{F}_{\text{CML}} = -\log P\left(\boldsymbol{\omega}^{\text{ref}} \middle| \mathbf{O}_{1:T}\right). \tag{7}$$

This aims to maximise the posterior probability of the reference word sequence, $\boldsymbol{\omega}^{\text{ref}}$. At the same time, the criterion also implicitly minimises the probabilities of alternative token sequences, including shorter token sequences, because of the softmax output of the model. However, this implicit minimisation of shorter token sequences may not be sufficiently strong, since it has been empirically observed that using length normalisation when performing recognition can still yield gains.

An alternative training criterion is to minimise the expected minimum edit distance relative to the reference [19], referred to as MBR training,

$$\mathcal{F}_{\text{MBR}} = \sum_{\boldsymbol{\omega}\in\mathbb{N}} \mathcal{L}\left(\boldsymbol{\omega}, \boldsymbol{\omega}^{\text{ref}}\right) \frac{P\left(\boldsymbol{\omega}|\mathbf{O}_{1:T}\right)}{\sum\limits_{\boldsymbol{\omega}'\in\mathbb{N}} P\left(\boldsymbol{\omega}'|\mathbf{O}_{1:T}\right)}. \tag{8}$$

Unlike (7), alternative hypotheses, including shorter token sequences, explicitly appear in (8). Therefore, minimising (8) explicitly minimises the probabilities of these alternative hypotheses. Since LAS and RNN-T tend to produce larger probabilities for shorter token sequences, it is likely that these will have greater contributions to the criterion. Thus the suppression of alternative hypotheses in (8) may have a greater impact on shorter than longer token sequences. This may reduce the model's bias toward shorter token sequences.

Applying length normalisation when performing recognition often yields gains. It may be possible to also benefit from Length Normalisation (LN) during MBR training, by replacing the hypothesis posteriors with the length-normalised scores,

$$\mathcal{F}_{\text{MBR-LN}} = \sum_{\boldsymbol{\omega}\in\mathbb{N}} \mathcal{L}\left(\boldsymbol{\omega}, \boldsymbol{\omega}^{\text{ref}}\right) \frac{P^{\frac{1}{|\boldsymbol{\omega}|}}\left(\boldsymbol{\omega}|\mathbf{O}_{1:T}\right)}{\sum\limits_{\boldsymbol{\omega}'\in\mathbb{N}} P^{\frac{1}{|\boldsymbol{\omega}'|}}\left(\boldsymbol{\omega}'|\mathbf{O}_{1:T}\right)}. \tag{9}$$

However, this criterion already explicitly reduces the bias of the model toward shorter token sequences, and therefore places less responsibility on the model to learn to reduce its own bias.

6. Experiments

Hybrid, LAS, and RNN-T models were each trained on 75 thousand hours of transcribed data from a variety of Microsoft applications. The models were evaluated on a variety of test sets, covering 13 application scenarios such as Cortana and far-field speech, using a total of 1.8 million words. All data was anonymised, with personally identifiable information removed. The results presented are the average Word Error Rates (WER) over all test scenarios.

The hybrid acoustic model was an ensemble of two layer-trajectory bi-directional Long Short-Term Memory (LSTM) networks [3], with 6 layers of 1024 and 832 nodes, and trained toward the $\mathcal{F}_{\text{MBR}}$ criterion. A 5-gram language model was used for recognition. The LAS comprised an encoder with 6 layers of 1024 bi-directional Gated Recurrent Unit (GRU) nodes, a decoder with 2 layers of 1024 GRU nodes, and an attention layer between the encoder and decoder. This LAS was first trained toward the $\mathcal{F}_{\text{CML}}$ criterion, then fine-tuned toward either the $\mathcal{F}_{\text{MBR}}$ or $\mathcal{F}_{\text{MBR-LN}}$ criterion. The RNN-T encoder network had 6 layers of 832 bi-directional LSTM nodes, each projected down to 400 dimensions [20]. The RNN-T prediction network had 2 layers of 1280 LSTM nodes, each projected down to 640 dimensions. The outputs of the encoder and prediction networks were combined through a single feed-forward layer with a softmax output. The RNN-T was trained toward the $\mathcal{F}_{\text{CML}}$ criterion, as work in [17] suggests that MBR training of the RNN-T may not yield significant gains. The LAS and RNN-T each used a different set of 4000 sub-word units as outputs.

MBR decoding and combination were performed using the Kaldi toolkit [21, 13]. N-best lists of size 16 were generated from each model. These were converted to the Kaldi lattice format, determinised, scaled, and normalised to sum to one. Separate posteriors were available for each hypothesis, but not for each word. MBR decoding or combination was performed on these lattices. Scaling factors, tuned on held out data, were applied to the hypothesis posteriors of each model.

6.1. MBR decoding of a single end-to-end NN model

MBR combination and decoding rely on the accuracy of the hypothesis posteriors. However, the posteriors of end-to-end NN models may be biased toward short token sequences. This may degrade the effectiveness of MBR decoding. Length normalisation and MBR training may reduce this bias. Table 1 assesses the interactions between MBR decoding, MBR training, and length normalisation, for LAS. The MBR decoding performance is compared against a baseline decoding method of choosing the hypothesis with the top score from the N-best list [6, 7], referred to as 1-best decoding. This is similar to Viterbi decoding [22] of a hybrid model.

Table 1: *Impact of MBR training and length normalisation on MBR decoding for LAS*

Training	Decoding length norm	WER (%) 1-best	WER (%) MBR
$\mathcal{F}_{\text{CML}}$	no	10.40	9.15
	yes	7.90	8.42
$\mathcal{F}_{\text{MBR}}$	no	8.95	8.82
	yes	7.92	8.53
$\mathcal{F}_{\text{MBR-LN}}$	no	9.29	8.76
	yes	7.85	8.42

With $\mathcal{F}_{\text{CML}}$ training, the performance of choosing the 1-best from the N-best list significantly improves when the list is reranked using the length-normalised scores. The insertion and deletion rates without length normalisation are 0.79 and 4.82%, and with length normalisation are 1.32 and 1.38%. This supports the observation in [6] that LAS tends to give posteriors that are biased toward short token sequences. When LAS is trained using either $\mathcal{F}_{\text{MBR}}$ or $\mathcal{F}_{\text{MBR-LN}}$, the gap between the 1-best decoding performances with and without applying length normalisation during decoding decreases, primarily accounted for by a decrease in the deletion rate. This decrease in the gap is

most significant for $\mathcal{F}_{\text{MBR}}$ training, which places more responsibility on the model to correct for its own bias.

MBR decoding yields gains compared to 1-best decoding when length normalisation is not used during decoding. One approach for MBR decoding to benefit from the re-ranking of length normalisation is to apply length normalisation to the hypothesis posteriors, re-normalise the resulting scores to sum to one over the N-best list, then perform MBR decoding using these new scores. This improves the MBR decoding performance over using MBR decoding without length normalisation. However, the standard 1-best decoding with length normalisation still performs the best. It is difficult to predict the behaviour of performing MBR decoding with length normalised scores, since the entropy of the length-normalised scores is not easily interpretable. Although MBR training may not yield significantly better performance than $\mathcal{F}_{\text{CML}}$, MBR training does reduce the model's bias toward short token sequences.

Table 2: *MBR decoding of different models*

	WER (%)	
Model	1-best	MBR
Hybrid	8.03	8.01
LAS	9.29	8.76
+ length norm	7.85	8.42
RNN-T	8.68	8.75
+ length norm	8.16	8.16

Table 2 assesses MBR decoding and length normalisation applied to the N-best lists for all three model architectures. The hybrid, LAS, and RNN-T models were trained with the $\mathcal{F}_{\text{MBR}}$, $\mathcal{F}_{\text{MBR-LN}}$, and $\mathcal{F}_{\text{CML}}$ criteria respectively. MBR decoding, without length normalisation, of the hybrid and RNN-T models does not yield significant gains over 1-best decoding. This may be due to the small N-best list size of 16. It may also be possible that the hybrid model may not benefit much from MBR decoding when it has been trained using a large quantity of data. The gap between the RNN-T 1-best decoding performances with and without length normalisation is smaller than that for LAS. This suggests that RNN-T may exhibit less bias toward short token sequences that LAS, even when RNN-T has only been trained with the $\mathcal{F}_{\text{CML}}$ criterion. Although MBR decoding may not yield gains over 1-best decoding, it is still important to assess its interaction with the LAS and RNN-T models, as MBR combination is likely to have a similar behaviour.

6.2. Combination

This section assesses combination between the hybrid, LAS, and RNN-T models. The 1-best decoding performances of each of the three models are shown in Table 3. The LAS here was trained with the $\mathcal{F}_{\text{MBR-LN}}$ criterion.

Table 3: *Single model performance*

Model	WER (%)
hybrid	8.03
LAS	7.85
RNN-T	8.16

MBR combinations between these models are shown in Table 4. Length normalisation was applied to both the LAS and RNN-T N-best lists before combination. Equal combination weights were used. The right most column shows the relative WER Reduction (WERR) against the best single model within each respective combination. Combinations between any two of the models yield gains. However, comparing the relative WERRs between the combinations may not reliably indicate the diversity between any two models, as the gain may be affected by the interactions between length normalisation, posterior scaling, and MBR decoding. Combining three models yields more gain than combining any two models, suggesting that the third model still contributes additional complementary diversity.

Table 4: *MBR combinations*

Combination	WER (%)	Relative WERR (%)
hybrid + LAS	7.32	6.8
hybrid + RNN-T	7.26	9.6
LAS + RNN-T	7.62	2.9
hybrid + LAS + RNN-T	6.89	12.2

The final experiment compares MBR combination with two other combination methods. ROVER combination here used only the 1-best hypotheses from each model. Only majority voting was used, without any confidence scores, because it is not straight forward to obtain per-token confidence scores from LAS and RNN-T. An un-tuned null confidence of 0.5 was used. Another combination approach is to form a union of the N-best lists from each model, then choose the top hypothesis from the merged N-best list. Length normalisation was applied to the posteriors from LAS and RNN-T. These scores were scaled, and then normalised to sum to one over each N-best list. The scores were then treated as posteriors, and merged across the multiple N-best lists as a sum.

Table 5: *Method of combining hybrid, LAS, and RNN-T models*

Combination method	WER (%)
1-best of merged N-best	7.59
ROVER	7.33
MBR	6.89

Initial experiments were performed to compare Kaldi MBR decoding with the SRILM N-best ROVER implementation [23, 14], and suggested that the former performed better. N-best ROVER and the Kaldi implementation of MBR decoding are different approximations to MBR decoding. Another approximation to MBR decoding is HTK confusion network decoding [24, 15]. However, it is not straight forward to use HTK confusion network decoding and combination with LAS hypotheses, because time alignments are not available. Out of these MBR decoding and combination approximations, only the results of the Kaldi implementation [13] are shown in this paper. The results of combining three models in Table 5 suggest that MBR combination yields a larger gain than both ROVER and choosing the top hypothesis from the merged N-best list.

7. Conclusion

This paper has studied combination between hybrid and end-to-end NN models. Combination yielded significant gains, suggesting that the models are complementary. MBR combination relies on the accuracy of the hypothesis posteriors. MBR training and length normalisation can reduce the bias of the hypothesis posteriors toward short token sequences.

8. References

[1] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Kluwer Academic Publishers, 1994.

[2] J. Li, R. Zhao, E. Sun, J. H. M. Wong, A. Das, Z. Meng, and Y. Gong, "High-accuracy and low-latency speech recognition with two-head contextual layer trajectory LSTM model," in *ICASSP*, Barcelona, Spain, May 2020, pp. 7699–7703.

[3] E. Sun, J. Li, and Y. Gong, "Layer trajectory BLSTM," in *Interspeech*, Graz, Austria, Sep 2019, pp. 1403–1407.

[4] S. H. K. Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP*, Brighton, UK, May 2019, pp. 6670–6674.

[5] G. Pundak and T. N. Sainath, "Lower frame rate neural network acoustic models," in *Interspeech*, San Francisco, USA, Sep 2016, pp. 22–26.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *ICASSP*, Shanghai, China, Mar 2016, pp. 4960–4964.

[7] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Representation Learning Workshop*, Edinburgh, UK, Jul 2012.

[8] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-Y. Chang, W. Li, R. Alvarez, Z. Chen, C.-C. Chiu, D. Garcia, A. Gruenstein, K. Hu, M. Jin, A. Kannan, Q. Liang, I. Mcgraw, C. Peyser, R. Prabhavalkar, G. Pundak, D. Rybach, Y. Shangguan, Y. Sheth, T. Strohman, M. Visontai, Y. Wu, Y. Zhang, and D. Zhao, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP*, Barcelona, Spain, May 2020, pp. 6059–6063.

[9] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, Calgary, Canada, Apr 2018, pp. 4774–4778.

[10] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S.-Y. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP*, Brighton, UK, May 2019, pp. 6381–6385.

[11] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving RNN transducer modeling for end-to-end speech recognition," in *ASRU*, Singapore, Dec 2019, pp. 114–121.

[12] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[13] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, Oct 2011.

[14] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, Oct 2000.

[15] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Speech Transcription Workshop*, May 2000.

[16] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, and C.-C. Chiu, "Two-pass end-to-end speech recognition," in *Interspeech*, Graz, Austria, Sep 2019, pp. 2773–2777.

[17] K. Hu, T. N. Sainath, R. Pang, and R. Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," in *ICASSP*, Barcelona, Spain, May 2020, pp. 7799–7803.

[18] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, New Orleans, USA, Mar 2017, pp. 4835–4839.

[19] J. Kaiser, B. Horvat, and Z. Kačič, "Overall risk criterion estimation of hidden Markov model parameters," *Speech Communication*, vol. 38, no. 3-4, pp. 383–398, Nov 2002.

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

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *ASRU*, Hawaii, USA, Dec 2011.

[22] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr 1967.

[23] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: update and outlook," in *ASRU*, Hawaii, USA, Dec 2011.

[24] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. C. Woodland, and C. Zhang, *The HTK book*, Dec 2015.