Revisiting Dynamic Adjustment of Language Model Scaling Factor for Automatic Speech Recognition

Hiroshi Sato*, Takafumi Moriya*, Yusuke Shinohara*, Ryo Masumura*, Takaaki Fukutomi†,
Kiyoaki Matsui*, Takanori Ashihara*, Yoshikazu Yamaguchi* Yushi Aono*
* NTT Media Intelligence Laboratories, NTT Corporation, Japan
E-mail: hiroshi.satou.bh@hco.ntt.co.jp
† NTT TechnoCross, Japan

***Abstract*—Automatic speech recognition (ASR) systems use the language model scaling factor to weight the probability output by the language model and balance it against those from other models including acoustic models. Although the conventional approach is to set the language model scaling factor to a constant value to suit a given training dataset to maximize overall performance, it is known that the optimal scaling factors varies depending on individual utterances. In this work, we propose a way to dynamically adjust the language model scaling factor to a single utterance. The proposed method utilized a recurrent neural network (RNN) based model to predict optimum scaling factors given ASR results from a training dataset. Some studies have already tackled this utterance dependency in the 2000s, yet few have improved the quality of ASR due to the difficulty in directly modeling the relationship between a series of acoustic features and the optimal scaling factor; a recent breakthrough in RNN technology has now made this feasible. Experiments on a real-world dataset show that the dynamic optimization of the language model scaling factor can improve ASR quality and that the proposed method is effective.**

I. Introduction

Language model scaling factor, or language model weight, is a parameter introduced to balance acoustic and language models usually used in Deep Neural Network-Hidden Markov Model (DNN-HMM) based hybrid ASR systems. Despite the recent attention being placed on end-to-end ASR modeling, DNN-HMM based systems are still popular for practical applications because their learnability can offset scant training data. In addition, joint nature of end-to-end ASR system that acoustic and linguistic elements are inseparable, makes it difficult to add new words or acoustic characteristics without re-training with whole dataset. In this respect, DNN-HMM systems composed of independent acoustic and language models has advantages.

DNN-HMM based ASR systems search for the best word sequence $\hat{\boldsymbol{w}}$, i.e., the one that maximizes posterior probability $p(\hat{\boldsymbol{w}}|\boldsymbol{X})$, given a series of acoustic features $\boldsymbol{X}$. The problem of maximizing the posterior probability for a word sequence $\boldsymbol{w}$ is usually transformed into maximization of a product of acoustic likelihood $p(\boldsymbol{X}|\boldsymbol{w})$, yielded by an acoustic model, and linguistic a priori probability $p(\boldsymbol{w})$, output by a language model, using Bayes formula.

However, in practice there is often a mismatch between the output range of these two models. This necessitates the introduction of a parameter called the language model scaling factor or language model weight, to balance these models. The best word sequence $\hat{\boldsymbol{w}}$ is chosen from hypotheses $\boldsymbol{w}$ according to Eq. (1) below.

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} (\log p(\boldsymbol{X}|\boldsymbol{w}) + \lambda \log p(\boldsymbol{w})) \tag{1}$$

The language model scaling factor λ is one of the decoding parameters that should be adjusted prior to recognition, and literature indicates that its optimal value depends on the target dataset. Accordingly, the optimization of decoding parameters, including the language model scaling factor, to a target dataset has widely been investigated, and thus the methods for solving the problems are well established [1]–[4]. However the optimal language model scaling factor varies not only between datasets but also between individual utterances in the target dataset. Therefore, dynamically adjusting the language model scaling factor to each utterance has the potential to better predict a correct word sequence. Thus this study investigates the dynamic-adaptation of the language model scaling factor to each utterance.

Dynamic optimization of the language model scaling factor is a problem that has not been exhaustively addressed. To the best of our knowledge, there are a few studies in this field. It has been suggested that the reliability of acoustic and language models calculated from acoustic observations and grammatical knowledge can be used to optimize the scaling factor for each-utterance [5]. However, no specific methods were proposed in [5]. In [6], a method to determine the language model scaling factor from a state of dialog or from each utterance was investigated and the relationship between the scaling factor and recognition accuracy was analyzed, but the improvement attained was limited because of the difficulty of clarifying which information should be utilized to determine optimality. All prior papers noted the difficulty of estimating the language model scaling factor, since at the time these papers were published, it was difficult to directly model the relationship between a series of acoustic features and the optimal scaling

factor; inevitably the researchers had to manually identify the features that represented the key characteristics.

However, the recent breakthroughs seen in deep neural networks and recurrent neural networks now enables the automatic extraction of key characteristics from a series of features and directly predict target labels from the characteristics. Thus in this research, we utilize the recurrent neural network to model the relationship between acoustic features and the optimal language model scaling factor.

In our approach, the scaling factor was not only used to adjust the output range of models but also to balance the reliability of each model given the utterances. For example, it is empirically known that if recordings are noisy and acoustic observation is not reliable, the increment of the scaling factor makes the system rely more on the language model outputs than acoustic outputs in determining the best hypothesis. Therefore, it is expected that the proposed method where the scaling factor is determined given acoustic observations can improve the recognition performance in the domain that has a wide variation in the quality of acoustic observation. In this paper, we show the effectiveness of our approach on real-world data collected from a personal assistant application.

The remainder of the paper is organized as follows. Section II describes related research. Section III details our analyses of the scaling factor in order to show the importance of dynamically adjusting it to each utterance. Section IV describes our method while Section V introduces the experimental settings. Section V explains the details of experimental settings. Section VI provides the results of our experiments and explains the meaning. Finally, we conclude in Section VII.

II. Related work

The idea of adjusting the language model scaling factor to each utterance was first posed as the word dependent language model scaling factor [7], [8]. In [7], the language model scaling factor was modified for each word by a word boosting factor that was trained adaptively to reduce recognition error against a training-set. A word and pronunciation dependent scaling factor was introduced to deal with the variation in discrimination performance of the acoustic model across pronunciation [9]. A context-dependent language model scaling factor was also proposed [10]. They incorporated interpolation weights of multiple language models into one framework, and the weights were trained in a discriminative manner using N-best candidates. These studies utilized lexical characteristics such as grammatical or phonetical information to adjust the language model scaling factor, while our appoarch is based on acoustic cues that impacts the reliability of the acoustic model output.

Uncertainty weighting [11], [12] was proposed to weight the information provided by acoustic observations according to observation reliability. It can be seen as a dynamic adjustment of the language model scaling factor in every frame according to acoustic observations. An uncertainty decoding scheme for DNN-HMM was recently investigated [13], [14]. A recent study weighs the acoustic observation at each frame by an uncertainty variance that is determined from an estimate of noise [14]. While approaches such as modifying the language model scaling factor according to noise estimates are proposed in [5], noise observation is not always enough to estimate the reliability of the acoustic model outputs. There are two reasons. One is that there are other acoustic phenomena that affect the reliability of acoustic outputs: for example, clipping and unclear pronunciation, etc. The other is that the uncertainty of 'observation' is not always enough, by itself, to accurately estimate the reliability of the acoustic model output; discriminability of the model against the uncertainty should also be taken into account. In the proposed method it is unnecessary to enumerate artificially the factor that affects the reliability of the model since the relationship between input features and the scaling factor is directly modeled. Moreover, the training is conducted on recognition hypotheses, which enabled to consider the discriminability of the models.

Some approaches have also been proposed to adjust decoding parameters, other than the language model scaling factor, to an utterance. Online control of decoding beam width has been investigated over the last few decades [15]–[17]. In [15], beam width is adaptively controlled in order to reduce the time needed to decode low-quality speech; more computation time is needed due to the more dispersed confidence score distributions of the hypotheses.

III. Importance of dynamic adjustment of the language model scaling factor

In this section we investigate the optimal value of the language model scaling factor for each utterance; we analyze their variation and its potential for reducing the recognition error. The model is trained on the data obtained from the target domain and used as a preprocessing step of ASR to determine the correct scaling factor at decoding.

A. Dataset and model

The analyses were conducted on professionally hand-transcribed and anonymized utterances collected from a personal assistant application. The utterances are naturally distorted by various noises such as background speech, music and TV noise, etc. Incorrectly recorded audio that only included background noise or speech that were not directed to the devices were excluded from the analyses. The dataset contains about 3.5k utterances which correspond to about 2 hours of recording, which were consist mostly of short utterances. The recordings were formatted in 16 kHz, 16 bit linear-PCM.

The acoustic model was trained on about 16000 hours recording of Japanese multi-conditioned data from a voice search application. The data was noised by living room (TV chatting noise) and kitchen (cooking and washing noise) noise recording to simulate home-use environment. The detailed model structure is described in [18]. The model was trained on cross entropy loss [19]. 4-gram language model was adopted in this research; it has about 700K size vocabulary and was trained on various text corpora. Decoding was performed by

the weighted finite-state transducer (WFST-) based decoder VoiceRex [20], [21].

B. Analysis

Utterances were recognized using the language model scaling factor values from 1 to 30 with increments of 1, and then one-best hypotheses at each value of the scaling factor were evaluated to calculate character error rate (CER) and sentence error rate (SER) at each scaling factor. We conducted two analysis of them.

1) performance of the constant scaling factor and adaptive scaling factor: Fig. 1 shows SER and CER at a constant scaling factor (values varied from 1 to 30) and at an oracle scaling factor that minimizes the character error rate for each utterance. The perfect prediction of the best scaling factor at every utterance would realize about 34.0 % relative reduction in SER and 44.5 % in CER, compared with using a constant scaling factor that is optimum for the entire dataset represented as a minimum value of the solid line. This result shows the potential of dynamic adjustment of the language score factor in reducing ASR errors, provided we can sufficiently extract and model the factors that affect the balance between the reliability of the acoustic model and the language model.

2) optimal scaling factor for each utterance: Fig. 2 shows the distribution of the language model scaling factor that makes the one-best hypothesis correct. Each utterance belongs to one of three groups: (i) samples whose one best hypothesis is correct at a certain scaling factor that minimizes SER over the entire dataset, (ii) samples whose one best hypothesis is not correct at the best scaling factor over the entire dataset but can be correct if other scaling factors are used, and (iii) samples whose hypotheses are never correct regardless of the scaling factor. In Fig. 2 we plot group (ii) samples as the minimum value of the displacement of the scaling factor for the sample hypotheses to be correct. For example, if the scaling factor that minimizes SER over the dataset was 15, and the hypothesis to this utterance became correct at the scaling factor of 17 or more, this utterance was aggregated into the bin marked +2 in the graph. The analysis results show that the scaling factors that yield true recognition distribute widely between 1 to 30, which contradicts the common assumption that the scaling factor can be set to a constant value.

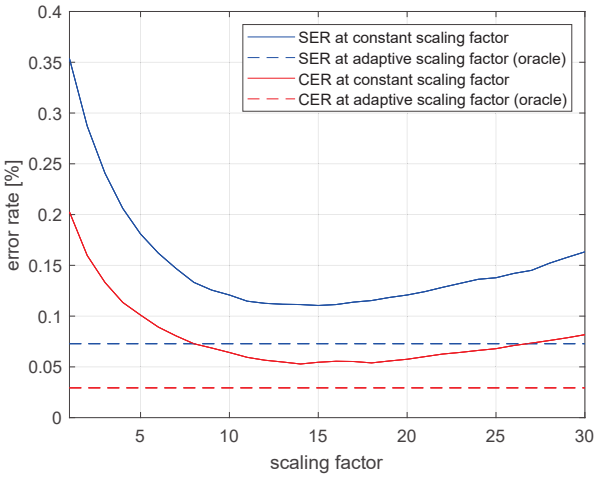

Fig. 1. Sentence error rate and character error rate at a constant scaling factor (values from 1 to 30) and an oracle scaling factor adjusted to each utterance. The former is shown by the solid line and the x-axis represents the corresponding constant value. The latter is shown by the dotted line. Optimal scaling factor would realize about 34.0 % relative reduction in SER and 44.5 % in CER to a constant scaling factor.

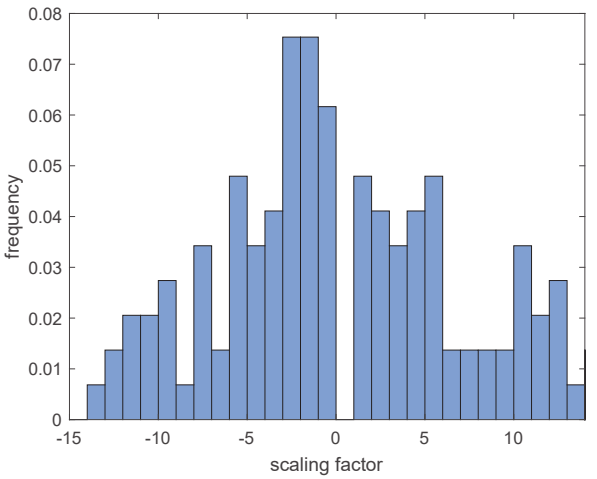

Fig. 2. The distribution of the language model scaling factor that validate the one-best hypothesis. X-axis means the minimum displacement of the scaling factor from the optimal value over the dataset where the one-best hypothesis for each utterance was correct. The optimum scaling factors distribute widely between 1 to 30.

IV. Methods

In this section, we describe the details of the framework to obtain a model that predicts the optimal language model scaling factor from sequences of acoustic features for each utterance.

A. General formulation

Eq. (1) is the general formulation that determines the word series $\hat{\boldsymbol{w}}$ that best matches the acoustic observation $\boldsymbol{X}$. The language model scaling factor, λ, generally takes a value within $[0, \infty)$. For convenience of modeling, we converted Eq. (1) into Eq. (2) below,

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}}((1-\tilde{\lambda})\log p(\boldsymbol{X}|\boldsymbol{w}) + \alpha\tilde{\lambda}\log p(\boldsymbol{w})) \quad (2)$$

where $\tilde{\lambda}$ is a scaled language model scaling factor that takes a value within $0 \leq \tilde{\lambda} \leq 1$, and α is a constant. From Eq. (1) and Eq. (2), $\tilde{\lambda}$ is represented by λ and α as follows; $\tilde{\lambda} = \lambda/(\lambda+\alpha)$, which means that scaling factor $\tilde{\lambda}$ becomes 0.5 when $\lambda = \alpha$. α was set to the value that minimizes the sentence error rate over the training dataset, and thus $\tilde{\lambda} = 0.5$ was the most general value of the scaling factor that is expected to have the highest probability of yielding correct one-best hypotheses.

B. Model

We introduced utterance-level language scaling factor prediction model $\tilde{\lambda} = \phi(\boldsymbol{X}; \boldsymbol{\theta})$ that predicts the optimal scaling factor $\tilde{\lambda}$ from a sequence of acoustic features $\boldsymbol{X}$, which was modeled by deep-neural-network. The model parameters $\boldsymbol{\theta}$ were trained to predict a better scaling factor $\tilde{\lambda}(\boldsymbol{\theta})$ that generate hypotheses $\hat{\boldsymbol{w}}$ with lower error rate. Evaluation against the $\tilde{\lambda}(\boldsymbol{\theta})$ was conducted in minimizing error rate basis using generated hypothesis $\hat{\boldsymbol{w}}(\tilde{\lambda}(\boldsymbol{\theta}))$ and transcription $\tilde{\boldsymbol{w}}$.

C. Loss

In order to assign an optimum scaling factor for each utterance, we adopted minimum error rate training using discriminative loss function calculated with N-best candidates [22]. Eq. (3) and (4) shown below are the formulation of discriminative loss fucntion.

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \frac{\sum_{k=1}^{K} E_{n,k} \cdot \exp(\beta \cdot l_{n,k}(\boldsymbol{\theta}))}{\sum_{k=1}^{K} \exp(\beta \cdot l_{n,k}(\boldsymbol{\theta}))} \quad (3)$$

$$l_{n,k}(\boldsymbol{\theta}) = (1-\tilde{\lambda_n}) \log p(\boldsymbol{X}_n|\boldsymbol{w}_{n,k}) + \alpha \tilde{\lambda_n}(\boldsymbol{\theta}) \log p(\boldsymbol{w}_{n,k}) \quad (4)$$

$\mathcal{L}(\boldsymbol{\theta})$ represents the total loss over N utterances and $E_{n,k}$ represents the error rate of k-th candidate of n-th data; character error rate was adopted in this paper. β is a parameter to adjust the smoothness of objective function. $l_{n,k}(\boldsymbol{\theta})$ represents the score of a candidate that was calcurated as weighted sum of acoustic and language score. In this paper, we optimize the model parameter for predicting a language scaling factor λ by minimizing this loss function. The minimization of the loss function decrease the score of a candidate of worse error rate and accordingly raise the probability of correct candidated to appear in higher rank.

V. Experimental details

We constructed a scaling factor prediction model and evaluated recognition accuracy in combination with an ASR system detailedly described in Section. III-A. Input features of the model were sequences of 40 dimensional log mel-scale filterbank coefficients that were extracted from utterances with the setting of 20 ms window width and 10 ms frameshift along with dynamic features (Δ and $\Delta\Delta$) . The structure of the model is presented in the Fig. 3. We used 2 layer bidirectional LSTM with 128 units at the bottom layer and introduced the self-attention layer over the outputs of LSTM layers [23]. A linear layer of 64 units with ReLU non-linear activation, then a linear layer of 1 unit with Sigmoid non-linear activation were stacked over the attention layer to get one-dimensional predictions $0 \leq \tilde{\lambda} \leq 1$. We used dropout for LSTM layer, attention layer and the first linear layer with the dropout rate of 0.5 at training. The optimizer was Adam with weight decay of 10^{-7} [24]. Learning rate was 0.0001. Hyperparameters were tuned on the validation dataset.

The dataset used in the experiments was anonymized real recording of utterances collected from a personal assistant application. The dataset contains about 5k utterances which correspond to about 3 hours of recording. In the experiments, background speech or noise that were recorded but not directed to the devices were not excluded for more realistic verification. The dataset was randomly split to train, validation and test dataset at the ratio of 7:1:2. The result shown in Section VI is the average of 5 times of trials. The data was split differently for each trial.

In this experiment, the number of candidates to calculate the discriminative loss was set as $N = 100$. The 100-best candidates were generated by recognizing at the language scaling factor of 15, which was the optimal value over this dataset. The smoothing parameter β in Eq. (3) was set as 0.1.

It was assumed that provide the number of candidates in the N-best hypotheses were large, the best hypothesis rescored at a certain scaling factor within 100-best, approximated the one-best hypothesis generated at the same scaling factor. ASR performance was evaluated by rescoring the 100 hypotheses at the predicted scaling factor instead of basing recognition on the predicted scaling factor.

TABLE I
The result of recognition on the dataset. The relative sentence error rate reduction of 11.1 % and character error rate reduction of 5.6 % were achieved compared with the use of a constant scaling factor.

	SER [%]	CER [%]
Fixed scaling factor	27.52	21.07
Adaptive scaling factor	24.44	19.89
Oracle scaling factor	21.66	18.27

VI. Results and Discussions

Table. I shows the results achieved on the dataset. The line "Fixed scaling factor" stands for the recognition performance with the constant language model scaling factor that minimized the sentence error rate over the entire training dataset. The line "Adaptive scaling factor" is the result acquired by the proposed method of predicting the scaling factor for each utterance. The line "Oracle scaling factor" is the performance

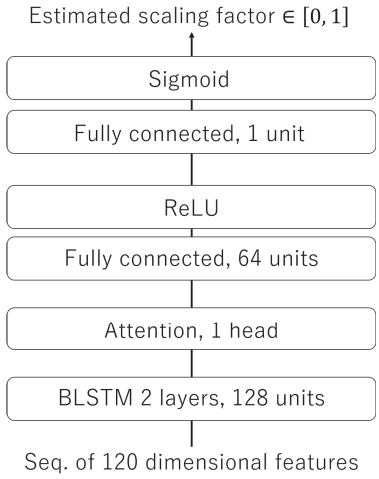

Fig. 3. The architecture of the RNN model.

TABLE II
EXAMPLES OF IMPROVEMENT OBSERVED WITH THE PROPOSED METHOD. IN (A), A CASE PARTICLE OMITTED IN A SPOKEN LANGUAGE WAS CORRRECLY FILLED, AND IN (B), AN EUPHONICAL CHANGE WAS CORRECTLY RECOGNIZED BY ADJUSTING A SCALING FACTOR TO CONFIDE MORE ON THE ACOUSTIC MODEL. IN (C), THE HYPOTHESIS WAS MODIFIED TO GRAMMATICALLY CORRECT SENTENCE BY INCRESING THE SCALING FACTOR. IN (D), A BACKGROUND SPEECH THAT WAS NOT REJECTED BY VAD WAS CORRECTLY REJECTED BY A HIGHER SCALING FACTOR.

	ref.	hyp. with constant scaling factor	hyp. with adaptive scaling factor	predicted scaling factor
(a)	メモ 保存 する Save the note.	メモ を 保存 する	メモ 保存 する	0.43
(b)	どこ の 県 に あんの In which prefecture?	どこ の 県 に ある の	どこ の 県 に あん の	0.31
(c)	何 が 見つから なかっ た の What was not found?	何 だ 見つから なかっ た の	何 が 見つから なかっ た の	0.60
(d)	NA	でもいい	NA	0.63

achieved when we can choose the best scaling factor for each utterance within the realistic range of $5 \leq \lambda \leq 30$.

In Table I, the sentence error rate reduction of 11.1 % and character error rate reduction of 5.6 % were achieved by dynamically predicting the optimal value (relative to adopting a fixed best value for the dataset). It can be said that the acoustic features contain the information that determines the optimum scaling factor and the proposed method succeeded in partially modeling the relationship between them. A possible explanation of the relatively small reduction rate of the proposed method compared with using oracle weight is that the language model scaling factor balances the reliability of the acoustic model and the language model, and the proposed method can only capture the variation in the reliability of the acoustic model. Since we only utilized acoustic features as the input of the model, the models were not able to capture the grammatical factors that affect the optimum scaling factor.

The Table II shows some examples that were corrected by the dynamic adjustment of the scaling factor. Rows (a), (b) are the examples that were validated by decreasing the scaling factor, which means to put more confidence in the acoustic model. In example (a), a case particle 'を' was not pronounced actually because it can be omitted in the spoken language. At a constant scaling factor, 'を' appeared in the hypothesis since grammatically it is more likely to use 'を'. With the proposed method, the case particle 'を' was correctly omitted by relying more on acoustic output. In example (b), 'ある' was grammatically correct but euphonically changed into 'あん' at utterance. This case was also validated by putting more confidence in an acoustic model according to the prediction of the reliability of the models judged from acoustic features. Since the recordings in this dataset were naturally distorted by various factors and diverse in acoustic reliability, the optimal language model scaling factor tended to be set higher to some clean data that were easy to recognize acoustically. The proposed method can correct this tendency and improve recognition in these cases.

Rows (c), (d) are the examples that became correct by increasing the scaling factor, which means to put more confidence in the language model. In case (c), the utterance was recognized as '何 だ' at a constant scaling factor, but it is grammatically incorrect even in spoken language. In this case, the recognition was corrected by relying more on language observation. In another case (d), a background speech that was not intended to be recognized but not rejected by VAD was recognized. In such cases, a high language weight scaling factor prone to output empty recognition results and with the proposed method, this tendency was captured and the utterances were correctly rejected.

VII. CONCLUSIONS

In this paper, we proposed a method to dynamically adjust the language model scaling factor to each utterance. An RNN-based model was introduced to predict optimum scaling factors directly from a series of acoustic features, which was previously thought to be an impossibility. Experiments showed that the dynamic optimization of the language model scaling factor has the potential to improves ASR quality and that the proposed method was actually effective for a real-world dataset.

REFERENCES

[1] T. Le Nguyen, D. Stein, and M. Stadtschnitzer, "Gradient-free decoding parameter optimization on automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 3261–3265.

[2] J. Schwenninger, D. Stein, and M. Stadtschnitzer, "Automatic parameter tuning and extended training material: Recent advances in the fraunhofer speech recognition system," 2013, pp. 3002–3011.

[3] B. Mak and T. Ko, "Automatic estimation of decoding parameters using large-margin iterative linear programming," in *Annual Conference of the International Speech Communication Association*, 2009.

[4] A. Ito, M. Kohda, and S. Makino, "Fast optimization of language model weight and insertion penalty from n-best candidates," *Acoustical science and technology*, vol. 26, no. 4, pp. 384–387, 2005.

[5] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech communication*, vol. 18, no. 3, pp. 205–231, 1996.

[6] G. Stemmer, V. Zeissler, E. Nöth, and H. Niemann, "Towards a dynamic adjustment of the language weight," in *International Conference on Text, Speech and Dialogue*. Springer, 2001, pp. 323–328.

[7] R. R. Sarukkai and D. H. Ballard, "Word set probability boosting for improved spontaneous dialog recognition," *IEEE transactions on speech and audio processing*, vol. 5, no. 5, pp. 438–450, 1997.

[8] X. Huang, M. Belin, F. Alleva, and M. Hwang, "Unified stochastic engine (use) for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1993, pp. 636–639.
[9] B. Hoffmeister, R. Liang, R. Schlüter, and H. Ney, "Log-linear model combination with word-dependent scaling factors," in *Annual Conference of the International Speech Communication Association*, 2009.
[10] S. Chang, A. Lahiri, I. Alphonso, B. Oğuz, M. Levit, and B. Dumoulin, "Discriminative training of context-dependent language model scaling factors and interpolation weights," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2015, pp. 45–51.
[11] N. B. Yoma, F. R. McInnes, and M. A. Jack, "Improving performance of spectral subtraction in speech recognition using a model for additive noise," *IEEE Transactions on speech and audio processing*, vol. 6, no. 6, pp. 579–582, 1998.
[12] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with splice for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I57–I60.
[13] C. Huemmer, A. Schwarz, R. Maas, H. Barfuss, R. F. Astudillo, and W. Kellermann, "A new uncertainty decoding scheme for dnn-hmm hybrid systems with multichannel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 5760–5764.
[14] J. Novoa, J. Fredes, V. Poblete, and N. B. Yoma, "Uncertainty weighting and propagation in dnn-hmm-based speech recognition," *Computer Speech & Language*, vol. 47, pp. 30–46, 2018.
[15] S. Kobashikawa, T. Hori, Y. Yamaguchi, T. Asami, H. Masataki, and S. Takahashi, "Efficient beam width control to suppress excessive speech recognition computation time based on prior score range normalization," in *Annual Conference of the International Speech Communication Association*, 2012.
[16] S. Abdou and M. S. Scordilis, "Beam search pruning in speech recognition using a posterior probability-based confidence measure," *Speech Communication*, vol. 42, no. 3-4, pp. 409–428, 2004.
[17] H. Van Hamme and F. Van Aelten, "An adaptive-beam pruning technique for continuous speech recognition," in *International Conference on Spoken Language Processing*, vol. 4. IEEE, 1996, pp. 2083–2086.
[18] T. Moriya, H. Kanagawa, K. Matsui, T. Fukutomi, Y. Shinohara, Y. Yamaguchi, M. Okamoto, and Y. Aono, "Efficient building strategy with knowledge distillation for small-footprint acoustic models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 21–28.
[19] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
[20] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "Voicerex - spontaneous speech recognition technology for contact-center conversations," *NTT Technical Review*, vol. 5, no. 1, pp. 22–27, 2007.
[21] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient wfst-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 4, pp. 1352–1365, 2007.
[22] F. J. Och, "Minimum rrror rate training in statistical machine translation," in *Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2003, pp. 160–167.
[23] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.