INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Joint End-to-End and DNN-HMM Hybrid Automatic Speech Recognition System with Transferring Sharable Knowledge

Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, Takanobu Oba, Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation
tomohiro.tanaka.ht@hco.ntt.co.jp

Abstract

This paper presents joint end-to-end and deep neural network-hidden Markov model (DNN-HMM) hybrid automatic speech recognition (ASR) systems that share network components. End-to-end ASR systems have been shown competitive performance compared with the DNN-HMM hybrid ASR systems in recent studies. These systems have different advantages, which are an estimation ability based on the totally optimized model of the end-to-end ASR system and a stable processing based on a frame-by-frame manner of the DNN-HMM hybrid ASR system. In our previous study, we proposed a method to utilize an end-to-end ASR system for rescoring hypotheses generated from a DNN-HMM hybrid ASR system. However, the conventional method cannot efficiently leverage the advantages since network components are independently modeled. In order to tackle this problem, we propose a joint end-to-end and DNN-HMM hybrid ASR systems that share the network to transfer knowledge of the systems. In the proposed method, end-to-end ASR systems utilize the information from an output of an internal layer in a DNN acoustic model in the DNN-HMM hybrid ASR system for enhancing the end-to-end ASR system. This enables us to efficiently leverage sharable information for improving the joint ASR system. Experimental results show that the proposed method outperforms the conventional method.

Index Terms: speech recognition, joint automatic speech recognition system, end-to-end speech recognition system, DNN-HMM hybrid ASR system

1. Introduction

The performance of automatic speech recognition (ASR) systems has been dramatically improved by deep neural networks (DNNs). The most successful method in the last decade is DNN-hidden Markov model (DNN-HMM) hybrid ASR system in which DNNs are applied to frame-level context-dependent acoustic models. The DNN acoustic models have shown significant performance improvements through the investigation of several network topologies [1–4]. On the other hand, end-to-end ASR systems that directly convert a speech into symbols (character, words, etc.) have been paid much attention. In recent studies, various end-to-end ASR methods including connectionist temporal classification [5–8], speech-to-text encoder-decoder models [9–11], and recurrent neural network transducers [12, 13] have been investigated.

Our motivation is to fully leverage these two different ASR systems for improving ASR performance. To this end, we focus on the fact that the both ASR systems have advantages and disadvantages. The advantage of the DNN-HMM hybrid ASR system is stable processing to estimate phoneme states in a frame-by-frame manner. But, cascading processing using not only the acoustic models but also pronunciation models and language models omits total optimization. By contrast, end-to-end ASR systems can achieve the total optimization. But, end-to-end ASR systems often suffer from the problem in which redundant generatons repeat and importance symbols vanish.

In our previous work, we proposed a joint end-to-end and DNN-HMM hybrid ASR system that can leverage the advantages of both ASR systems [14]. In the joint ASR system, a end-to-end ASR system based on an encoder-decoder model was utilized for rescoring hypotheses generated from the DNN-HMM hybrid ASR system. This enables us to utilize the stable ASR processing of the DNN-HMM hybrid ASR system and the estimation ability of the totally optimized model of end-to-end ASR system.

However, in the previous work, DNN-HMM hybrid ASR system and end-to-end ASR system are independently trained although these two ASR systems process the same input speech. The part of the DNN acoustic model is regarded as a network specialized in frame-level phoneme estimation. We expect that the network in the DNN acoustic model output continuous vectors including information to discriminate phonemes so that they are valuable for enhancing the end-to-end ASR system.

In this paper, we propose a novel joint end-to-end and DNN-HMM hybrid ASR system that involves a shared network in the DNN acoustic model. In the proposed system, continuous vectors extracted by the part of DNN acoustic model is utilized as auxiliary features for the end-to-end ASR system. The continuous vectors including knowledge of estimation of phonemes enhance the end-to-end ASR system. As a results, a higher accuracy of ASR can be provided.

We carry out our experiments with the Corpus of Spontaneous Japanese (CSJ) [15]. Experimental results show that the proposed joint systems yield better ASR performance. We verify that the joint system provides better ASR performance compared with that of the DNN-HMM hybrid system and the end-to-end ASR system.

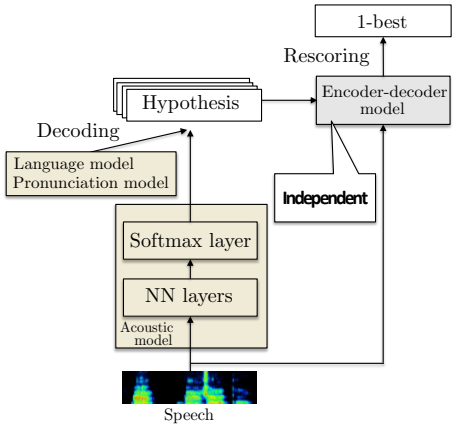
This paper is organized as follows. Section 2 gives the details of a joint end-to-end and DNN-HMM hybrid ASR system and related work. Experiments are shown in Section 3. Section 4 concludes the paper.

2. A Joint End-to-End and DNN-HMM hybrid ASR system

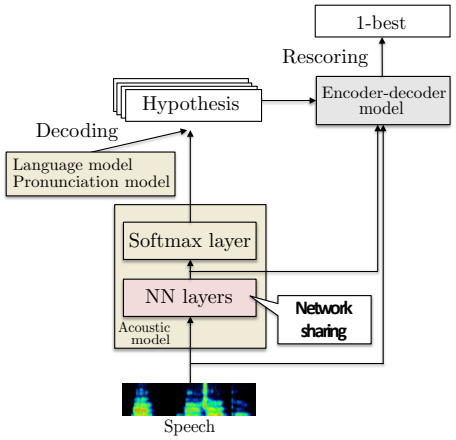
2.1. Modeling

In this section, we explain a joint end-to-end and DNN-HMM hybrid ASR system. (a) in Fig 1 illustrates the ASR procedure of a joint end-to-end and DNN-HMM hybrid ASR system. The ASR score of joint systems is calculated by linearly interpolation with log probabilities calculated from the DNN-HMM hybrid system and end-to-end ASR system. Given the input sequence $\boldsymbol{x} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_J\}$, 1-best ASR result $\hat{\boldsymbol{w}}$ is determined

http://dx.doi.org/10.21437/Interspeech.2019-2263



(a): Joint system without network sharing.



(b): Joint system with network sharing.

Figure 1: *ASR procedure of joint end-to-end and DNN-HMM hybrid ASR systems.*

by

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{NBEST}}{\arg\max} \{\alpha \log P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Lambda}) + (1-\alpha)\log P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Theta})\}, \quad (1)$$

$$= \underset{\boldsymbol{w}\in\mathcal{NBEST}}{\arg\max} \{\alpha \log P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Lambda}) + (1-\alpha)\{\log P(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\theta}_{am})P(\boldsymbol{w};\boldsymbol{\theta}_{lm})\}\}, \quad (2)$$

where $P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Lambda})$ and $P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Theta})$ are the probability calculated from the end-to-end ASR system with their parameters $\boldsymbol{\Lambda}$ and the probability calculated from the DNN-HMM hybrid ASR system with their parameters $\boldsymbol{\Theta}$, α is the interpolation weight of the end-to-end ASR system, and $\mathcal{NBEST}$ denotes N-best list generated from the DNN-HMM hybrid ASR system. $P(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\theta}_{am})$ and $P(\boldsymbol{w};\boldsymbol{\theta}_{lm})$ are calculated from an acoustic model with parameters $\boldsymbol{\theta}_{am}$ and a language model with parameters $\boldsymbol{\theta}_{lm}$.

At first, probability estimation of the DNN acoustic model is described. Given the acoustic feature sequence $\boldsymbol{f} = \{\boldsymbol{f}_1, \cdots, \boldsymbol{f}_J\}$, the input sequence $\boldsymbol{x}$ is concatenated vectors created from multiple frames of acoustic features as

$$\boldsymbol{x}_j = [\boldsymbol{f}_{j-M}, \cdots, \boldsymbol{f}_j, \cdots, \boldsymbol{f}_{j+M}], \quad (3)$$

where M denotes the context size in input. Given the input sequence $\boldsymbol{x}$, internal outputs $\boldsymbol{I} = \{\boldsymbol{I}_1, \cdots, \boldsymbol{I}_J\}$ of the acoustic model are calculated as

$$\boldsymbol{I}_j = \texttt{NN}(\boldsymbol{x}_j; \boldsymbol{\theta}_{nn}), \quad (4)$$

where $\texttt{NN}(\cdot)$ represents the function of a part of the neural network in the acoustic model and $\boldsymbol{\theta}_{nn}$ is the parameters in $\boldsymbol{\theta}_a m$. Then, the DNN acoustic model estimates the probability distribution of HMM states of context-dependent phones $\boldsymbol{p} = \{p_1, \cdots, p_J\}$. The probability distribution $\boldsymbol{o}_j$ of the j-th frame is calculated with a softmax function as

$$\boldsymbol{o}_j = \texttt{SOFTMAX}(\boldsymbol{I}_j; \boldsymbol{\theta}_o), \quad (5)$$

$$P(p_j|\boldsymbol{x}_j; \boldsymbol{\theta}_{am}) = \boldsymbol{o}_{j,p_j}, \quad (6)$$

where $\boldsymbol{\theta}_s$ is the parameters in $\boldsymbol{\theta}_{\boldsymbol{am}}$ and $\boldsymbol{o}_{j,p_j}$ is the probability of the state p_j. The conditional probability of $\boldsymbol{x}_j$ in j-th frame is calculated as

$$P(\boldsymbol{x}_j|p_j; \boldsymbol{\theta}_{am}) = \frac{P(p_j|\boldsymbol{x}_j; \boldsymbol{\theta}_{am})}{P(p_j)} P(\boldsymbol{x}_j), \quad (7)$$

where $P(p_j)$ is unigram probability of phoneme state and $P(\boldsymbol{x}_j)$ is generative probability of the input. $P(p_j)$ is the count of training data and $P(\boldsymbol{x}_j)$ can be regared as a constant value. The conditinal probability $P(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\theta}_{am})$ in Eq. (3) is calsulaated as

$$P(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\theta}_{am}) = \sum_{\boldsymbol{p}} (\prod_j P(\boldsymbol{x}_j|p_j;\boldsymbol{\theta}_{am})) P(\boldsymbol{p}|\boldsymbol{w}), \quad (8)$$

where $P(\boldsymbol{p}|\boldsymbol{w})$ is the conditional probability of the phoneme states. $P(\boldsymbol{p}|\boldsymbol{w})$ is modeled by context-dependent phones and pronunciation model which decides relationship between phonemes and words. Finally, N-best lists in Eq. (3) are obtained by the decoding based on weighted finite-state transducer (WFST).

On the other hand, the end-to-end ASR system based on encoder-decoder model estimates the conditional probability $P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Lambda})$ in Eq. (3) as

$$P(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\Lambda}) = \prod_{i=1}^{\boldsymbol{I}} P(w_i|w_{i-1}, \boldsymbol{s}_{i-1}, \boldsymbol{v}_i; \boldsymbol{\Lambda}), \quad (9)$$

where $\boldsymbol{\Lambda}$ represents the model parameters. The end-to-end ASR system utilizes acoustic feature sequence $\boldsymbol{f}$ and the internal output of the part of acoustic model $\boldsymbol{I}$ as the input features:

$$\boldsymbol{x}'_j = \begin{cases} \boldsymbol{f}_j, \\ [\boldsymbol{f}_j, \boldsymbol{I}_j], \\ \boldsymbol{I}_j. \end{cases} \quad (10)$$

When $\boldsymbol{f}_j$ is used alone, the joint system corresponds to our previous work [14]. The acoustic feature is input to the encoder on the basis of bi-directional LSTM as

$$\overrightarrow{\boldsymbol{h}}_j = \overrightarrow{\texttt{LSTM}}(\boldsymbol{x}'_j, \overrightarrow{\boldsymbol{h}}_{j-1}, \boldsymbol{\lambda}_{lf}), \quad (11)$$

$$\overleftarrow{\boldsymbol{h}}_j = \overleftarrow{\texttt{LSTM}}(\boldsymbol{x}'_j, \overleftarrow{\boldsymbol{h}}_{j+1}, \boldsymbol{\lambda}_{lb}), \quad (12)$$

where $\overrightarrow{\texttt{LSTM}}(\cdot)$ and $\overleftarrow{\texttt{LSTM}}(\cdot)$ represent LSTM functions of forward and backward LSTM. $\boldsymbol{\lambda}_l f$ and $\boldsymbol{\lambda}_l b$ are the trainable model parameters. The encoder hidden state $\boldsymbol{h}_i$ is calculated by concatenating $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ as

$$\boldsymbol{h_i} = [\overrightarrow{\boldsymbol{h}}_i^\top, \overleftarrow{\boldsymbol{h}}_i^\top]^\top. \quad (13)$$

The context vector $\boldsymbol{u}_i$ is constructed in each time step when estimating generative word probabilities in the decoder as

$$\boldsymbol{v}_i = \sum_{j=1}^{J} \alpha_{j,i} \boldsymbol{h}_j, \tag{14}$$

where $\alpha_{j,i}$ is calculated as

$$\alpha_{j,i} = \frac{\exp(\boldsymbol{e_{j,i}})}{\sum_{j=1}^{J} \exp(\boldsymbol{e}_{j,i})}, \tag{15}$$

where $e_{j,i}$ is calculated with the previous α and matrix $\boldsymbol{G}$ as

$$\boldsymbol{g}_j = \boldsymbol{G} * \boldsymbol{\alpha}_{j-1}, \tag{16}$$

$$\boldsymbol{e}_{j,i} = \texttt{tanh}(\boldsymbol{s}_i, \boldsymbol{h}_j, \boldsymbol{g}_{j,i}, \boldsymbol{\lambda}_e), \tag{17}$$

where $\boldsymbol{s}_i$ is the hidden state in the decoder, "*" indicates the convolutional function and $\boldsymbol{G}$ and $\boldsymbol{\lambda}_e$ are the trainable model parameters. In the decoder, the distributed representation $\boldsymbol{d}_{i-1}$ is calculated by the weight matrix as

$$\boldsymbol{d}_{i-1} = \texttt{EMBED}(w_{i-1}, \boldsymbol{\lambda}_d). \tag{18}$$

The hidden state in the decoder is calculated by LSTM function as

$$\boldsymbol{s}_i = \texttt{LSTM}([\boldsymbol{d}_{i-1}, \boldsymbol{v}_{i-1}], \boldsymbol{s}_{i-1}, \boldsymbol{\lambda}_s). \tag{19}$$

Then, $\boldsymbol{o}_j$ is calculated by concatenating the decoder hidden state with a context vector and the hyperbolic tangent function as

$$\boldsymbol{u}_i = \texttt{tanh}([\boldsymbol{s}_i, \boldsymbol{v}_i]^\top, \boldsymbol{\lambda}_u), \tag{20}$$

where $\boldsymbol{s}_i$ is the hidden state in the decoder. Finally, the decoder estimates the probability distribution in the target hypothesis with a conditional probability as

$$\boldsymbol{O}_i = \texttt{SOFTMAX}(\boldsymbol{u}_i, \boldsymbol{\lambda}_o), \tag{21}$$

$$P(w_i | w_{i-1}, \boldsymbol{s}_{i-1}, \boldsymbol{v}_i; \boldsymbol{\Lambda}) = \boldsymbol{O}_{i,w_i}, \tag{22}$$

where $\boldsymbol{\lambda}_o$ is the trainable model parameter and $\boldsymbol{O}_{i,w_i}$ denotes the conditional probaility of symbol w_i.

The trainable parameters of the neural network in the joint end-to-end and DNN-HMM hybrid ASR system are optimized separately. At first, the parameters of the DNN acoustic model in the DNN-HMM hybrid system is updated. Given training data set $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{p}_1), \cdots, (\boldsymbol{x}_N, \boldsymbol{p}_N)\}$, the trainable parameters in the DNN acoustic model $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{nn}, \boldsymbol{\theta}_o\}$ are upadated to minimize cross entropy between estimated probabilities and their references as

$$\mathcal{L}_{AM} = -\sum_{\mathcal{D}} \sum_{j=1}^{|\boldsymbol{x}|} \sum_{k} \hat{\boldsymbol{o}}_{j,k} \log \boldsymbol{o}_{j,k}, \tag{23}$$

where $\hat{\boldsymbol{o}}_{j,k}$ and $\boldsymbol{o}_{j,k}$ are the reference probability and estimated probability of state of the phoneme k at the j-th frame, respectively. Then, the parameters in the end-to-end ASR system are updated. Given training data set $\mathcal{D} = \{(\boldsymbol{x}'_1, \boldsymbol{w}_1), \cdots, (\boldsymbol{x}'_N, \boldsymbol{w}_N)\}$, the trainable model parameters $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_{lf}, \boldsymbol{\lambda}_{lb}, \boldsymbol{F}, \boldsymbol{\lambda}_e, \boldsymbol{\lambda}_d, \boldsymbol{\lambda}_s, \boldsymbol{\lambda}_t, \boldsymbol{\lambda}_o\}$ are optimized to minimize cross entropy loss between estimated probabilities and their references as

$$\mathcal{L}_{E2E} = -\sum_{\mathcal{D}} \sum_{i=1}^{|\boldsymbol{w}|} \sum_{m} \hat{\boldsymbol{O}}_{i,m} \log \boldsymbol{O}_{i,m}, \tag{24}$$

where $\hat{\boldsymbol{O}}_{i,m}$ and $\boldsymbol{O}_{i,m}$ are the reference probability and the estimated probability of the i-th symbol m.

Table 1: *Details of data for training, development and evaluation*

Data	# of characters	# of words	Hours
Training	12,573,004	7,798,998	644.84
Development	44,915	28,424	2.42
Evaluation	74,779	45,889	4.00

2.2. Related work

Our proposed method is closely related to neural network-based language models (NNLMs). In ASR, NNLMs are generally used for rescoring hypotheses or lattice generated from DNN-HMM hybrid ASR systems. Among the NNLMs, recurrent neural network-based language models (RNNLMs) [16, 17] and long short-term memory recurrent neural network-based langauge model (LSTM-RNNLMs) [18] have been shown to improve ASR performance significantly. RNNLMs can efficiently capture long-term dependencies of words by embedding long-term contexts into hidden representations. Our proposed systems utilize not only hidden representation of the context but also the acoustic features and internal outputs from acoustic models in the DNN-HMM hybrid ASR system.

In our proposed systems, encoder-decoder models utilize internal outputs of DNN acoustic models. It is known that bottleneck features are one of the outputs from internal layers of DNNs. Bottleneck features have been widely used in several tasks such as ASR [19, 20]. The internal layer constructs a compressed continuous representation of the task-related information. In the case of ASR, the bottleneck features are created from a DNN trained to predict phoneme states. Kombrink et al. [20] used bottleneck features for the input feature of Gaussian mixture model-HMM-based ASR systems. Our proposed systems introduce internal outputs of DNN acoustic models into end-to-end ASR systems.

3. Experiment

3.1. Setups

All experiments were carried out on CSJ [15], which is a Japanese lecture corpus. Table 1 shows the details of the data for training, development and evaluation.

We prepared an acoustic-to-character based encoder-decoder model for end-to-end ASR system. The end-to-end ASR systems had bi-directional LSTM with 4 hidden layers and 320 units in each layer and direction in the encoder and uni-directional LSTM with 1 hidden layer and 320 LSTM units in the decoder. The vocabulary size was 3251 symbols corresponding to the dimensions of the output target. The beam size was set to 20 for the beam search decoding and the candidates hypotheses were re-ranked on the basis of the length normalized scores [21]. The End-to-end system was trained with 40 mel-scale filter-bank features, delta and delta-delta features (120-dim). We used AdaDelta algorithm [22] to optimize the model parameters and early stopping was conducted by development set accuracy.

We used a convolutional neural network (CNN)-LSTM acoustic model in the DNN-HMM hybrid ASR system. In the CNN-LSTM acoustic model, each static and dynamic component was sliced within 11 frames, which was composed as 3 feature maps. We used 1 convolutional layer with 128 features maps in which 5×11 frequency-time filters. For pooling, 2×1 frequency-time max pooling was performed. In addition, the CNN output was fed into 2 LSTM layers, each of which had

Table 2: *Character error rates (%) on evaluation set in different systems and input fetures for end-to-end ASR systems. "AM" represents the acoustic model in the DNN-HMM hybrid ASR system.*

	System	Input features for end-to-end	%CER
(1)	End-to-end	FBANK	12.95
(2)	End-to-end	Internal output of AM	13.25
(3)	End-to-end	FBANK + Internal output of AM	12.45
(4)	DNN-HMM	-	10.85
(5)	DNN-HMM + LSTM-RNNLM	-	10.17
(6)	Joint (end-to-end + DNN-HMM)	FBANK	9.28
(7)	Joint (end-to-end + DNN-HMM)	Internal output of AM	9.35
(8)	Joint (end-to-end + DNN-HMM)	FBANK + Internal output of AM	9.18

1024 cells. The LSTM output was fed into a softmax layer. We prepared a 3-gram language model. The DNN-HMM hybrid ASR system included a WFST based decoder [23].

In the joint DNN-HMM hybrid ASR systems, the hyper-parameters of end-to-end and DNN-HMM hybrid systems are the same as the systems mentioned above. The internal outputs of LSTM in the acoustic model (1024-dim) are used as the input features of the encoder-decoder network. We used 1144-dimensional vectors when both FBANK (120-dim) and the internal outputs of the acoustic model (1024-dim) are used as auxiliary input features for the end-to-end ASR system.

We prepared a word-based LSTM-RNNLM as a rescoring model for comparison. The LSTM-RNNLM had 2 hidden layers and 520 LSTM units in each layer. The vocabulary size was 67780 words corresponding to the dimensions of the input and output. The dropout ratio was set to 0.3 in each hidden layer.

We used the 100-best list generated from each utterance for rescoring. The interpolation weights in rescoring were changed from 0 to 1 in the 0.1 step, and the best one is selected by character error rate (CER) on the development set. In the case of using LSTM-RNNLM, the score was interpolated with the 3-gram language model score.

3.2. Results

3.2.1. Character Error Rate Evaluation

Table 2 shows the results in terms of CER when using different systems and input features for end-to-end ASR systems. In lines (1)–(3), the end-to-end ASR system trained from FBANK and the internal outputs was superior to the end-to-end ASR system trained from one of the two features. This results indicate that the internal outputs of acoustic models help the end-to-end ASR system to improve ASR performance. On the other hand, end-to-end ASR system trained from internal outputs alone showed the worst CER. It is assumed that the neural network in the acoustic model lost necessary information for generation in end-to-end ASR systems. In lines (1), (4) and (6)–(8), the joint systems outperformed the end-to-end ASR system and DNN-HMM system. This result suggests that the end-to-end and DNN-hybrid ASR system have different characteristics derived from the target label in training. In lines (5)–(8), our proposed joint systems outperformed LSTM-RNNLM trained from lexical features even when the internal outputs of the acoustic model were used alone. This indicates that acoustic information and internal outputs are effective for the prediction of symbol probability. In lines (6)–(8), by the combination of internal outputs of the acoustic model and acoustic features, we obtained a 0.10-points CER reduction from the joint system trained from FBANK alone. This is because the joint system can efficiently leverage the information extracted from acoustic models.

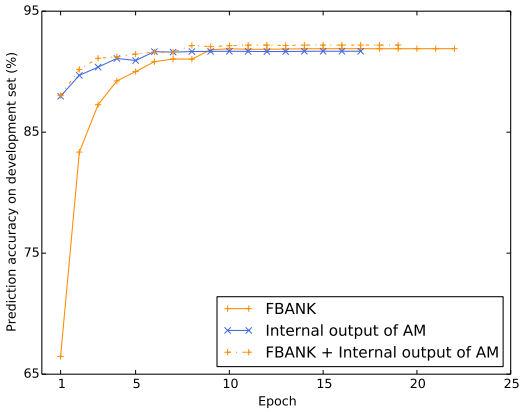


Figure 2: *Prediction accuracy on development set during training in different inputs of end-to-end ASR systems. The accuracies were calculated with giving previous correct characters.*

3.2.2. Prediction Accuracies during Training of Encoder-Decoder Models

Figure 2 shows prediction accuracies on the development set during training of end-to-end ASR systems. The accuracies were calculated with giving previous correct characters. When the internal outputs of the acoustic model were utilized for the input features, the training of the end-to-end ASR systems started with higher accuracies than those of the model trained from only FBANK. This is because the internal outputs include the information necessary to predict phones since the acoustic model was optimized to maximize the prediction accuracy. It is assumed that the information helps the end-to-end ASR systems to predict characters accurately. The best accuracy was obtained when both FBANK and the internal outputs are used for the input features.

4. Conclusions and future work

This paper presented a joint end-to-end and DNN-HMM hybrid ASR system with a shared network in a DNN acoustic model in the DNN-HMM hybrid ASR system. In the joint system, the end-to-end ASR system can leverage the information from the acoustic model through a shared network. The shared network introduces information to predict phones into the end-to-end ASR system. Experiments were carried out on a Japanese lecture corpus by an ASR task. The joint system achieved better ASR performance than conventional methods without sharing networks. The results showed the shared network helped the end-to-end ASR system to predict characters more accurately. Future work includes joint training of the DNN-HMM hybrid ASR system and the end-to-end ASR system to improve their respective ASR performance.

5. References

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4277–4280, 2012.

[2] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8614–8618, 2013.

[3] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, 2013.

[4] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 273–278, 2013.

[5] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," *In Proc. International Conference on Machine Learning, (ICML)*, pp. 1764–1772, 2014.

[6] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv: 1412.5567*, 2014.

[7] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: end-to-end speech recognition using deep RNN models and wfst-based decoding," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174, 2015.

[8] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, and M. Picheny, "Building competitive direct acoustics-to-word models for english conversational speech recognition," *In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4759–4763, 2018.

[9] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *In Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 577–585, 2015.

[10] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949, 2016.

[11] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *In Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 7–11, 2018.

[12] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," *In Proc of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, 2013.

[13] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," *In Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 193–199, 2017.

[14] T. Tanaka, R. Masumura, T. Moriya, and Y. Aono, "Neural speech-to-text language models for rescoring hypotheses of DNN-HMM hybrid automatic speech recognition systems," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, pp. 196–200, 2018.

[15] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," *In Proc. ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium*, pp. 244–248, 2000.

[16] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1045–1048, 2010.

[17] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition," *In proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2877–2880, 2011.

[18] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 194–197, 2012.

[19] F. Grézl, M. Karafiát, S. Kontar, and J. Cernocký, "Probabilistic and bottle-neck features for LVCSR of meetings," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 757–760, 2007.

[20] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," *In Proc. International Speech Communication Association (INTERSPEECH)*, pp. 237–240, 2011.

[21] F. Cromierès, C. Chu, T. Nakazawa, and S. Kurohashi, "Kyoto university participation to WAT," *In Proc. the 3rd Workshop on Asian Translation (WAT)*, pp. 166–174, 2016.

[22] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv:1212.5701*, 2012.

[23] 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 Trans. Audio, Speech & Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.