Relevant Phonetic-aware Neural Acoustic Models using Native English and Japanese Speech for Japanese-English Automatic Speech Recognition

Ryo Masumura*, Suguru Kabashima†, Takafumi Moriya*,
Satoshi Kobashikawa*, Yoshikazu Yamaguchi* and Yushi Aono*
* NTT Media Intelligence Laboratories, NTT Corporation, Japan
† The University of Tokyo, Japan
E-mail: ryou.masumura.ba@hco.ntt.co.jp

***Abstract*—This paper proposes relevant phonetic-aware neural acoustic models that leverage native Japanese speech and native English speech to create improved automatic speech recognition (ASR) of Japanese-English speech. In order to accurately transcribe Japanese-English speech in ASR, acoustic models are needed that are specific to Japanese-English speech since Japanese-English speech exhibits pronunciations that differ from those of native English speech. The major problem is that it is difficult to collect a lot of Japanese-English speech for constructing acoustic models. Therefore, our motivation is to efficiently leverage the significant amounts of native English and native Japanese speech material available since Japanese-English is definitely affected by both native English and native Japanese. Our idea is to utilize them indirectly to enhance the phonetic-awareness of Japanese-English acoustic models. It can be expected that the native English speech is effective in enhancing the classification performance of English-like phonemes, while the native Japanese speech is effective in enhancing the classification performance of Japanese-like phonemes. In the proposed relevant phonetic-aware neural acoustic models, this idea is implemented by utilizing bottleneck features of native English and native Japanese neural acoustic models. Our experiments construct the relevant phonetic-aware neural acoustic models by utilizing 300 hours of Japanese-English speech, 1,500 hours of native Japanese speech, and 900 hours of native English speech. We demonstrate effectiveness of our proposal using evaluation data sets that involve four levels of Japanese-English.**

I. Introduction

The progress of globalization has increased the need to use English as an official language for non-native speakers. For example, in international meetings, foreigners communicate each other through English. One problem is that foreigners often cannot understand the English of someone from another country. In order to support full communication, we need automatic speech recognition (ASR) that can transcribe English speech into text. This paper focuses on improving the ASR performance for the input of Japanese-English speech.

Recent ASR technologies have been dramatically improved by deep learning technologies. In particular, neural acoustic models have attained significant performance superiority compared to Gaussian mixture model based methods [1], [2]. It is reported that neural acoustic models offer substantial ASR performance for practical systems if a lot of target domain training data is available.

In order to accurately transcribe non-native speech, including Japanese-English speech, ASR systems must be specialized to handle the target non-native speech domain since the pronunciations of non-native speech differ from those of native speech [3]–[6]. The main problem is that it is difficult to collect adequate amounts of target non-native speech. In particular, Japanese-English speech exhibits a wide range in pronunciation style, from beginner level to professional level. In fact, neural acoustic models trained using native English speech are completely unsuitable for improving the ASR performance of Japanese-English speech.

Our idea is to promote the phonetic-awareness of neural acoustic models by indirectly utilizing native Japanese speech and native English speech. It can be considered that phonemes appearing in the native Japanese speech are relevant to those that appear in beginner level Japanese-English. In the same way, it can be considered that phonemes appearing in the native English speech will be relevant to those found in fluent Japanese-English. Therefore, we can expect that the native English speech is effective in enhancing the classification performance of English like phonemes, and the native Japanese speech is effective in enhancing the classification performance of Japanese like phonemes. Utilizing relevant phonetic-awareness allows well-trained acoustic models to be constructed from limited Japanese-English speech material.

In this paper, we propose relevant phonetic-aware neural acoustic models for Japanese-English ASR. The proposal leverages neural acoustic models that are individually constructed from native Japanese speech and native English speech. The relevant phonetic-aware neural acoustic models have two components; awareness extraction networks and a classification network. In the awareness extraction networks, the native Japanese and the native English neural acoustic models are used in converting input acoustic features into relevant phonetic-aware features. The classification network handles the relevant phonetic-aware features to estimate phonemes.

The proposed method is closely related to neural acoustic models with auxiliary features. In ASR, major auxiliary fea-

tures have been speaker-awareness [7]–[9], noise-awareness [10]–[12], reverberant-awareness [13], and distance-awareness [14]. Phonetic-awareness is often used in speaker recognition [15], spoken language identification [16], and voice activity detection [17]. Different from the previous work, the proposal utilizes multiple relevant phonetic-aware features extracted from neural acoustic models for enhancing other neural acoustic models. In addition, the proposal is related to multilingual neural acoustic models [18]–[21]. While multilingual neural acoustic models jointly employ multilingual speech for constructing acoustic models, the proposal utilizes pre-trained neural acoustic models for constructing target domain acoustic models. This enables us to directly leverage relevant domain knowledge.

For evaluation purposes, we create Japanese-English evaluation data sets that involve four levels of Japanese-English. Tests that utilize 300 hours of Japanese-English speech, 1,500 hours of native Japanese speech, and 900 hours of native English speech for constructing acoustic models demonstrate that the proposal offers improved Japanese-English ASR performance. To the best of our knowledge, this paper is the first study on Japanese-English acoustic modeling that effectively utilizes both native Japanese data and native English data.

This paper is organized as follows. Section 2 of this paper describe neural acoustic models. The proposal is detailed in Section 3. Section 4 describes our experiments. We conclude in Section 5 with a brief summary.

II. Neural Acoustic Models

This section describes neural acoustic models; they are often used in deep neural network hidden Markov model (DNN-HMM) hybrid ASR systems. In neural acoustic models, phonetic state sequence $\boldsymbol{S} = \{s_1, \cdots, s_T\}$ is estimated from input acoustic feature sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ in a frame-by-frame manner. Phonetic state $s_t \in \{1, \cdots, K\}$ represents context-dependent phones, each of which corresponds to a HMM state. The neural acoustic models define conditional probability $P(\boldsymbol{S}|\boldsymbol{X}, \boldsymbol{\Theta})$ where $\boldsymbol{\Theta}$ is the model parameter.

In neural acoustic models, frame-level input features are often composed by stacking a currently-being-processed frame and its left-right contexts. Predictive probabilities of phonetic states in the t-th frame, $\boldsymbol{o}_t$, are given by:

$$\boldsymbol{o}_t = \mathcal{F}(\boldsymbol{i}_t; \boldsymbol{\Theta}), \tag{1}$$

$$\boldsymbol{i}_t = [\boldsymbol{x}_{t-M}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+M}^\top]^\top, \tag{2}$$

where M denotes context size in the input layer, and $\mathcal{F}()$ is a nonlinear transformational function based on DNNs or unidirectional long short-term memory recurrent neural networks (LSTM-RNNs). Unidirectional LSTM-RNNs can automatically store previous long-range information in hidden layers.

Given training data sets $\mathcal{D} = \{(\boldsymbol{x}_1, s_1), \cdots, (\boldsymbol{x}_{|\mathcal{D}|}, s_{|\mathcal{D}|})\}$ which are forcibly aligned in a preliminary step, the parameter can be optimized by minimizing the cross entropy between reference and estimated probabilities:

$$\hat{\boldsymbol{\Theta}} = \arg\min - \sum_{t=1}^{|\mathcal{D}|} \sum_{k=1}^{K} \hat{\boldsymbol{o}}_{t,k} \log \boldsymbol{o}_{t,k}, \tag{3}$$

where $\hat{\boldsymbol{o}}_{t,k}$ and $\boldsymbol{o}_{t,k}$ are the reference probability and the estimated probability of the k-th phonetic state in the t-th frame, respectively.

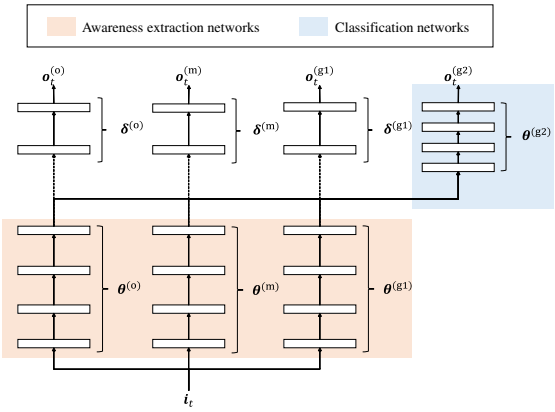

Fig. 1. Network structure of relevant phonetic-aware neural acoustic models.

III. Relevant Phonetic-aware Neural Acoustic Models

This section details relevant phonetic-aware neural acoustic models. The proposal indirectly utilizes neural acoustic models that are individually constructed from native Japanese speech and native English speech. The relevant phonetic-aware neural acoustic models consist of two components; multiple awareness extraction networks that can extract relevant phonetic-aware features from acoustic features, and a classification network that predicts posterior probability of each phonetic state from the relevant phonetic-aware features.

We define the language uttered by a speaker and the mother language of the speaker as $\mathcal{L}^{(\mathtt{o})}$ and $\mathcal{L}^{(\mathtt{m})}$, respectively. In ASR of Japanese-English speech, $\mathcal{L}^{(\mathtt{o})}$ means English, and $\mathcal{L}^{(\mathtt{m})}$ means Japanese. In addition, we define the following three speech data sets.

- $\mathcal{D}^{(\mathtt{o})}$: speech of $\mathcal{L}^{(\mathtt{o})}$ uttered by native speakers of $\mathcal{L}^{(\mathtt{o})}$.
- $\mathcal{D}^{(\mathtt{m})}$: speech of $\mathcal{L}^{(\mathtt{m})}$ uttered by native speakers of $\mathcal{L}^{(\mathtt{m})}$.
- $\mathcal{D}^{(\mathtt{g})}$: speech of $\mathcal{L}^{(\mathtt{o})}$ uttered by native speakers of $\mathcal{L}^{(\mathtt{m})}$.

In ASR of Japanese-English speech, $\mathcal{D}^{(\mathtt{o})}$ is native English speech, $\mathcal{D}^{(\mathtt{m})}$ is native Japanese speech, and $\mathcal{D}^{(\mathtt{g})}$ is Japanese-English speech.

Fig. 1 shows the model structure of relevant phonetic-aware neural acoustic models. Model parameter $\boldsymbol{\Theta}$ corresponds to $\{\boldsymbol{\theta}^{(\mathtt{o})}, \boldsymbol{\theta}^{(\mathtt{m})}, \boldsymbol{\theta}^{(\mathtt{g1})}, \boldsymbol{\theta}^{(\mathtt{g2})}\}$. The networks are detailed in the following subsections.

A. Awareness Extraction Networks

Awareness extraction networks are individually constructed from $\mathcal{D}^{(\mathtt{o})}$, $\mathcal{D}^{(\mathtt{m})}$, and $\mathcal{D}^{(\mathtt{g})}$. To this end, we compose DNNs

with a bottleneck layer which is an output layer of the awareness extraction network. For the DNNs, a frame-level input feature is composed by stacking the frame currently being processed and its left-right contexts according to Eq. (2). Posterior probabilities of the phonetic states in the t-th frame are given by:

$$\boldsymbol{o}_t^{(\mathtt{o})} = \mathcal{G}(\mathcal{H}(\boldsymbol{i}_t; \boldsymbol{\theta}^{(\mathtt{o})}); \boldsymbol{\delta}^{(\mathtt{o})}) \in \mathbb{R}^{K^{(\mathtt{o})}}, \tag{4}$$

$$\boldsymbol{o}_t^{(\mathtt{m})} = \mathcal{G}(\mathcal{H}(\boldsymbol{i}_t; \boldsymbol{\theta}^{(\mathtt{m})}); \boldsymbol{\delta}^{(\mathtt{m})}) \in \mathbb{R}^{K^{(\mathtt{m})}}, \tag{5}$$

$$\boldsymbol{o}_t^{(\mathtt{g1})} = \mathcal{G}(\mathcal{H}(\boldsymbol{i}_t; \boldsymbol{\theta}^{(\mathtt{g1})}); \boldsymbol{\delta}^{(\mathtt{g1})}) \in \mathbb{R}^{K^{(\mathtt{g})}}, \tag{6}$$

where $\mathcal{H}()$ represents non-linear functions of the awareness extraction network whose outputs correspond to relevant phonetic-aware features. $\mathcal{G}()$ represents non-linear functions of an additional network connected to the awareness extraction network that is used only for training. $\boldsymbol{\theta}^{(\mathtt{o})}$, $\boldsymbol{\theta}^{(\mathtt{m})}$, and $\boldsymbol{\theta}^{(\mathtt{g1})}$ are the model parameters of the awareness extraction network, and $\boldsymbol{\delta}^{(\mathtt{o})}$, $\boldsymbol{\delta}^{(\mathtt{m})}$, and $\boldsymbol{\delta}^{(\mathtt{g1})}$ are the model parameters of the additional network. $K^{(\mathtt{o})}$, $K^{(\mathtt{m})}$, and $K^{(\mathtt{g})}$ are the number of phonetic states in $\mathcal{D}^{(\mathtt{o})}$, $\mathcal{D}^{(\mathtt{m})}$, and $\mathcal{D}^{(\mathtt{g})}$, respectively. Each model parameter can be trained by optimizing the following cross entropy loss function:

$$\hat{\boldsymbol{\theta}}^{(\mathtt{o})}, \hat{\boldsymbol{\delta}}^{(\mathtt{o})} = \arg\min_{\boldsymbol{\theta}^{(\mathtt{o})}, \boldsymbol{\delta}^{(\mathtt{o})}} - \sum_{t=1}^{|\mathcal{D}^{(\mathtt{o})}|} \sum_{k=1}^{K^{(\mathtt{o})}} \hat{\boldsymbol{o}}_{t,k}^{(\mathtt{o})} \log \boldsymbol{o}_{t,k}^{(\mathtt{o})}, \tag{7}$$

$$\hat{\boldsymbol{\theta}}^{(\mathtt{m})}, \hat{\boldsymbol{\delta}}^{(\mathtt{m})} = \arg\min_{\boldsymbol{\theta}^{(\mathtt{m})}, \boldsymbol{\delta}^{(\mathtt{m})}} - \sum_{t=1}^{|\mathcal{D}^{(\mathtt{m})}|} \sum_{k=1}^{K^{(\mathtt{m})}} \hat{\boldsymbol{o}}_{t,k}^{(\mathtt{m})} \log \boldsymbol{o}_{t,k}^{(\mathtt{m})}, \tag{8}$$

$$\hat{\boldsymbol{\theta}}^{(\mathtt{g1})}, \hat{\boldsymbol{\delta}}^{(\mathtt{g1})} = \arg\min_{\boldsymbol{\theta}^{(\mathtt{g1})}, \boldsymbol{\delta}^{(\mathtt{g1})}} - \sum_{t=1}^{|\mathcal{D}^{(\mathtt{g})}|} \sum_{k=1}^{K^{(\mathtt{g})}} \hat{\boldsymbol{o}}_{t,k}^{(\mathtt{g1})} \log \boldsymbol{o}_{t,k}^{(\mathtt{g1})}, \tag{9}$$

where $\hat{\boldsymbol{o}}_{t,k}^{(\mathtt{o})}$, $\hat{\boldsymbol{o}}_{t,k}^{(\mathtt{m})}$, and $\hat{\boldsymbol{o}}_{t,k}^{(\mathtt{g1})}$ are the reference probabilities, and $\boldsymbol{o}_{t,k}^{(\mathtt{o})}$, $\boldsymbol{o}_{t,k}^{(\mathtt{m})}$, and $\boldsymbol{o}_{t,k}^{(\mathtt{g1})}$ are the estimated probabilities of the k-th phonetic state in the t-th frame, respectively. After optimization, $\hat{\boldsymbol{\theta}}_{t,k}^{(\mathtt{o})}$, $\hat{\boldsymbol{\theta}}_{t,k}^{(\mathtt{m})}$, and $\hat{\boldsymbol{\theta}}_{t,k}^{(\mathtt{g1})}$ are only utilized for constructing the classification network.

B. Classification Network

The classification network takes, as input, multiple relevant phonetic-aware features extracted from the awareness extraction networks. The t-th input of the classification network, $\boldsymbol{u}_t$, is defined as:

$$\boldsymbol{u}_t = [\mathcal{H}(\boldsymbol{i}_t; \hat{\boldsymbol{\theta}}^{(\mathtt{o})})^\top, \mathcal{H}(\boldsymbol{i}_t; \hat{\boldsymbol{\theta}}^{(\mathtt{m})})^\top, \mathcal{H}(\boldsymbol{i}_t; \hat{\boldsymbol{\theta}}^{(\mathtt{g1})})^\top]^\top. \tag{10}$$

The posterior probabilities of the phonetic states in the t-th frame, $\boldsymbol{o}_t^{(\mathtt{g2})}$, are given by:

$$\boldsymbol{o}_t^{(\mathtt{g2})} = \mathcal{F}(\boldsymbol{u}_t; \boldsymbol{\theta}^{(\mathtt{g2})}) \in \mathbb{R}^{K^{(\mathtt{g})}}, \tag{11}$$

where $\boldsymbol{\theta}^{(\mathtt{g2})}$ is the model parameter of the classification network. $\mathcal{F}()$ is a nonlinear transformational function based on DNNs or unidirectional LSTM-RNNs. In order to optimize the model parameter of the classification network, only $\mathcal{D}^{(\mathtt{g})}$ is necessary. The model parameter can be trained by optimizing the following cross entropy loss function:

$$\hat{\boldsymbol{\theta}}^{(\mathtt{g2})} = \arg\min_{\boldsymbol{\theta}^{(\mathtt{g2})}} - \sum_{t=1}^{|\mathcal{D}^{(\mathtt{g})}|} \sum_{k=1}^{K^{(\mathtt{g})}} \boldsymbol{o}_{t,k}^{(\mathtt{g2})} \log \boldsymbol{o}_{t,k}^{(\mathtt{g2})}, \tag{12}$$

where $\hat{\boldsymbol{o}}_{t,k}^{(\mathtt{g2})}$ is the reference probability, and $\boldsymbol{o}_{t,k}^{(\mathtt{g2})}$ is the estimated probability of the k-th phonetic state in the t-th frame.

TABLE I
EVALUATION DATA SETS.

Proficiency level	Category	# of words	# of speakers
Level A	Beginner English	18,439	8
Level B	Traveler English	62,357	25
Level C	Daily English	49,819	20
Level D	Professional English	17,661	7

TABLE II
TRAINING DATA SETS.

	hours	# of phonetic states
Native English	885.9	2,601
Native Japanese	1,496.9	3,072
Japansese-English	311.5	2,601

IV. EXPERIMENTS

A. Setups

In our experiments, we prepared Japanese-English evaluation data sets that included short read speech segments uttered by 60 speakers. The data sets can be categorized into four proficiency levels; beginner level, traveler level, daily level, professional level. The details are shown in Table 1. In addition, we prepared native English training data sets, native Japanese training data sets, and Japanese-English training data sets for constructing the acoustic models. Each data set included various tasks and various speakers. In order to utilize them for neural acoustic models, the phonetic state sequences were annotated by force alignment using GMM-HMMs which were individually constructed from the training sets. We constructed a Japanese GMM-HMM from the Japanese speech and an English GMM-HMM from the both Japanese-English and native English speech sets. The training data sets are detailed in Table 2.

Our neural acoustic models used 38 dimensional mel-frequency cepstrum coefficients (12MFCC, 12ΔMFCC, 12ΔΔMFCC, Δpower and ΔΔpower) as acoustic features; they were extracted using 20 msec windows shifted by 10 msec. The input features, i.e. Eq. (2), were 418 dimensional acoustic features formed by stacking the current processed frame and its ±5 left-right context. For the evaluation, we constructed following acoustic models.

- **Baseline neural acoustic models**: We constructed DNN and LSTM-RNN acoustic models. **BASE-DNN** had 8 hidden layers with 2,048 sigmoid activation units. **BASE-LSTM** had 2 hidden layers with 2,048 sigmoid activation units and 2 hidden layers with 1,024 LSTM units. We trained them by using Japanese-English speech and native English speech in isolation, and both together.

TABLE III
EXPERIMENTAL RESULTS IN TERMS OF WORD ERROR RATE (%).

	Methods	Training data for awareness extraction networks	Training data for classification network	Proficiency level A	B	C	D
(1).	BASE-DNN	-	Japanese-English	18.7	14.4	14.8	15.6
(2).	BASE-DNN	-	Native English	75.0	53.8	49.6	27.8
(3).	BASE-DNN	-	Japanese-English, Native English	21.9	16.4	17.5	15.2
(4).	RPA-DNN	Japanese-English, Native Japanese	Japanese-English	17.3	13.3	13.9	14.3
(5).	RPA-DNN	Japanese-English, Native English	Japanese-English	17.1	13.0	13.8	14.5
(6).	RPA-DNN	Japanese-English, Native Japanese, Native English	Japanese-English	16.6	12.7	13.5	14.0
(7).	BASE-LSTM	-	Japanese-English	17.0	14.8	15.7	16.5
(8).	BASE-LSTM	-	Native English	66.4	48.7	45.7	27.7
(9).	BASE-LSTM	-	Japanese-English, Native English	19.6	16.0	16.3	16.2
(10).	RPA-LSTM	Japanese-English, Native Japanese	Japanese-English	16.2	14.3	15.2	15.4
(11).	RPA-LSTM	Japanese-English, Native English	Japanese-English	16.0	14.0	15.1	15.7
(12).	RPA-LSTM	Japanese-English, Native Japanese, Native English	Japanese-English	15.4	13.7	14.5	15.0
(13).	BASE-DNN+LSTM	-	Japanese-English	15.0	12.8	13.6	14.5
(14).	RPA-DNN+LSTM	Japanese-English, Native Japanese, Native English	Japanese-English	**13.9**	**11.9**	**12.6**	**13.5**

For optimization, we used discriminative pre-training to construct an initial network and then fine-tuned it using mini-batch stochastic gradient descent (MB-SGD). The validation set was used for early stopping. Note that **BASE-DNN+LSTM** is the posterior combination of BASE-DNN and BASE-LSTM.

- **Relevant phonetic-aware neural acoustic models**: For awareness extraction networks, we constructed DNN acoustic models with 5 hidden layers. The fourth hidden layer was a bottleneck layer whose unit size was set to 64, and the other hidden layers had 1,024 sigmoid units. We constructed them from native English, native Japanese, and Japanese-English speech. As the classification network, we examined DNN and LSTM-RNN acoustic models. **RPA-DNN** had 4 hidden layers with 2048 sigmoid units. **RPA-LSTM** had 2 hidden layers with 1,024 LSTM units. In order to train both the awareness extraction networks and the classification network, we used discriminative pre-training to construct the initial networks and then fine-tuned them using MB-SGD. The validation set was used for early stopping. Note that **RPA-DNN+LSTM** is the posterior combination of RPA-DNN and RPA-LSTM.

The baseline neural acoustic models are regarded as relevant phonetic-aware neural acoustic models without awareness extraction networks.

For evaluating ASR performance, we prepared a WFST-based ASR decoder [22]. We constructed 3-gram language models with 1M words taken from English Web texts and transcribed texts of spontaneous speech.

B. Results

Table 3 shows the results in terms of word error rate with respect to proficiency levels. Lines (1)–(3) are BASE-DNN, lines (4)–(6) are RPA-DNN, lines (7)–(9) are BASE-LSTM, and lines (10)–(12) are RPA-LSTM. In addition, line (13) plots the posterior combination results of combining line (1) with line (7), while line (14) shows the posterior combination results of combining line (6) with line (12).

First, in lines (1) and (7), BASE-LSTM trained using Japanese-English was superior to BASE-DNN trained using Japanese-English for high proficiency levels, while inferior for low proficiency levels. This indicates that LSTM-RNNs and DNNs have different strengths for Japanese-English ASR. In lines (1)–(3) and (7)–(9), baseline neural acoustic models trained using Japanese-English outperformed those trained using native English. In fact, the baseline models trained using native English were not suitable for the low proficiency level. In addition, baseline models trained using both Japanese-English and native English speech were inferior to those trained using only native Japanese-English for the low proficiency level. These results confirm that it is difficult to directly utilize native English speech for improving Japanese-English ASR performance.

Next, with regard to lines (4)–(6) and (10)–(12), relevant phonetic-aware acoustic models that indirectly leveraged native English and native Japanese outperformed the baseline models. RPA-DNN outperformed BASE-DNN constructed using Japanese-English, and RPA-LSTM outperformed BASE-LSTM constructed using Japanese-English. In particular, the relevant phonetic-awareness of native Japanese was effective for low proficiency level speakers while that of native English was effective for high proficiency level speakers. In addition, we could achieve useful performance improvements by utilizing both native English speech and native Japanese speech simultaneously. This indicates that both types of relevant phonetic-awareness complement each other. The best results were obtained by RPA-DNN+LSTM which yielded statistically significant performance improvements ($p < 0.05$) compared to BASE-DNN+LSTM for all proficiency levels.

V. CONCLUSIONS

In this paper, relevant phonetic aware neural acoustic models were proposed. Our proposed method achieved to efficiently leverage native Japanese speech and native English speech for improving Japanese-English ASR. ASR Evaluation using four proficiency level speakers showed the proposed method yielded significant performance improvement.

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, and T. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, pp. 82–97, 2012.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dpendent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE transactions on Audio, Speech and Language Processing*, vol. 20, pp. 30–42, 2012.

[3] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 540–543, 2003.

[4] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2, pp. 141–153, 2004.

[5] D. Vergyri, L. Lamel, and J.-L. Gauvaion, "Automatic speech recognition of multiple accented English data," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1652–1655, 2010.

[6] Y. Huang, D. Yu, C. Liu, and Y. Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2977–2981, 2014.

[7] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using I-vector," *In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 55–59, 2013.

[8] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6334–6338, 2014.

[9] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 171–176, 2014.

[10] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7398–7402, 2013.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2670–2674, 2014.

[12] K. H. Lee, W. H. Kang, T. G. Kang, and N. S. Kim, "Integrated DNN-based model adaptation technique for noise-robust speech recognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5245–5249, 2017.

[13] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5014–5018, 2015.

[14] Y. Miao and F. Metze, "Distance-aware DNNs for robust speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 761–765, 2015.

[15] M. McLaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottlececk features for speaker and language rocognition," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5575–5579, 2016.

[16] R. Masumura, T. Asami, H. Masataki, and Y. Aono, "Parallel phonetically aware DNNs and LSTM-RNNs for frame-by-frame discriminative modeling of spoken language identification," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5260–5264, 2017.

[17] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5710–5714, 2016.

[18] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7319–7323, 2013.

[19] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-lingual knowledge transfer using multilingual deep neural network with shared hidden layers," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 7304–7308, 2013.

[20] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 8619–8623, 2013.

[21] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 714–718, 2017.

[22] 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.