Multi-task Learning for Acoustic Modeling Using Articulatory Attributes

Yueh-Ting Lee*, Xuan-Bo Chen*, Hung-Shin Lee†, Jyh-Shing Roger Jang*, Hsin-Min Wang†
* Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
† Institute of Information Science, Academia Sinica, Taiwan
ayueh.lee@mirlab.org, bruce.chen@mirlab.org

***Abstract*—In addition to the phone sequences, articulatory attributes in spoken utterances have demonstrated salient cues for supervised training of acoustic models in automatic speech recognition (ASR). In this paper, a multi-task learning (MTL) scheme for neural network-based acoustic modeling is proposed. It aims to simultaneously minimize the cross-entropy losses of the triphone-states and articulatory attributes, given their corresponding true alignments. Supposing the articulatory information associated with the physical process is not as abstract and composite as the phonetic descriptions, the layer-wise neuron sharing occurs only in the first few layers. Moreover, instead of the fully-connected feed-forward networks (FFNs), the well-known structure of time-delay neural networks (TDNNs) is adopted to efficiently model the long-term contexts of each acoustic input frame. The results of experiments on the MATBN Mandarin Chinese broadcast news corpus show that our proposed framework achieves relative character error rate reductions of 3.3% and 5.7% over the non-MTL TDNN-based system and the MTL-FFN-based system, respectively.**

***Index Terms*—multi-task learning, articulatory attributes, deep neural networks, time-delay neural networks, LVCSR**

I. Introduction

In recent years, automatic speech recognition (ASR) has gradually become an indispensable application in the industry of artificial intelligence. Because of the increasing computational capability and accessibility to big audio data, deep neural networks (DNNs) have become the mainstream architecture for acoustic modeling in large vocabulary continuous speech recognition (LVCSR) instead of the traditional Gaussian mixture model (GMM)-based models [1].

Most speech recognition decoders today are based on phones (or phonemes), which, in other perspectives, are often given excessive legitimacy in the speech processing community, particularly with regard to the assumption that the sequence of acoustic observations can be forcibly aligned with the sequence of phones. These phones are viewed as the basic units of speech, but it is now widely believed that they can be broken down into smaller, essential and fundamental units [2]. Although there is no consensus on what these units are, we will take the most popular view, called the articulatory features. Some recent research works have incorporated articulatory knowledge into the acoustic models to better classify phones or improve the performance of LVCSR systems [3–5].

For example, in [3], a framework called automatic speech attribute transcription (ASAT) was proposed for developing detection-based ASR based on attribute detection and knowledge integration. In ASAT, bottom-up knowledge integration was accomplished through a two-step process: the events or attributes of speech were first detected by a learning machine, such as an artificial neural network (ANN), and then the detected cues were integrated into the ASR system using evidence mergers or lattice rescoring techniques [6, 7]. This knowledge, lurking in a speech utterance or phrase, describes the mouth, lips, and tongue attributes of each phoneme and has been summarized by linguists and phoneticians for decades.

However, with the development of DNNs, the models have been integrated in the hidden Markov model (HMM) framework, and the DNN-HMM ASR systems are superior to detection-based ASR systems. So some studies integrated the articulatory attributes into DNN-HMM systems through multi-task learning (MTL) techniques and treated the articulatory attributes as auxiliary roles. For instance, Zheng *et al.* processed acoustic and phonetic information in both model and feature domains using three different feed-forward networks (FFNs) [8]. In the model domain, attribute classification was used as the secondary task (or subtask) to help improve the performance of phone recognition with an FFN by lifting its discriminative ability on pronunciation. In the feature domain, attribute-based features were extracted from another FFN trained for attribute classification, where triphone-state classification was taken as the subtask. Finally, the attribute-aware features and the acoustic features were combined to train the third MTL-based FNN for acoustic modeling. It is worth noting that in Zheng's work, attribute classification was formulated as multiple independent binary classification problems, one for each attribute.

On the practical side, there is no doubt that Kaldi[1] is one of the most popular open-source toolkits [9], providing industrial engineers and academic researchers working in the speech processing area with the most advanced and regularly updated training recipes and a fair ground for comparison. Based on the topology of HMMs and weighted finite-state transducers (WFSTs), Kaldi also generates high-quality word/phone lattices that are sufficiently efficient for real-time decoding. Therefore, we attempt to build an acoustic model based on the recently popular time-delay neural networks (TDNNs) with the multi-task learning strategy in the `nnet3` setup in Kaldi [10, 11].

[1] http://kaldi-asr.org

TABLE I
ARTICULATORY ATTRIBUTES AND THEIR ASSOCIATED PHONES IN MANDARIN GRAPHEME-PHONEMES IN HANYU PINYIN.

category	attribute	grapheme-phoneme
place	bilabial	`b p m`
	labiodental	`f`
	alveolar	`d t l n`
	dental	`z c s ii`
	retroflex	`zh ch sh r err iii`
	palatal	`j q x a o e er i u v`
	velar	`g k h nn ng`
manner	stop	`b p d t g k`
	fricative	`f s sh r x h`
	affricative	`z zh c ch j q`
	nasal	`m n nn ng`
	lateral	`l`
	n/a	all vowels
backness	back	`o er u`
	central	`a err iii`
	front	`e i v ii nn ng`
	n/a	all consonants
height	high	`i ii iii u v`
	low	`a ng`
	middle high	`o er nn`
	middle low	`e err`
	n/a	all consonants
roundedness	rounded	`o u v ng`
	unrounded	`a er e err i ii iii nn`
	n/a	all consonants

TABLE II
THE PHONEME SEQUENCE AND ITS CORRESPONDING MANNER, PLACE + BACKNESS, PLACE + HEIGHT, AND PLACE + ROUNDEDNESS SEQUENCES, TAKING THE CHINESE PHRASE “我們” (WE) FOR EXAMPLE.

phrase	**我們 (we)**				
phoneme	`u`	`o`	`m`	`er`	`nn`
manner	vowel	vowel	nasal	vowel	vowel
place + backness	back	back	bilabial	back	front
place + height	high	middle high	bilabial	middle high	middle high
place + roundedness	rounded	rounded	bilabial	unrounded	unrounded

In this paper, we formulate acoustic modeling into a multi-task learning problem (abbreviated as MTL-TDNN), where the TDNNs are exploited as the primary structure of the acoustic model, and the classification of the articulatory attributes is the subtask. MTL-DNN based models use back propagation to affect the features extracted from the shared hidden layers by training subtasks. In TDNNs, the higher layers have the ability to learn broader temporal relationships. We connect the substasks to the first few layers because the attributes of a phone do not span a long time. In this way, the layer-wise neuron sharing occurs only in the first few layers, and the articulatory information has a greater impact on the features extracted from the lower layers than the features extracted from the higher layers.

In summary, the highlights of this paper in comparison with other relevant works are threefold.

1) Unlike most MTL implementations for NNs, where all tasks share parameters of all layers except the last one or two layers, in our model, layer-wise neuron sharing occurs only in the first few layers. The experiment results show that, on the one hand, it is necessary to increase the number of shared layers to enhance the regularization strength induced by the articulatory information to help avoid overfitting, and on the other hand, deep layers (more than 3 layers) may not be able to capture the abstract or composite representations of the articulatory features.
2) Instead of the FFNs, a more complex structure, like the TDNNs, is adopted in our framework. We have modified the `nnet3` setup [12] to optimize multiple objectives simultaneously. The modification can be easily introduced into other models, such as the long short-term memory with TDNN (TDNN-LSTM) and TDNN-F [13, 14].
3) Unlike the treatment of articulatory labels in [8], we consider the exclusive and non-exclusive properties between the articulatory attributes. We split the attributes into four blocks (we use blocks instead of groups in this paper because each group corresponds to a block in the output layer) so that the attributes in each block are certainly exclusive. This makes the model discriminative among competing attributes in each block [15, 16].

The remainder of this paper is organized as follows. In Section II, we introduce he articulatory attributes with respect to the Mandarin phoneme set. Then, Section III illustrates the mechanisms of MTL-TDNN, single-task learning (STL)-TDNN, and several variants of MTL-TDNN. In Section IV, we evaluate the baseline MTL-DNN model and our framework on the phoneme recognition and LVCSR tasks. Finally, we provide concluding remarks and future work in Section V.

II. ARTICULATORY ATTRIBUTES

Globalization brings the need for second language learning in recent years. The articulatory attributes from an articulatory model are considered as a good feedback to non-native language learners in computer-assisted pronunciation training (CAPT) [17, 18]. Articulatory models can be categorized into geometrical [19–22] and biomechanical [23, 24] types. In this paper, we focus on the geometrical model. In a geometrical model, the vocal tract is represented by its initial geometry, and a set of parameters estimated from the electromagnetic articulography (EMA) data directly deforms this geometry. It has been proved helpful in many areas, such as speech therapy [25], speech comprehension improvement [26] and pronunciation perceptual training [27].

In this study, we use the phone-based symbols to represent the phoneme set, which is a subset of International Phonetic Alphabet (IPA) and only contains Mandarin phonemes [28]. The attributes of speech can be comprehended by a collection of information from fundamental speech sounds. We use the place and manner attributes of each phone. The place attributes identify the place, location, spot and mouth organs involved in

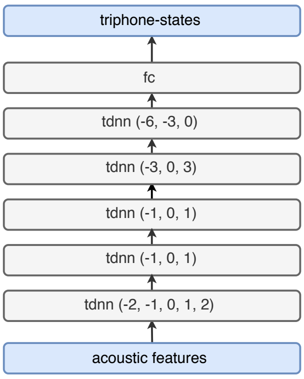

Fig. 1. The architecture of STL-TDNN, where "tdnn $(\cdot)$" denotes the information about splicing indices of the TDNN-based layer, and "fc" denotes the fully-connected layer.

the triggering and production of speech sounds. The manner attributes describe the manner in which these mouth organs trigger or produce speech sounds. The phoneticians sum up the 21 phonological features (attributes), which are listed in Table I. We split these attributes into four blocks so that the attributes in each block are exclusive, that is, only one attribute is labeled as 1 and the other attributes are labeled as 0 in a block. Therefore, we apply the softmax function and the categorical cross-entropy loss function in each block of the subtask output layer in our DNN model instead of the simple multi-label layer in [8].

We prepare four kinds of articulatory transcriptions: namely manner, place + backness, place + height, and place + roundedness, to represent all attributes [15, 16]. The three types of the place attributes are identical if the phone is a consonant. If the phone is a vowel, the attributes only label the place of tongue (backness and height) and the roundedness, because these attributes provide sufficient clues to distinguish vowels. Table II gives an example of the attribute labels mapped from the phone labels.

III. Multi-task Learning

Multi-task learning (MTL) is a machine learning technique that improves single-task learning (STL) by training the model with several related tasks using a shared representation [29]. The effectiveness of MTL depends on the relationship between individual tasks and the shared learning structure across tasks.

One aspect of the effectiveness of subtask learning, which is similar to the dropout strategy and sparse penalty in a sense, can be explained as a regularization to avoid over-fitting [30]. As a result, MTL is effective especially when the training data is limited, in which case the over-fitting problem is more likely to occur. By adding additional articulatory attribute targets, subtasks weaken the excessive dependence between the model and the primary task. Subtask learning can also improve the model performance by applying additional information, such as accent and speaker. Taking MTL-DNN as an example, subtask learning increases the discrimination of the hidden layer outputs on these additional tasks, which leads to a more discriminative hidden layer for the primary classification task.

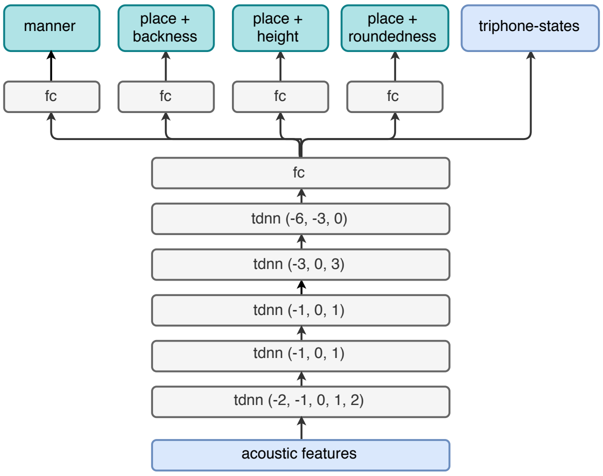

Fig. 2. The architecture of MTL-TDNN-A, where the four subtasks for classifying articulatory attributes are respectively added after the penultimate layer of STL-TDNN in Fig. 1. Note that the five tasks share almost all hidden layers.

A. Using attribute classification as the subtasks

In a conventional DNN-based acoustic model, based on which a triphone-state classification task is performed to provide the posteriors of the triphone states for the subsequent HMM decoder. Given an input vector $\mathbf{x}$, the posterior probability of the i-th triphone state $s_i^{(p)}$ from the output layer is computed using the softmax function as follows:

$$P(s_i^{(p)}|\mathbf{x}) = \frac{\exp(y_i^{(p)})}{\sum_{j=1}^{N^{(p)}} \exp(y_j^{(p)})}, \forall i = 1, ..., N^{(p)}, \qquad (1)$$

where $y_i^{(p)}$ denotes the i-th output of the triphone-state classification task, and $N^{(p)}$ is the number of triphone states, which is 4,464 in this paper.

When using multi-task learning, we consider the triphone-state classification as the primary task, and use the attribute classification as the subtasks. By forced alignment with a pre-trained phone-based GMM model, each frame of a training speech utterance is labeled with the phone and the articulatory attributes. As mentioned above, these attributes are divided into 4 blocks: namely manner, place + backness, place + height and place + roundedness, each with 6, 10, 11, and 9 attributes. Given an input vector $\mathbf{x}$, the posterior of the i-th attribute for each subtask is also computed using the softmax function as follows:

$$P(s_i^{(a)}|\mathbf{x}) = \frac{\exp(y_i^{(a)})}{\sum_{j=1}^{N^{(a)}} \exp(y_j^{(a)})}, \forall i = 1, ..., N^{(a)}, \qquad (2)$$

where $y_i^{(a)}$ denotes the i-th output of the attribute classification subtask, and and $N^{(a)}$ is the number of the attributes in the subtask.

We use the cross-entropy as the training criterion. The cross-entropy of the primary task (or a substask) is calculated as follows:

$$E^{(\cdot)} = \sum_{\mathbf{x}} \sum_{i=1}^{N^{(\cdot)}} d_i^{(\cdot)} \log P(s_i^{(\cdot)}|\mathbf{x}), \qquad (3)$$

where d_i denotes the target value of the i-th triphone state (or attribute), which is 1 when $\mathbf{x}$ belongs to the i-th triphone state (or attribute) and is 0 otherwise, and N is the number of triphone states (or attributes). $E^{(p)}$ is the cross-entropy of the primary task, and $E^{(a_1)}$, $E^{(a_2)}$, $E^{(a_3)}$, and $E^{(a_4)}$ are the cross-entropy of the four subtasks, respectively. The cross-entropy of the overall attribute classification subtask $E^{(a)}$ is the average loss of the four attribute layers:

$$E^{(a)} = \frac{1}{4}(E^{(a_1)} + E^{(a_2)} + E^{(a_3)} + E^{(a_4)}). \quad (4)$$

Finally, the MTL-DNN is trained by minimizing the weighted summation of $E^{(p)}$ and $E^{(a)}$:

$$E = (1 - \alpha)E^{(p)} + \alpha E^{(a)}, \quad (5)$$

where α is the weight that controls the proportion of gradient calculated from the secondary task.

B. Time-delay neural networks (TDNNs)

In our MTL-DNN architectures, we use the time-delay neural network (TDNN) [10, 11] as the hidden layers. When processing a wider temporal context, in a standard DNN, the initial layer learns an affine transform for the entire temporal context. However, in a TDNN architecture, the initial transforms are learned on narrow contexts and the deeper layers process the hidden activations from a wider temporal context. Hence, the higher layers have the ability to learn wider temporal relationships. Each layer in a TDNN operates at a different temporal width, increasing with higher layers of the network.

In a typical TDNN, hidden activations are computed at all time steps. However there are large overlaps between input contexts of activations computed at neighboring time steps. Under the assumption that neighboring activations are correlated, they can be sub-sampled.

Fig. 1 shows the STL-TDNN architecture. There are 6 hidden layers, including 5 TDNN-based layers and 1 fully-connected layer. The layer-wise context shows the information about sub-sampling splicing indices of TDNN-based layers. For example, the first TDNN-based layer is fairly typical, which splices together frames $(t-2)$ through $(t+2)$ at the input layer. The third layer splices sub-sampling contexts, including $(t-3)$, t and $(t+3)$ vectors output by the previous layer.

We propose two types of MTL-TDNN architectures. MTL-TDNN-A (Fig. 2) is the conventional architecture, which directly adds the attribute classification subtasks to STL-TDNN. We modify the MTL-TDNN-A architecture to MTL-TDNN-B (Fig. 3). The attribute layers connect to the third TDNN layer. In sight of regularization, training the substasks will affect the output of the shared hidden layers by back propagation. In TDNN, each layer outputs a sequence of feature vectors. The feature vectors computed by the lower TDNN layers are raw, while the feature vectors computed by the higher layers contain more triphone-state information. We connect the substask layers to a lower layer in order to affect the raw feature vectors directly. It can increase the effect of regularization in MTL-TDNN.

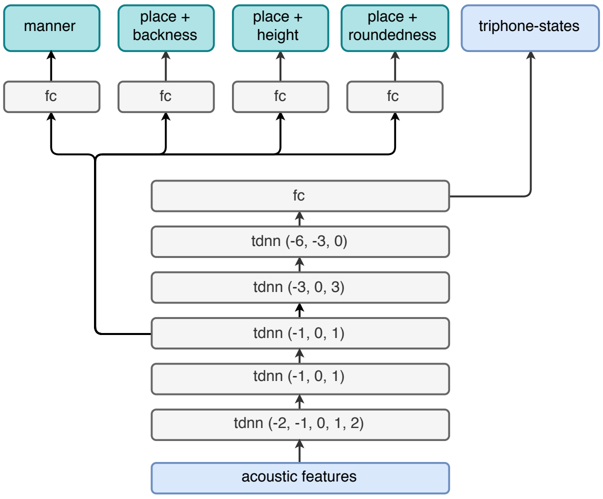

Fig. 3. The architecture of MTL-TDNN-B, where the four subtasks are connected to the third hidden layer. It means that the parameters of the following higher hidden layers are only affected by the loss from the triphone-state classification layer in the training phase.

IV. Experiments

A. Corpora

1) MATBN: MATBN is a publicly available Mandarin Chinese broadcast news corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003, which has been segmented into separate stories and transcribed manually [31]. Each story contains the speech of one studio anchor, as well as several field reporters and interviewees. In our experiments, we used a subset of 25-hour speech data to train the acoustic models and tested them on two testing sets, namely `dev` and `test`, each consisting of 1.4 hours of speech. We performed LVCSR experiments on the MATBN corpus and evaluated the performance in terms of the character error rate (CER).

2) TCC-300: TCC-300 is a collection of microphone speech provided by 3 universities in Taiwan : NTU, NCKU and NCTU [32]. The speech data from each university were recorded by 100 speakers (50 males and 50 females). In the NTU corpus, the recording script has been carefully designed, considering the syllables in a large text corpus and their frequencies. It contains 6,509 utterances, 52,218 syllables, and 141,536 phones. We evaluated the acoustic models by conducting free syllable/phone decoding without language model constraints on the TCC-300 corpus.

3) Lexicon and language model: The lexicon contains 91,573 Chinese words, including 66,290 words from the CKIP[2] lexicon, 5,404 words extracted automatically from the Central News Agency (CNA) news stories in 2001 and 2002 in Chinese Gigaword [33], and 19,879 words from Word List with Accumulated Word Frequency in Sinica Corpus[3] (WLWAWFS 3.0). The word-based trigram language model was trained with Kneser-Ney backoff smoothing using the

[2] http://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

[3] http://elearning.ling.sinica.edu.tw/eng_jindai.html

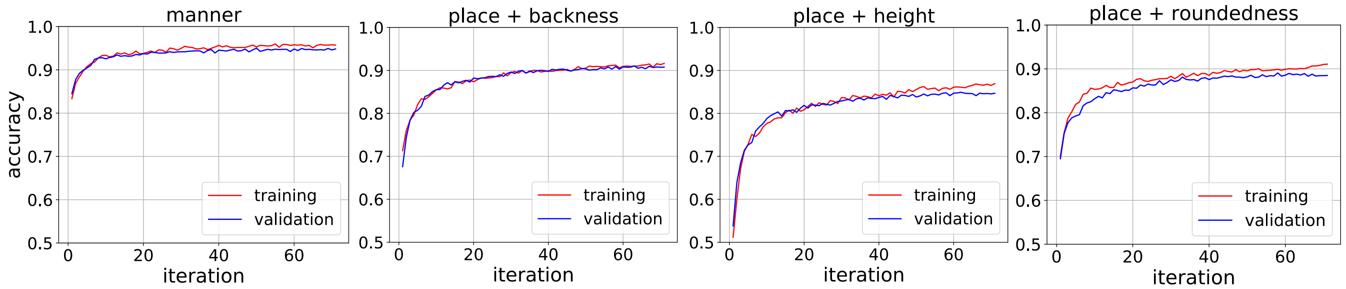

Fig. 4. The training history of attribute classification in MTL-TDNN-B, where each subplot shows the training and validation accuracies with respect to training iterations. Note that the validation set contains 300 utterances excerpted from the training set.

TABLE III
CHARACTER ERROR RATES (%) FOR TWO BASELINES (STL-TDNN AND MTL-TDNN-ML) AND OUR PROPOSED MODELS (MTL-TDNN-A AND MTL-TDNN-B) EVALUATED ON MATBN. (2) AND (3) AFTER MTL-TDNN-B DENOTE THE NUMBER OF SHARED HIDDEN LAYERS.

			+sMBR	
Model	dev	test	dev	test
STL-TDNN	7.49	7.39	7.45	7.44
MTL-TDNN-ML	7.68	7.65	7.69	7.64
MTL-TDNN-A	7.26	7.36	7.25	7.36
MTL-TDNN-B (2)	7.31	**7.3**	7.29	**7.28**
MTL-TDNN-B (3)	**7.24**	**7.3**	**7.2**	7.31

TABLE IV
SYLLABLE ERROR RATES (SER) AND PHONE ERROR RATES (PER) FOR STL-TDNN AND OUR PROPOSED MODELS (MTL-TDNN-A AND MTL-TDNN-B) EVALUATED ON TCC-300. THE FIRST THREE HIDDEN LAYERS ARE SHARED IN MTL-TDNN-B.

Model	**SER (%)**	**PER (%)**
STL-TDNN	26.05	15.88
MTL-TDNN-A	25.42	14.97
MTL-TDNN-B	**24.47**	**14.94**

SRILM toolkit [34]. The textual training corpus was compiled from the CNA news stories from 2006 to 2010 in Chinese Gigaword.

B. Units for acoustic modeling

In Mandarin speech recognition, the initial-final with tone phonetic alphabet (e.g., Formosa Phonetic Alphabet, ForPA) is commonly used as the phonetic units for acoustic modeling [35]. However, in order to fetch the articulatory attributes corresponding to each phone, we used the phone set derived from International Phonetic Alphabet (IPA) for Mandarin speech [28]. We added the tonal symbols to the vowel phonemes.

C. Input features

To extract acoustic features, spectral analysis was applied to a 25 ms frame of speech waveform every 10 ms. For each frame, 40 high-resolution MFCCs, derived by DCT conducted on 40 Mel-frequency bins and normalized by utterance-based mean subtraction, were used as the input to the NN-based acoustic models. Since Mandarin is a tonal language, 3 pitch-related features were concatenated to the 40-dimensional MFCCs [36] for the Mandarin ASR task. Moreover, we appended a 100-dimensional i-vector to each acoustic frame.

D. Baseline systems

The GMM-HMM system was pre-trained to generate reliable frame-to-state (or pdf-id) alignments for subsequent neural network training. We used the fifth round triphone system (i.e., `tri5`[4]) to decode and generate the alignments for each utterance.

One of our baseline systems was STL-TDNN built with 6 hidden layers, each containing 650 hidden nodes. The output layer was a softmax-activated layer with 4,464 nodes for MATBN, and the maximum change in the parameters per mini-batch was set to 1.5. The mini-batch sizes were 256 and 128. The initial and final effective learning rates were set to 0.0015 and 0.00015, respectively, and the total number of training epochs was set to 3, while additional 4 training epochs were used for fine tuning by state-level minimum Bayes risk (sMBR). [37]

Another baseline is the MTL-TDNN system, named MTL-TDNN-ML, which used a single multi-label classification subtask in multi-task learning. Here the number of output nodes of the substask is 21 because there are 21 unique articulatory attributes in total.

E. Proposed systems

Both MTL-TDNN-A and MTL-TDNN-B contain 150 hidden nodes before each attribute layer block. The weight α was set to 10^{-6}. The other hyper-parameters were the same as the baseline systems.

F. Results

Table III shows the character error rates (CER) achieved by STL-TDNN and three MTL-TDNN systems evaluated on MATBN. Compared to the STL-TDNN baseline system, the traditional MTL-TDNN-ML system did not gain any improvements. In contrast, the proposed MTL-TDNN-A system achieved relative error rate reductions of 3% and 0.4% on the `dev` and `test` sets, respectively. The MTL-TDNN-B (3)

[4] https://github.com/kaldi-asr/kaldi/blob/master/egs/formosa/s5/run.sh

system achieved relative error rate reductions of 3.3% and 1.2% on `dev` and `test`, respectively. MTL-TDNN-B slightly outperformed MTL-TDNN-A. When the number of shared hidden layers in MTL-TDNN-B was reduced from 3 to 2, the performance was slightly degraded. Surprisingly, the sMBR fine tuning did not consistently improve these four models.

Compared to MTL-TDNN-ML, MTL-TDNN-A achieved relative error rate reductions of 5.5% and 3.8% on `dev` and `test`, respectively. MTL-TDNN-B (3) achieved relative error rate reductions of 5.7% and 4.6% on `dev` and `test`, respectively. The experimental results confirm the advantage of dividing the articulartory attributes into four exclusive blocks. The results also confirm our assumption that we should share only the first few layers in multi-task learning because the articulatory information associated with the physical process is not as abstract and composite as the phonetic descriptions.

The training histories of the four subtasks are showed in Fig. 4. The validation accuracies of the manner, place + backness, place + height, and place + roundedness subtasks are 0.951, 0.913, 0.84, and 0.88, respectively. Although α is relatively small, it still helps to yield good performance in the articulartory attribute classification subtasks.

Table IV shows the syllable error rates and the phone error rates of the free syllable/phone decoding experiments on the TCC-300 corpus. Tone errors were ignored because the articulatory attributes were not concerned with the difference in tone. Compared to the STL-TDNN baseline system, the MTL-TDNN-A system achieved relative reductions of 2.4% and 5.7% in terms of syllable and phone error rates, respectively. The MTL-TDNN-B system achieved relative reductions of 6.1% and 5.9% in terms of syllable and phone error rates, respectively. The number of shared hidden layers in MTL-TDNN-B was set to 3 in this experiment. The experimental results reaffirm the appropriateness of sharing only the first few layers of the neural network in multi-task learning.

V. Conclusions and Future Work

In this paper, we have proposed two MTL-TDNN architectures for acoustic modeling, which are trained by adding four articulatory attribute classification subtasks in multi-task learning. We split the articulatory attributes into four blocks to perform four classification subtasks separately, instead of directly performing a single multi-label classification subtask. We have evaluated the proposed framework on the MATBN Mandarin Chinese LVCSR task and the TCC-300 Mandarin free syllable/phone decoding task. MTL-TDNN-ML (using a single multi-label classification subtask in multi-task learning) did not enhance the performance compared with STL-TDNN (the single-task learning counterpart of MTL-TDNN). The proposed models MTL-TDNN-A and MTL-TDNN-B achieved relative error rate reductions, although not very significant. MTL-TDNN-B performed best in the ASR experiments and also performed well in attribute classification.

As future directions, we will experiment with other corpora in English or other languages. With the IPA phoneme set, which can present all phones in different languages, our system can be used in cross-language LVCSR. We will also implement the MTL-TDNN models on the Kaldi `chain` setup. In addition, articulatory information not only improves the LVCSR task but may also be useful in the CAPT field for mispronunciation detection and diagnosis.

References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech Using Neural Networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.

[3] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. R. Rabiner, "An Overview on Automatic Speech Attribute Transcription (ASAT)," in *Proc. Interspeech*, 2007.

[4] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-daniel, J. Morris, S. M. Siniscalchi, Y. Tsao, and Y. Wang, "Detection-Based ASR in the Automatic Speech Attribute Transcription Project," in *Proc. Interspeech*, 2007.

[5] C. Zhang, Y. Liu, and C.-H. Lee, "Detection-based Accented Speech Recognition Using Articulatory Features," in *Proc. IEEE ASRU*, 2011.

[6] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting Attribute and Phone Estimation Accuracies with Deep Neural Networks for Detection-based Speech Recognition," in *Proc. ICASSP*, 2012.

[7] S. M. Siniscalchi, J. Li, and C.-H. Lee, "A Study on Lattice Rescoring with Knowledge Scores for Automatic Speech Recognition," in *Proc. Interspeech*, 2006.

[8] H. Zheng, Z. Yang, L. Qiao, J. Li, and W. Liu, "Attribute Knowledge Integration for Speech Recognition Based on Multi-task Learning Neural Networks," in *Proc. Interspeech*, 2015.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE ASRU*, 2011.

[10] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[11] V. Peddinti, D. Povey, and S. Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *Proc. Interspeech*, 2015.

[12] P.-T. Huang, H.-S. Lee, S.-S. Wang, K.-Y. Chen, Y. Tsao, and H.-M. Wang, "Exploring the Encoder Layers of

Discriminative Autoencoders for LVCSR," in *Proc. Interspeech*, 2019.

[13] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal Low-rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech*, 2018.

[14] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low Latency Acoustic Modeling Using Temporal Convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.

[15] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Pronunciation Error Detection Using DNN Articulatory Model Based on Multi-lingual and Multi-task Learning," in *Proc. ISCSLP*, 2016.

[16] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Efficient Learning of Articulatory Models Based on Multi-Label Training and Label Correction for Pronunciation Learning," in *Proc. ICASSP*, 2018.

[17] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving Non-native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-based Speech Attribute Modeling," in *Proc. ICASSP*, 2016.

[18] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving Mispronunciation Detection for Non-native Learners with Multisource Information and LSTM-based Deep Models," in *Proc. Interspeech*, 2017.

[19] S. Maeda, "Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-tract Shapes Using an Articulatory Model," in *Speech Production and Speech Modelling*. Springer, 1990, pp. 131–149.

[20] P. Badin, G. Bailly, M. Raybaudi, and C. Segebarth, "A Three-dimensional Linear Articulatory Model Based on MRI Data," in *Proc. ETRW*, 1998.

[21] O. Engwall, "Combining MRI, EMA and EPG Measurements in a Three-dimensional Tongue Model," *Speech Communication*, vol. 41, no. 2-3, pp. 303–329, 2003.

[22] P. Birkholz, D. Jackel, and B. Kroger, "Construction And Control Of A Three-Dimensional Vocal Tract Model," in *Proc. ICASSP*, 2006.

[23] Y. Payan and P. Perrier, "Synthesis of VV Sequences with a 2D Biomechanical Tongue Model Controlled by the Equilibrium Point Hypothesis," *Speech communication*, vol. 22, no. 2-3, pp. 185–205, 1997.

[24] J.-M. Gérard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan, "A 3D Dynamical Biomechanical Tongue Model to Study Speech Motor Control," *arXiv preprint physics/0606148*, 2006.

[25] S. Fagel and K. Madany, "A 3-D Virtual Head as a Tool for Speech Therapy for Children," in *Proc. Interspeech*, 2008.

[26] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can You 'read' Tongue Movements? Evaluation of the Contribution of Tongue Display to Speech Understanding," *Speech Communication*, vol. 52, no. 6, pp. 493–503, 2010.

[27] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three Dimensional Articulator Model for Speech Acquisition by Children with Hearing Loss," in *Proc. UAHCI*, 2007.

[28] I. P. Association and Others, *Handbook of the International Phonetic Association: A Guide to theUuse of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[29] R. A. Caruana, "Multitask Learning: A Knowledge-Based Source of Inductive Bias," in *Proc. ICML*, 1993.

[30] T. Evgeniou and M. Pontil, "Regularized Multi-task Learning," in *Proc. ACM SIGKDD*, 2004.

[31] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 10, no. 2, pp. 219–236, 2005.

[32] "MAT Speech Database – TCC-300," 2005. [Online]. Available: http://www.aclclp.org.tw/doc/tcc300_brief.pdf

[33] D. Graff and K. Chen, "Chinese Gigaword (LDC2003T09)," 2003.

[34] A. Stolcke, "SRILM: An Extensible Language Modeling Toolkit," in *Proc. Interspeech*, 2002.

[35] R.-y. Lyu, M.-s. Liang, and Y.-c. Chiang, "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka , and Mandarin," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 9, no. 2, pp. 1–12, 2004.

[36] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," *Proc. ICASSP*, 2014.

[37] D. Povey and K. Vesel, "Sequence-discriminative Training of Deep Neural Networks," in *Proc. Interspeech*, 2013.