MULTI-TASK DEEP NEURAL NETWORK ACOUSTIC MODELS WITH MODEL ADAPTATION USING DISCRIMINATIVE SPEAKER IDENTITY FOR WHISPER RECOGNITION

Jingjie Li[1], *Ian McLoughlin*[1], *Cong Liu*[2], *Shaofei Xue*[1], *Si Wei*[2]

[1]National Engineering Laboratory of Speech and Language Information Processing
University of Science and Technology of China
[2] iFlytek Research

{jingjie,xuesf}@mail.ustc.edu.cn, ivm@ustc.edu.cn, {congliu2,siwei}@iflytek.com

ABSTRACT

This paper presents a study on large vocabulary continuous whisper automatic recognition (wLVCSR). wLVCSR provides the ability to use ASR equipment in public places without concern for disturbing others or leaking private information. However the task of wLVCSR is much more challenging than normal LVCSR due to the absence of pitch which not only causes the signal to noise ratio (SNR) of whispers to be much lower than normal speech but also leads to flatness and formant shifts in whisper spectra. Furthermore, the amount of whisper data available for training is much less than for normal speech. In this paper, multi-task deep neural network (DNN) acoustic models are deployed to solve these problems. Moreover, model adaptation is performed on the multi-task DNN to normalize speaker and environmental variability in whispers based on discriminative speaker identity information. On a Mandarin whisper dictation task, with 55 hours of whisper data, the proposed SI multi-task DNN model can achieve 56.7% character error rate (CER) improvement over a baseline Gaussian Mixture Model (GMM), discriminatively trained only using the whisper data. Besides, the CER of the proposed model for normal speech can reach 15.2%, which is close to the performance of a state-of-the-art DNN trained with one thousand hours of speech data. From this baseline, the model-adapted DNN gains a further 10.9% CER reduction over the generic model.

Index Terms— Whisper recognition, model adaption, speaker code, multi-task DNN, Silent speech interface

1. INTRODUCTION

Nowadays, large vocabulary continuous automatic speech recognition (LVCSR) services have become common in smartphones and in smart-cars, primarily enabled by the success of deep neural network (DNN) acoustic models [1, 2, 3, 4] coupled with the availability of large speech training databases. However, even when such applications are both convenient and accurate, people tend to avoid using them in public places, e.g. library, museum or during meetings. Some users may not want to interfere with others, while some are concerned by the lack of privacy when talking to such devices in the presence of others.

In recent years, silent speech interfaces [5] have been proposed, to enable speech-like communication without requiring normal speech. Whispers are one such mechanism, used frequently in human-to-human conversation where normal speech is either inappropriate or impossible [6]. Because of the lack of pitch in whispers, normal speech LVCSR systems tend to perform extremely poorly with whisper input [7]. Therefore, much effort has been devoted to whisper recognition. For example, a body-conducted microphone [8] was developed to detect soft whispers by reducing the effect of environmental noise. In order to train a GMM-HMM system with limited whisper data, initial GMM acoustic models have been trained using large normal speech databases, and then adapted using maximum likelihood linear regression (MLLR), constrained MLLR [9] or VTS [10]. To partially mitigate the phonological information loss caused by the lack of pitch in whispers, articulatory features have also been used in parallel with spectral features [11]. In [12], several feature extraction strategies were studied to compensate energy, spectral slope and formant shift in whisper, for a small vocabulary whisper recognition task.

In this paper, we aim to derive effective strategies for building a practical whisper-LVCSR (wLVCSR) using a DNN-HMM paradigm. Even though DNN acoustic models trained using whisper data can provide obvious recognition improvement over baseline GMM acoustic models, the resulting DNN acoustic model is still far from satisfactory for practical use. Inspired by the idea of multi-lingual DNNs [13, 14, 15, 16], that aim to build ASR systems quickly with only small amounts of target language speech data, multi-task

The authors gratefully acknowledge the support of the Fundamental Research Funds for the Central Universities, China under grant nos. KY2100060002 and WK2100000002.

DNN acoustic models are proposed in this paper, to exploit both the enormous amount of existing speech data as well as existing highly accurate DNN acoustic speech models. We also propose a model adaptation strategy using discriminative speaker identity [17, 18, 19] to adapt the multi-task DNN acoustic models for whispers using just a few utterances.

2. BASELINE WHISPER ASR PERFORMANCE

2.1. Mandarin whisper corpus

Whisper data was recorded in quiet office environments (about 30dBA background noise) using different kinds of smartphones. The training set comprises 40,323 whisper utterances (55 hours of up to 150 utterances from 287 speakers). The test set comprises a further 1,568 utterances from 12 speakers. 20 utterances selected randomly from each speaker act as enrolment data. A normal speech test set comprises 600 utterances from 4 speakers, each of whom also participated in the whisper test recordings. In all experiments, decoding is performed by a trigram language model consisting of 26 million n-grams. We also use a separate 1 million utterance corpus (about 1,000 hours, containing 3,056 speakers) of Mandarin speech to train a highly accurate normal speech DNN acoustic model.

2.2. Baseline systems

In order to determine how different acoustic models, trained by different datasets, affect the performance of whisper recognition, we first define a baseline GMM-HMM system (GMM_{MFCC}). Regular 42-dimension features (13 static MFCCs, Δ, $\Delta\Delta$ and 3 pitch features) are extracted from whispers and modeled by a baseline GMM-HMM system with 3004 context dependent (CD) senones, each with 20-Gaussians trained by maximum likelihood (ML) and then trained with minimum phone error rate (MPE) criteria.

As noted in Section 1, the spectra of whispers deviates from normal speech largely due to lack of pitch. Thus MFCCs used for normal speech ASR are consequently unlikely to be optimal for whispers. Therefore, we created a bottleneck DNN (BN) to extract discriminative features [20] for whispers. This is a DNN with 5 hidden layers of 2048 nodes, apart from a 42 node BN layer in the middle. The softmax output has 3004 nodes corresponding to the senones in the baseline GMM. The network is initialised by RBM layer-wise pre-training, fine tuned by back propagation over 10 iterations. We then used the features extracted by this BN to train a second GMM recogniser (denoted GMM_{BN}).

Besides this, a regular DNN (DNN_w) with 6 hidden layers and each layer with 2048 nodes is trained in the same way to the BN network described above. Both DNNs use 75-dimension features (24 static Mel filter-bank features, Δ, $\Delta\Delta$ and 3 pitch features) and include an input context of 10 neighbouring frames (± 5) yielding a final dimensionality of 825. All whisper data is aligned by the baseline GMM.

Table 1. CER of different models for whispers and speech.

	Whisper, % (relative)	Normal, %
GMM_{MFCC}	38.7	90.3
GMM_{BN}	32.3 (17%)	73.8
DNN_w	**28.1** (28%)	73.6
DNN_{nA}	62.6 (-62%)	24.5
DNN_{nB}	31.2 (20%)	**10.5**

Table 2. PER of different tones recognised by DNN_{nB} on the whisper and normal speech test sets.

	Whisper, %	Normal, %
tone1	26.9	6.2
tone2	26.2	8.3
tone3	23.7	9.2
tone4	21.6	4.4
consonants	11.8	2.4

To explore further, two additional DNNs (DNN_{nA}, DNN_{nB}) were trained using 300 and 1,000 hours of normal speech respectively, aligned by a highly accurate DNN with 9004 CD states. Both have 6 hidden layers of 2048 nodes and 825-dimensional concatenated input features.

Table 1 shows the recognition performance in CER for each of these systems, on both whisper and normal speech test sets. Clearly, the performance on whisper and normal speech for both GMM_{BN} and DNN_w out-perform the baseline GMM. The GMM_{BN} score demonstrates that bottleneck features are more representative for whisper recognition than traditional MFCCs. The further improvement of DNN_w shows the discriminative capability of the DNN. Notice that even though DNN_{nB} performance with speech is excellent at 10.5%, its whisper performance is only 31.2%. It is interesting that this highly accurate acoustic model does not work well for whisper input, echoing findings in [7].

It is interesting to consider the phoneme error rates (PER) for different tones using the DNN_{nB} model in Table 2. The normal speech PER is much smaller than for whispers, furthermore, the model recognises consonants more precisely than vowels for both whispers and speech. Mandarin is a tonal language where vowels are voiced with one of four lexical tones [21]. Consonants are non-tonal and mainly unvoiced. In normal speech, tone is mainly conveyed by pitch, but since whispers are pitchless, the tones become less distinct. With this in mind, the results shown in Table 2 are reasonable.

3. MULTI-TASK DNN FOR WHISPER ASR

As Section 2 demonstrated, the layer-by-layer structure of DNN acoustic models can derive deep lingual information from whispers which allows them to perform much better

than a baseline GMM acoustic model. The challenge now is to further improve the recognition accuracy achieved by whisper DNN acoustic models, since 70%, is clearly insufficient for practical use. Obtaining more whisper data might yield a possible solution, but is time-consuming and expensive. Therefore, inspired by the idea of multi-lingual DNNs, a multi-task DNN acoustic model is proposed, to overcome the limitations of whisper training data, by exploiting the much better resources available for normal speech.

In this paper, the idea of a multi-task DNN is proposed in three different methods: (1) to align whisper data using DNN_{nB} rather than the baseline GMM system. Whisper data would then share CD states with normal speech and the aligned data can be used to train a DNN acoustic model ($M\_RBM\_DNN_w$). (2) instead of pre-training a DNN using a layer-wise RBM based algorithm, a whisper DNN model can be directly initialised by the accurate normal speech model DNN_{nB}. This can also be viewed as a way to enable DNN_{nB} to recognise whispers by re-tuning it with whisper data (M_DNN_w). (3) another method is to re-tune DNN_{nB} not only with whisper data but also mixed with normal speech. Specifically, 50 and 100 hours of normal speech are selected randomly from the normal speech corpus, mixed with whisper data, aligned by DNN_{nB}, and then deployed to retrain two mixed speech style DNNs ($M\_DNN_{w55\_n50}$, $M\_DNN_{w55\_n100}$). The results of all strategies are summarised in Table 3, where we can see large relative error reductions are achieved by the proposed multi-task DNNs over the baseline GMM.

Note that M_DNN_w is the first model whose CER is less than 20% on this test. If whisper data is re-aligned by the obtained M_DNN_w, and then used to re-tune the DNN_{nB}, the resulting model (Mul_DNN_{wB}) achieves a slight performance gain over M_DNN_w. Therefore, DNN_{nB} seems able to align this whisper training set quite accurately.

Without degrading recognition performance on whisper input, the two mixed speech style DNNs at the bottom can reduce the CER of normal speech from 34% to less than 15%, which is close to DNN_{nB} on speech. This is important since it is the first model that achieves good accuracy for both whisper and normal speech tasks.

The PER of tones recognised by $M\_DNN_{w55\_n50}$ for whispers and normal speech is given in table 4. Compared to the results shown in Table 2, $M\_DNN_{w55\_n50}$ improves recognition of different tones of whispers very obviously, with only a small performance loss on normal speech.

Table 3. CER of different multi-task DNNs on whisper and normal speech tests.

	Whisper, % (relative)	Normal, %
GMM_{MFCC}	38.7	90.3
$M\_RBM\_DNN_w$	20.1 (48%)	59.6
M_DNN_w	17.2 (56%)	34.0
M_DNN_{wB}	16.9 (57%)	34.1
$M\_DNN_{w55\_n50}$	**16.8** (57%)	15.2
$M\_DNN_{w55\_n100}$	17.0 (56%)	**13.7**

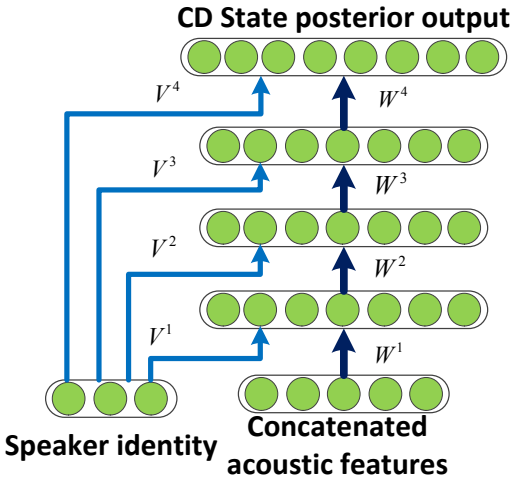

Fig. 1. Model adaptation of DNNs using speaker identity information.

4. FAST MODEL ADAPTATION

Now having an effective speaker independent (SI) DNN acoustic model for whispers, we investigate the effectiveness of model adaptation using speaker identity information.

4.1. Fast model adaptation based on speaker identity

Fig. 1 illustrates the proposal to perform speaker model adaptation based on discriminative speaker identity. The core idea to feed each speaker identity (which may also carry environmental background and channel inference) into hidden and output layers of the initial generic DNN, through a set of generic connection weights (all $\mathbf{V}^l$), shared by all speakers. Thus speaker identiies can be used to adapt the SI DNN to more precisely locate target speakers in the model space.

Let us denote $\mathbf{W}^l$ as the lth layer weights in the generic DNN, with L hidden layers, and $\mathbf{S}^c$ as the identity of the cth

Table 4. PER of different tones recognised by model $M\_DNN_{w55\_n50}$ for whisper and normal speech.

	Whisper, %	Normal, %
tone1	16.5	8.0
tone2	14.1	9.2
tone3	14.2	10.4
tone4	9.0	4.8
consonants	5.6	3.3

speaker. Therefore, the output of layer l, $\mathbf{h}^l$ can be computed as follows:

$$\boldsymbol{h}^l = \sigma(\boldsymbol{W}^l \boldsymbol{h}^{l-1} + \boldsymbol{V}^l \boldsymbol{S}^c), (1 \leq l \leq L+1) \quad (1)$$

During the model adaptation phase, only $\mathbf{V}^l$ and $\mathbf{S}^c$ are updated though stochastic descent in a supervised way using speaker labels and CD state labels for each utterance, while keeping all $\mathbf{W}^l$ unchanged. If E is the objective function during error back-propagation, the derivative of any connection weight $\mathbf{V}^l_{ij}$ that connects the ith speaker identity node to node j in layer l of the generic DNN can be computed as:

$$\frac{\partial E}{\partial \boldsymbol{V}^l_{ij}} = \frac{\partial E}{\partial \boldsymbol{h}^l_j}(1 - \boldsymbol{h}^l_j)\boldsymbol{h}^l_j \boldsymbol{S}^c_i \quad (2)$$

If speaker code is used as speaker identity, the derivative of $\boldsymbol{S}^c_i$ in the ith node of the speaker code can be computed:

$$\frac{\partial E}{\partial \boldsymbol{S}^c_i} = \frac{1}{L}\sum_{l=1}^{L+1}\sum_{j=1}^{J}\frac{\partial E}{\partial \boldsymbol{h}^l_j}(1 - \boldsymbol{h}^l_j)\boldsymbol{h}^l_j \boldsymbol{V}^l_{ij} \quad (3)$$

$\boldsymbol{V}^l$ are learned across all data from each speaker in the training set, and $\boldsymbol{S}^c$ is only updated by the data belonging to speaker c. If an i-vector is used as speaker identity $\boldsymbol{S}^c$, it can be extracted following standard procedure as in [19], and would stay unchanged during training. For any new speaker in the test set, only a few utterances are needed to compute the speaker code or to extract the i-vector. This enables fast speaker model adaptation.

4.2. Performance of fast model adaptation based on speaker identity for whisper recognition

In this paper, the performance of model adaptation using only a few whisper utterances is evaluated. The SI M_DNN_w from Section 3 is used as the initial generic DNN to be adapted. During the training of adaptation matrices $\boldsymbol{V}^l$ and $\boldsymbol{S}^c$ (if speaker code is used), initial learning rate is set to 0.5, and is halved after the first three epochs. The momentum is kept as 0.9, mini-batch size is 1024 and the training process iterates for 5 epochs. Two different speaker identity methods, i-vector and speaker code, are evaluated in the experiments. For speakers in the training set, all the whisper utterances are used to extract a 100-dimension i-vector or to learn speaker code (SC). For speakers in the test set, only a small number of utterances, e.g., 5, 10, 15, 20, are randomly selected from the enrollment set, to infer speaker identities. To learn speaker code for speakers in the test set, the learning rate was 0.02, bunch size set to 128, and learning repeated over 5 epochs. In addition, the performance impact of different sizes of speaker code (100, 500, 1000) is evaluated. Experimental results on the whisper test set are summarised in Table 5.

From the second row in the table, if less than 20 utterances are used to extract the i-vector, there is no CER reduction. But if all utterances for each speaker in the test and enrollment sets are used to extract an i-vector (denoted 'max'), an 11.3% relative CER reduction can be achieved. This result can be explained by noting that speaker identity information is degraded in whispers, so fewer utterances may be insufficient to extract a robust i-vector. On the other hand, if SC is used as speaker identity, the model adaptation can provide stable and effective CER improvement using only a few whisper utterances. From the results in the 3rd to 5th rows in Table 5, if only a small amount of enrollment data is available, a smaller size of SC tends to perform slightly better. Vice versa, larger SC sizes can achieve better performance when more adaptation data is available. For example, if a 1000-dimension speaker code is used with all 20 whisper utterances from each speaker in the enrollment set, 10.9% relative CER reduction can be achieved over the baseline generic DNN. Finally, this advantage of speaker code over i-vector might be because more robust speaker identity information can be learned by the deep structure shown in Fig. 1.

Table 5. CER (in %) for M_DNN_w model adaptation using 5, 10, 15, 20 utterances with different speaker identity methods on the whisper test set (baseline performance is 17.23%).

utterances →	5	10	15	20	max
100-dim i-vec	17.55	17.36	17.51	17.51	15.48
100-dim SC	**16.28**	16.22	16.09	15.93	15.54
500-dim SC	16.35	16.23	16.07	15.91	15.43
1000-dim SC	16.39	**15.98**	**16.06**	**15.36**	**15.09**

5. CONCLUSION

In this paper, we have proposed and evaluated SI multi-task DNN acoustic models for wLVCSR by taking advantage of the availability of extremely large normal speech resources. The absence of pitch in the vowel component of whispers not only leads to the SNR of whisper being much lower than normal speech, but also causes speaker identity and linguistic information loss. When applying a model adaptation strategy to the multi-task DNN models, further improvement in whisper recognition accuracy rate is shown to be possible by using discriminative speaker identity information. On the whispered Mandarin Chinese dictation task, with 55 hours whisper data, the proposed SI multi-task DNN model achieves a 57% improvement in CER over a baseline GMM which is discriminatively trained using the same data. The model adapted speaker dependent DNN can further reduce CER over the SI model by almost 11%, even when using only a small number of whisper utterances for adaptation. The results results presented here describe Mandarin speech recognition results, due primarily to the availability of good training data. The method should be applicable to any other language, including English, with sufficient training data.

6. REFERENCES

[1] Dong Yu, Li Deng, and G Dahl, "Roles of pretraining and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[2] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 20, no. 1, pp. 30–42, 2012.

[3] Geoffrey Hinton, Li Deng, and Dong et. al. Yu, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] Jia Pan, Cong Liu, Zhiguo Wang, Yu Hu, and Hui Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling.," in *ISCSLP*, 2012, pp. 301–305.

[5] Thomas et. al. Hueber, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[6] McLoughin I, Jingjie Li, and Song Yan, "Reconstruction of continuous voiced speech from whispers," in *INTERSPEECH*, 2013.

[7] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[8] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proc. ICASSP*, 2003, vol. 5, pp. V–708.

[9] Denis Babani, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Acoustic model training for non-audible murmur recognition using transformed normal speech data," in *Proc. ICASSP*, 2011, pp. 5224–5227.

[10] Chen-Yu Yang, Georgina Brown, Liang Lu, Junichi Yamagishi, and Simon King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation.," in *ISCSLP*, 2012, pp. 220–223.

[11] Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel, "Whispery speech recognition using adapted articulatory features.," in *Proc. ICASSP (1)*, 2005, pp. 1009–1012.

[12] Shabnam Ghaffarzadegan, Hynek Boril, and John HL Hansen, "UT-vocal effort ii: Analysis and constrained-lexicon recognition of whispered speech," .

[13] S. Dupont, C. Ris, O. Deroo, and S. Poitoux, "Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents," in *Workshop on Automatic Speech Recognition and Understanding*, Nov 2005, pp. 29–34.

[14] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304–7308.

[15] Georg Heigold, Vincent Vanhoucke, Andrew Senior, Patrick Nguyen, M Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, 2013, pp. 8619–8623.

[16] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation," in *INTERSPEECH*, 2014.

[17] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Workshop on Automatic Speech Recognition and Understanding*, Dec 2013, pp. 55–59.

[18] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, and Lirong Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code," in *Proc. ICASSP*, 2014, pp. 6339–6343.

[19] Andrew Senior and Ignacio Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014, pp. 225–229.

[20] Bing Jiang, Yan Song, Si Wei, Jun-Hua Liu, Ian Vince McLoughlin, and Li-Rong Dai, "Deep bottleneck features for spoken language identification," *PloS one*, vol. 9, no. 7, pp. e100795, 2014.

[21] Ian Vince McLoughlin, "Vowel Intelligibility in Chinese," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 18, no. 1, pp. 117–125, 2010.