DEEP BOTTLENECK FEATURES AND SOUND-DEPENDENT I-VECTORS FOR SIMULTANEOUS RECOGNITION OF SPEECH AND ENVIRONMENTAL SOUNDS

Sakriani Sakti, Seiji Kawanishi, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

{ssakti,kawanishi.seiji.km6,neubig,koichiro,s-nakamura}@is.naist.jp

ABSTRACT

In speech interfaces, it is often necessary to understand the overall auditory environment, not only recognizing what is being said, but also being aware of the location or actions surrounding the utterance. However, automatic speech recognition (ASR) becomes difficult when recognizing speech with environmental sounds. Standard solutions treat environmental sounds as noise, and remove them to improve ASR performance. On the other hand, most studies on environmental sounds construct classifiers for environmental sounds only, without interference of spoken utterances. But, in reality, such separate situations almost never exist. This study attempts to address the problem of simultaneous recognition of speech and environmental sounds. Particularly, we examine the possibility of using deep neural network (DNN) techniques to recognize speech and environmental sounds simultaneously, and improve the accuracy of both tasks under respective noisy conditions. First, we investigate DNN architectures including two parallel single-task DNNs, and a single multi-task DNN. However, we found direct multi-task learning of simultaneous speech and environmental recognition to be difficult. Therefore, we further propose a method that combines bottleneck features and sound-dependent i-vectors within this framework. Experimental evaluation results reveal that the utilizing bottleneck features and i-vectors as the input of DNNs can help to improve accuracy of each recognition task.

Index Terms— Simultaneous recognition of speech and environmental sounds, bottleneck features, sound-dependent i-vector.

1. INTRODUCTION

We live in a world that is filled with a variety of sounds, and our sense of hearing enables us to perceive this world of acoustic vibrations all around us. It is a complex process of picking up sound and attaching meaning to it, but understanding of these sounds provides us with important channels of communication [1]. One of the objectives in human-machine communication is to develop a machine that can achieve human-like performance on this understanding of sound, which is sometimes referred to as "Machine Hearing" [2].

Automatic speech recognition (ASR) systems can be considered as one part of machine hearing that focuses on recognizing a specific variety of sound – human speech signals. Research in ASR has progressed from developing simple machines that recognize speech in clean conditions to developing more sophisticated systems that respond to real spoken language in real environments. However, ASR becomes difficult when recognizing speech with environmental sounds. Standard solutions generally treat environmental sounds as noise, and remove them for better ASR performance. This is motivated by the cocktail party effect [3] where humans have the ability to selectively attend to a single speaker among various sources of conversation and background noise. Techniques to recover clean desired speech signals in noisy and reverberant environment include spectral substraction [4, 5], minimum mean-square error (MMSE) estimation [6, 7], Kalman filtering [8, 9], and recently deep neural networks (DNN) [10, 11]. Several challenge-based workshops focusing on related noisy speech tasks such as the REVERB Challenge [12, 13] or the CHiME Speech Separation and Recognition Challenge [14, 15] have also been held to provide common data and benchmarks suitable for comparing and contrasting the performance of different methods in constructing noise-robust ASR systems.

Environmental sound recognition, on the other hand, focuses on recognizing and classifying various (usually non-speech) environmental sounds. Previous methods include Dat et al.'s robust sound event classifier based on the generalized Gaussian distribution Kullback-Leibler kernel support vector machine (SVM) [16], and Dikmen et al.'s sound event detection using non-negative dictionaries learned from annotated overlapping events [17]. Recent studies have also utilized DNN techniques to recognize sound events [18, 19].

However, in reality, such separate situations almost never exist. Furthermore, more and more applications are interested in providing information not only on the main object but also the surrounding or related objects [20]. For speech applications, it would be useful to understand the overall auditory environment, not only recognizing what is being said, but also being aware of the actions surrounding the utterance. But, unfortunately, very few works have focused on this direction. Previously, there has been work on computers capable of lis-

tening to several things simultaneously [21]. From the perspective of psychoacoustic studies, a study by Morell et al. [22] also showed that people vary widely in their ability to process what they hear, and some people are able to listen a phone message in one ear while a friend is talking into their other ear. Furthermore, Kashino et al. also found that humans can listen up to two things simultaneously but no more [23].

Therefore, in this study, we attempt to investigate these possibilities and construct a system that has the capability to recognize both speech and environmental sounds simultaneously. Particularly, we examine the possibility of training DNNs to jointly perform the multiple tasks of speech and environmental sound recognition, and improve the accuracy of both tasks under the respective noisy conditions. First, we investigate DNN architectures including two parallel single-task DNNs and a single multi-task DNN. However, direct multi-task learning of simultaneous speech and environmental recognition is difficult, due to the fact that as additional environmental sounds get louder, it becomes easier to recognize environmental sounds, but more difficult to perform speech recognition. Therefore, we further propose a method that combines bottleneck features and sound-dependent i-vectors within the DNN framework to overcome this problem.

2. DNN-BASED SIMULTANEOUS RECOGNITION OF SPEECH AND ENVIRONMENTAL SOUND

2.1. Single-task and Multi-task Learning

In the single-task DNN scenario, two DNNs are trained independently for the tasks of speech recognition and environmental sound recognition. The structure of both networks are kept the same, except for the output layers which indicate HMM triphone units or sound class units for speech recognition and environmental sound recognition, respectively.

The single multi-task DNN, on the other hand, is trained to perform both speech recognition and environmental sound recognition at the same time, using a shared representation in the hidden layers. The structure of the multi-task DNN, in comparison with the single-task DNN, is also similar, except that the output layer performs both single-task speech recognition and environmental sound recognition.

The loss function of two parallel single-task DNNs is estimated by the loss function of the speech recognition task and environmental sound recognition task separately, while the loss function of a single multi-task DNN is defined as the combination of two loss functions as follows:

$$\epsilon_{MT} = \epsilon_S + \epsilon_E \tag{1}$$

where ϵ_{MT} is the total loss function of the multi-task DNN, which is estimated from a combination of two loss functions which are the loss functions of the speech recognition task ϵ_S and the loss function of the environmental sound recognition task ϵ_E. The error from both tasks will be back-propagated through the hidden layers of the network.

After the training is complete, in traditional multi-task learning, the portion of the network associated with the secondary tasks is discarded, and the network will perform only the primary task. In contrast, in this study, we treat both tasks as primary tasks that needed to be solved with equal priority. Therefore, there is no portion of the network that is discarded.

2.2. Bottleneck and Sound-dependent I-vector Features

To address the problems in direct multi-task learning of simultaneous speech and environmental recognition, we propose a method that combines bottleneck features and sound-dependent i-vectors.

- **Bottleneck Features**

  The bottleneck features are simply vectors consisting of the activations at a bottleneck layer, which has a relatively small number of hidden units compared to the other hidden layers in the network. As described earlier, we also proposed to use two parallel single-task DNNs and a single multi-task DNN for producing the bottleneck layer. Therefore, the difference of the structures with DNNs that are used for speech and environmental sound recognition is that it has one hidden layer with a relatively small number of hidden units compared to the other hidden layers.

  The loss function and training method are the same as those in previous single-task and multi-task learning approaches. The input can be taken from speech features only or through combination with i-vector features which will be described in next section. After the training is complete, the layers after the bottleneck features are discarded. As there are two bottleneck features produced when performing the speech or environmental recognition task, those features are stacked together into one set of bottleneck features and used as input for another single-task and multi-task DNN of speech and environmental sound recognition.

- **Sound-dependent I-vector Features**

  I-vectors are derived from traditional joint factor analysis (JFA) [24], where an utterance supervector for a speaker should be decomposable into speaker independent, speaker dependent, channel dependent, and residual components. However, instead of finding two separate vector subspaces which represent the speaker and channel variabilities as in JFA, Dehak et al. proposed to find a single low dimensional vector subspace of the GMM supervector space that represents both speaker and channel variabilities, which is known as i-vectors [25].

  As the i-vector approach was initially introduced for speaker recognition tasks, it is common to find a low dimensional i-vector that represents speaker variability

only. However, in this study, we extract i-vectors that represent environmental sound characteristics using the following equations:

$$M_s = m + Tw_s \tag{2}$$

where M_s is the utterance supervector that depends on environmental sound components, m is the mean supervector of a universal background model (UBM), T is a low rank rectangular total variability matrix, and w_s is the i-vector that represents a low dimensional vector subspace of environmental sound variabilities.

Given input speech feature frames X, the i-vector w_s can be defined by the mean of the posterior distribution $P(w_s|X)$, where this posterior distribution is a Gaussian distribution. This is a maximum a posteriori probability (MAP) estimate of w_s, while the matrix T is estimated using the expectation maximization (EM) algorithm [26].

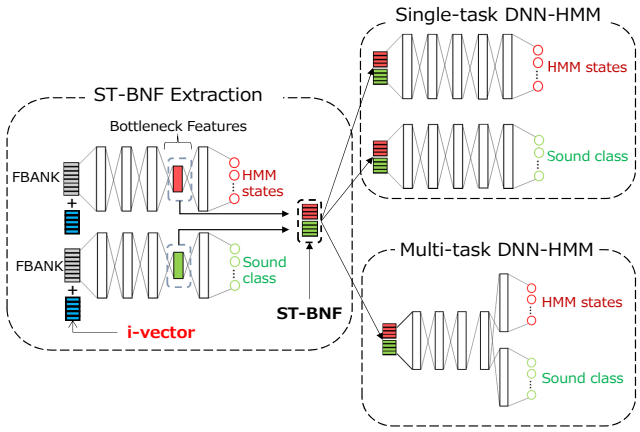
2.3. Overall Architecture

As we have DNNs that generate bottleneck features, and DNNs that recognize speech and environmental sounds, the overall architecture of the proposed approach consists of two main DNN components:

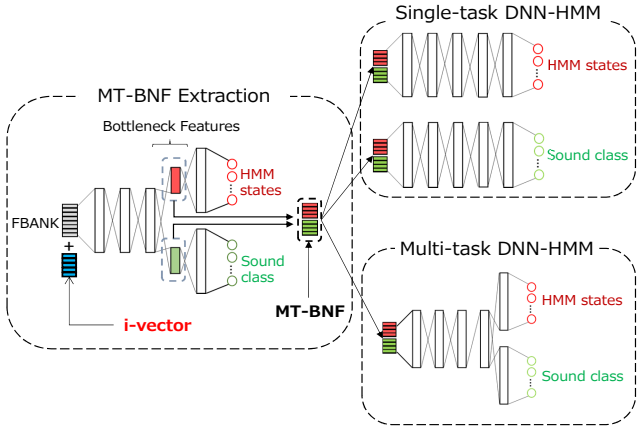
- **DNN-Extractor**: a DNN trained as an extractor that employ i-vectors and produce discriminative bottleneck features
- **DNN-Classifier**: a DNN trained as a classifier for the intended simultaneous recognition of speech and environmental sounds

Here, for both the first and second DNN-Extractor and -Classifier components, we investigate DNN architectures with two parallel single-task DNNs and a single multi-task DNN.

The combination of two DNN components with two types of architectures including i-vector features is illustrated in Fig. 1, where we have: (a) two parallel single-task DNNs that produce bottleneck features (ST-BNF) for speech and sound recognition respectively; and (b) a single multi-task DNN that produces bottleneck features (MT-BNF) for single-task and multi-task speech and sound recognition. The output from DNN-Classifier when performing speech recognition task will be used to estimate HMM emission probabilities within the DNN-HMM hybrid framework, while output from DNN-Classifier when performing environmental sound recognition will be used directly as environmental sound recognition output.



(a) *ST-BNF and i-vector for single-task and multi-task speech and sound recognition.*



(b) *MT-BNF and i-vector for single-task and multi-task speech and sound recognition.*

Fig. 1: *Overall DNN structure for simultaneous recognition of speech and environmental sound.*

3. EXPERIMENTAL SET-UP

3.1. Corpus Construction

- **Clean Speech Corpus**
  The clean speech corpus is derived from the Corpus of Spontaneous Japanese (CSJ)[27] which is a richly annotated speech and language database of spontaneous monologue speech (academic presentations, public speaking). It consists of approximately 7.5 million words, which were provided by more than 1,400 speakers of ages ranging from twenties to eighties. In this study, we used 20,000 utterance (approximately 34 hours of speech) for the training set, 1600 utterances (approximately 2.5 hours) for the development set, and 1000 utterances (approximately 1.5 hours) for the test set. The speech format is mono-channel with 16-kHz sampling frequency and 16-bit quantization.

- **Collection of Environmental Sound Data**
  The environmental sound data is collected from sound snap[1] – a web site providing data of various sound effects and background music (BGM). In sound snap, there are approximately 150 sound categories, each including many sound wav files and sound tag information. In this study, we only selected 10 environmental sound types of sound categories that often appear in our everyday life (car engine, horn, drill, alarm, bell, flow of river, keyboard, applause, whistle, crowd noise), and 18 different wav files in each category, where we use 10 files for training set, 4 files for the development set, and 4 files for the test set. The sound format is also mono-channel with 16-kHz sampling frequency and 16-bit quantization.

- **Mixing Speech and Environmental Sound**
  In order to have speech data with background environmental sounds, we mix the clean speech of the CSJ corpus with the selected environmental sound data using 1 sound file for 50 utterances in training set, 1 sound file for 40 utterances in development set, and 1 sound file for 10 utterances in the test set. Then we construct mixed speech and sound data with SNR of 0db, 5db, 10db, and 20db, resulting in 20,000 speech-sound files for the training set, 1600 speech-sound files for the development set, and 1000 speech-sound files for the test set, respectively.

3.2. Recognition System Set-up

Our speech recognition system is based on Kaldi [28], a free open-source toolkit for speech recognition research. To construct the DNN models, we use a Python toolkit for deep learning with Kaldi [29].

- **Front-End Processing**
  We trained the systems with a front-end based on 40-dimensional log-scale filter-banks (FBANK). The front-end provides features every 10ms with 25ms width. To incorporate the temporal structures and dependencies, 11 adjacent (center, 5 left, and 5 right) frames of FBANKs are stacked into one single feature vector leading to 440 dimensional super vectors (11x40 dimensions). These are then projected down to an optimal 40 dimensions by applying linear discriminant analysis (LDA). After that, the resulting features are further de-correlated using a maximum likelihood linear transformation (MLLT) [30], which is also known as the global semi-tied covariance (STC)[31] transform. Moreover, speaker adaptive training (SAT) [32] is performed using a single feature-space maximum likelihood linear regression (fMLLR) [33] transform estimated per speaker.

[1] Sound snap – http://www.soundsnap.com/

- **Baseline GMM-HMM Speech Recognition**
  Standard GMM-HMM acoustic models are trained on the provided features describe above. All models are context-dependent cross-word triphones with a standard three-state left-to-right HMM topology without skip states. The HMM units are derived from 39 phonemes of Japanese, with a total of 2,160 HMM triphone states. The pronunciation dictionary was constructed with the CSJ pronunciation dictionary of Japanese. The resulting pronunciation dictionary contains 50k words. Using the SRILM toolkit [34], we built trigram language models.

- **DNN Speech and Sound Recognition**
  Two parallel single-task DNNs and one multi-task DNN are constructed in parallel for speech and environmental sound recognition. The DNN topology consists of 5 fully connected hidden layers with 1,024 nodes in each layer, and a softmax output layer on the top. The output layer has 2,160 nodes for speech recognition corresponding to the number of HMM triphone states, and 10 nodes for environmental sound recognition.

- **Bottleneck Features**
  Similar to DNN-HMM acoustic modeling, two parallel single-task DNNs (ST-BNF) and one multi-task DNN (MT-BNF) with bottleneck hidden layer are constructed in parallel for speech and environmental sound recognition. Each bottleneck layer of speech and sound recognition has 42 nodes, resulting in a total of 84-dimension bottleneck features.

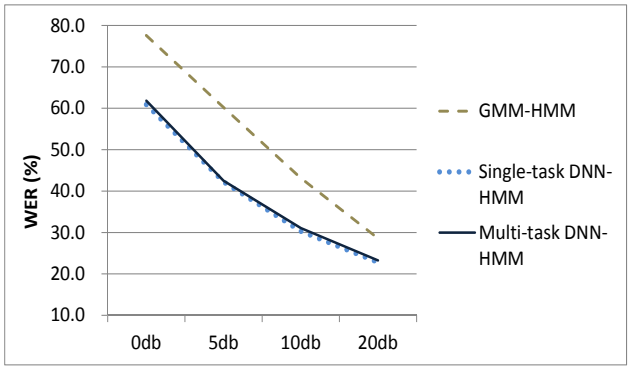
- **I-vector Features**
  GMM-UBM was learned using the same 20,000 mixed speech and sound files in training set. Then, we trained a 100-dimensional i-vector for each sound file.

4. RESULTS AND DISCUSSION

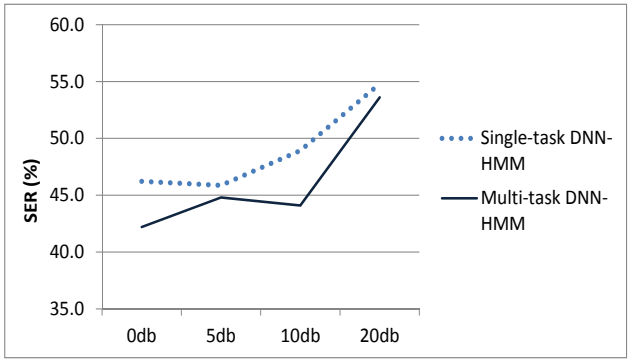
First, we construct a simple instantiation of our proposed approach in which we only use DNN-Classifier components without bottleneck features and i-vectors. In this case, the FBANK features are provided as input to the parallel single-task DNNs and the single multi-task DNN directly. Figure 2 shows their performance in comparison with the baseline GMM-HMM system, where: Fig. 2(a) shows the word error rate (WER) of the system for the speech recognition task, and Fig. 2(b) shows the sound error rate (SER) of the systems for the environmental sound recognition task.

From the results, we can see that the multi-task DNN give only similar performance with two parallel single-task DNNs in the speech recognition task, and even slightly lower performance in the environmental sound recognition task. This reveals that direct multi-task learning for speech and environmental sound recognition is difficult, and the system is likely

not able to effectively share mutual knowledge. These phenomena can be clearly seen where in 0dB SNR the WER is very high while the SER is low, but in 20dB the WER can be reduced while the SER increases. Nevertheless, in both systems, two parallel single-task DNNs and a single multi-task DNN outperformed the GMM-HMM baseline.



(a) *WER of DNN-HMM speech recognition in comparison with GMM-HMM baseline.*



(b) *SER of DNN environmental sound recognition.*

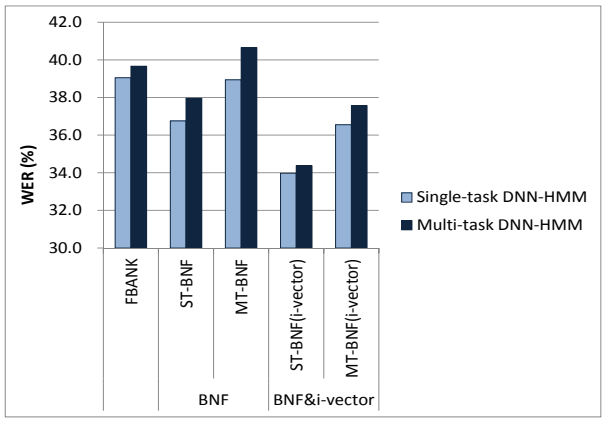
Fig. 2: *The performance of DNN-Classifier components with two parallel single-task DNNs and a single multi-task DNN.*

Next, we investigated the our full-fledged proposed system with the two DNN-Extractor and DNN-Classifier components, each constructed either with two parallel single-task DNNs or a single multi-task DNN, as illustrated in Fig. 1. The input of DNN-Classifier components includes:

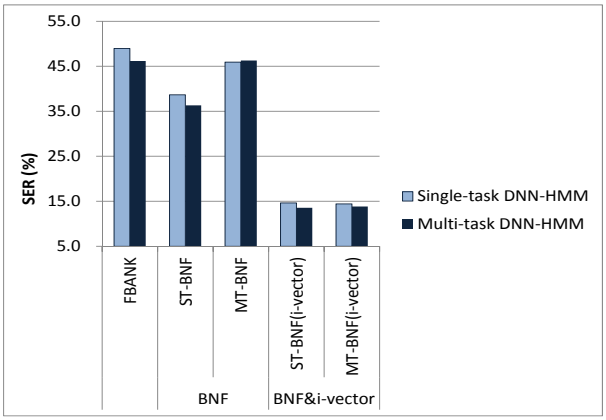
- FBANK
- Single-task bottleneck features (named as "ST-BNF")
- Multi-task bottleneck features (named as "MT-BNF")
- ST-BNF with FBANK plus i-vector input (named as "ST-BNF(i-vector)")
- MT-BNF with FBANK plus i-vector input (named as "MT-BNF(i-vector)")

With combination of various inputs and structure, we will have 10 different systems in total.

We performed experiments with SNR of 0db, 5db, 10db, and 20db as before, showing the difference in average error rate from the error rate of the baseline for each SNR. The recognition performance of all systems can be seen in Fig. 3, where: Fig. 3(a) shows the WER of the DNN-HMM systems for the speech recognition task, and Fig. 3(b) shows the SER of the DNN systems for the environmental sound recognition task.



(a) *WER of DNN-HMM speech recognition.*



(b) *SER of DNN environmental sound recognition.*

Fig. 3: *Bottleneck and i-vector features for single-task and multi-task speech and sound recognition*

The results reveal that the error rate reduces with bottleneck features, specially using ST-BNF. This shows that ST-BNF could provide a better discriminative features than the FBANK features. Furthermore, the error rate could be further reduced using i-vector features. The best system is provided by ST-BNF(i-vector) with single-task DNN-HMM reaching 5% absolute WER reduction from the standard single-task DNN-HMM and 18.4% absolute WER reduction from the GMM-HMM speech recognition baseline, while for

environmental sound recognition, the best system is provided by ST-BNF(i-vector) with multi-task DNN reaching 32.6% absolute SER reduction from DNN-Classifier without bottleneck features and i-vectors. Overall, the results may indicate that DNN-Extractor with bottleneck features and i-vector is able to reduce the variabilities that exist in mixed speech and sound, and extract only the important information for the next stage of DNN-Classifier.

5. RELATED WORKS

DNN-based methods have demonstrated superior performance on a number of natural language processing and spoken language processing tasks. Specifically with regards to ASR tasks, DNNs are mainly utilized either to directly estimate hidden Markov model (HMM) emission probabilities in the DNN-HMM hybrid approach [35], or in a tandem fashion that produce discriminative features for training Gaussian mixture models (GMMs) in the standard GMM-HMM framework [36]. These features are generated by the estimated target probabilities from DNN's output layer or the activations from a narrow hidden bottleneck layer (known as bottleneck features) [37]. Recently, Wu et al. showed that a DNN can be improved through the use of stacked bottleneck features combined with multi-task learning, but the approach was applied for the speech synthesis task [38]. In this study, we examine the use of DNNs in tandem-hybrid approach combinations, in which we use DNNs to produce bottleneck features, not for training the GMM-HMM model, but for training a DNN-HMM hybrid model within the framework of multi-task learning of both speech and environmental sound recognition.

The standard idea of multi-task learning in the area of ASR is to train a single neural network to perform a primary task, commonly the speech recognition task, plus at least one secondary task in parallel manner. It is well known that if the multiple tasks have some knowledge to share, learning them together can help to improve the generalization of the model and lead to an improvement in performance. Various related works include the use of multi-task DNNs for the phoneme recognition task with phone labeling, state context, or phone context learning tasks [39], triphone modeling with trigrapheme modeling tasks [40], noisy speech recognition with gender classification tasks [41], as well as multilingual speech recognition tasks [42, 43]. In most cases, the tasks have similar characteristic in which the secondary tasks contains information that may support the primary task. Our research, on the other hand, attempts to perform multi-task learning to perform speech recognition with environmental sound recognition. Similarly to multi-task learning for multilingual speech recognition, we treat both tasks as primary tasks that needed to be solved with equal priority. As discuss earlier, difficulty arises due to the fact that as additional environmental sounds get louder, it becomes easier to recognize environmental sounds, but more difficult to perform speech recognition.

To improve robustness to mismatched conditions in DNNs, one popular technique is using i-vector features as input to the model [44]. The most common way to do so is to use i-vectors for adapting a DNN to a specific speaker [45]. This way, DNNs are able to learn speaker-specific differences. However, recently, many studies also showed the advantages of involving i-vectors to handle other sources of variability, i.e., channel and noise-related variabilities [46, 47]. In this study, we employ i-vectors to capture environmental sound characteristics by combining bottleneck features and sound-dependent i-vectors within the DNN framework.

6. CONCLUSIONS

In this study, we investigated the possibility of recognizing speech and environmental sound simultaneously based on multi-task deep learning. Particularly, we investigated various DNN-HMM hybrid architectures, including two parallel single-task DNNs and a single multi-task DNN. Experimental evaluation results revealed that direct multi-task DNN training for simultaneous speech and environmental sound recognition is hard, and therefore utilizing bottleneck feature and i-vectors as the input of DNNs can help to improve accuracy of each recognition task. For speech recognition, the best system is provided by ST-BNF(i-vector) with a single-task DNN-HMM reaching 5% absolute WER reduction from standard single-task DNN-HMM and 18.4% absolute WER reduction from GMM-HMM speech recognition baseline, while for environmental sound recognition, the best system is provided by ST-BNF(i-vector) with a multi-task DNN reaching 32.6% absolute SER reduction from DNN-Classifier without bottleneck features and i-vectors.

7. ACKNOWLEDGEMENTS

Part of this research was supported by JSPS KAKENHI Grant Number 24240032 and 26870371.

8. REFERENCES

[1] American Speech-Language-Hearing Association (ASHA), "Hearing and balance," http://www.asha.org/public/hearing/, 2013.

[2] Richard Lyon, "Machine hearing: An emerging field," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131–139, 2010.

[3] E. C. Cherry, "Some experiments on the recognition of speech with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by additive noise," in *Proc. ICASSP*, Washington, USA, 1979, pp. 208–211.

[5] S. Kamath and P. Loizou, "A multiband spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. ICASSP*, Orlando, USA, 2002.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *Proc. ICASSP*, Tampa, USA, 1985, pp. 443–445.

[7] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed priors," in *Proc. ICASSP*, Orlando, USA, 2002, pp. 504–512.

[8] K. Paliwak and A. Basu, "A speech enhancement based on kalman filtering," in *Proc. ICASSP*, Dallas, USA, 1987, pp. 177–180.

[9] M. Gabrea, "Robust adaptive kalman filtering-based speech enhancement algorithm," in *Proc. ICASSP*, Montreal, Canada, 2004, pp. 301–304.

[10] X.-G. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto-encoder," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 436–440.

[11] Y. Xu, J. Du, L.-R. Dai, and Lee C.-H., "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[12] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE WASPAA*, New Paltz, USA, 2013.

[13] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, 2016.

[14] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[15] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, Scottsdale, USA, 2015.

[16] T.-H. Dat, N.-W. Terence, J.-W. Dennis, and Ren L.-Y., "Generalized Gaussian distribution Kullback-Leibler kernel for robust sound event recognition," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 5949–5953.

[17] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 2013, pp. 1–4.

[18] E. Cakir, T. Heittola, H. Huttunen, and Virtanen T., "Polyphonic sound event detection using multilabel deep neural networks," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN)*, Anchorage, Alaska, USA, 2015, pp. 1–7.

[19] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 559–563.

[20] I.E. Muller, "Scan anything and let your phone do the rest," Tech. Rep., MIT Technology Review, 2011.

[21] H.G. Okuno, T. Nakatani, and T. Kawabata, "Cocktail-party effect with computational auditory scene analysis," *Symbiosis of Human and Artifact (Elsevier)*, vol. 2, pp. 503–508, 1995.

[22] R. Morell, "Ability to listen to two things at once is largely inherited, says twin study," *Human Genetics*, 2007.

[23] M. Kashino and T. Hirahara, "One, two, many - judging the number of concurrent talkers," *Journal of Acoustical Society of America*, vol. 99, no. 4, 1996.

[24] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., CRIM, Montreal, Canada, 2005.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, pp. 788–798, 2010.

[26] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE*, vol. 13, no. 3, pp. 345–354, 2005.

[27] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, 2000, pp. 947–952.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Moticek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Hawaii, USA, 2011.

[29] Y. Miao, "Kaldi+PDNN: building DNN-based ASR systems with Kaldi and PDNN," arXiv:1401.6984, 2014.

[30] R. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. of ICASSP*, 1998, pp. 661–664.

[31] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[32] R. Schwartz T. Anastasakos, J. Mcdonough and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[33] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[34] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of ICSLP*, Denver, USA, 2002, pp. 901–904.

[35] H.-A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic, Norwell, MA, USA, 1993.

[36] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1635–1638.

[37] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. ICASSP*, Hawaii, USA, 2007, pp. 757–760.

[38] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 4460–4464.

[39] M.-L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 6965–6969.

[40] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 5592–5596.

[41] Y. Lu, F. Lu, S. Sehgal, S. Gupta, J. Du, C.-H. Tham, P. Green, and V. Wan, "Multitask learning in connectionist speech recognition," in *Proc. of the 10th Australian International Conference on Speech Science & Technology*, Sydney, Australia, 2004, pp. 312–315.

[42] 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. ICASSP*, Vancouver, Canada, 2013, pp. 8619–8623.

[43] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7304–7308.

[44] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, Olomouc, Czech Republic, 2013, pp. 55–59.

[45] P. Cardinal, N. Dehak, Y. Zhang, and J. Glass, "Speaker adaptation using the i-vector technique for bottleneck features," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2867–2871.

[46] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 225–229.

[47] S. Ganapathy, S. Thomas, D. Dimitriadis, and S. Rennie, "Investigating factor analysis features for deep neural networks in noisy speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1898–1902.