A GENERATIVE-DISCRIMINATIVE HYBRID APPROACH TO MULTI-CHANNEL NOISE REDUCTION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Hendrik Meutzner[§,‡] *Shoko Araki*[§] *Masakiyo Fujimoto*[§] *Tomohiro Nakatani*[§]

[§] NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237 Japan
[‡] Cognitive Signal Processing Group, Institute of Communication Acoustics
Ruhr-University Bochum, Universitätsstrasse 150, 44801 Bochum, Germany

ABSTRACT

In the recent years, discriminative models have become a very attractive utility and gained a lot of attention in the speech research community, encompassing both front and back-end methods, thanks to their prominent discriminative power and the availability of improved training strategies. When it comes to the recognition of speech that is distorted by highly non-stationary environmental noise, robust front and back-end methods are required in order to achieve a satisfactorily high speech recognition performance. Furthermore, when dealing with severe noise conditions, multi-channel front-end methods can be advantageous for suppressing environmental background noise, as compared to single-channel methods. In this work, we improve an existing multi-channel noise reduction approach, referred to as *DOminance-based Locational and Power-spectral cHaracteristics INtegration (DOLPHIN)*, by using a generative-discriminative hybrid model, that makes use of spatial and spectral features. We show that the proposed method outperforms the existing DOLPHIN approach, which is solely based on generative models, in terms of the word error rate reduction achieved on the CHiME-3 challenge data.

Index Terms— Speech recognition, multi-channel noise reduction, generative-discriminative hybrid models.

1. INTRODUCTION

When an automatic speech recognition (ASR) system captures a speech signal in noisy environments, the performance of ASR degrades due to the noise included in the captured signal. Although it has recently been shown that the use of deep neural network acoustic models (DNN-AMs) greatly improves the noise robustness of ASR, the performance is still not satisfactory under severe noise conditions [1, 2, 3, 4].

To mitigate such noise influence on ASR, a generative model-based (single-channel) noise reduction approach has been extensively studied [5, 6, 7, 8]. With this approach, a Gaussian mixture model (GMM), referred to as a speech GMM, is trained in advance on clean speech spectral features. Then, for given noisy speech, this approach estimates posteriors of Gaussians in the speech GMM that correspond to the unknown clean speech spectral features, and estimates the clean speech spectral features based on the estimated posteriors. This approach has shown to improve the ASR performance when we use a DNN-AM trained on clean speech features, but it is not the case when we use a more robust AM, namely a DNN-AM trained on multi-condition data [9].

A major problem that limits the performance of the above approach lies in its poor accuracy of estimating the clean speech Gaussian posteriors from the noisy speech. To solve this problem, a new approach, referred to as a generative-discriminative hybrid approach, has recently been proposed [9, 10], and shown to improve the ASR performance even with a multi-condition DNN-AM. With this approach, e.g., in [9], a DNN for Gaussian posterior estimation is used jointly with the speech GMM. The DNN is trained in advance so that it can estimate Gaussian posteriors in the speech GMM from noisy speech, and then from given noisy speech, clean speech features are estimated using Gaussian posteriors obtained using the DNN. Because the accuracy of the Gaussian posterior estimation is improved using the DNN, the accuracy of noise reduction as a whole is also improved.

In this paper, we propose to extend an existing generative model-based multi-channel noise reduction approach, referred to as *DOminance-based Locational and Power-spectral cHaracteristics INtegration (DOLPHIN)* [11], using the generative-discriminative hybrid approach. DOLPHIN utilizes not only the spectral features of the signals but also the spatial features of the signals that can be derived from multi-channel microphone signals. Based on both types of features, a set of parameters are estimated, which are referred to as soft masks and indicate whether speech is stronger than noise at individual time frequency bins, and used to estimate clean speech spectral features more accurately than a single channel generative approach. For the proposed generative-discriminative hybrid approach, we extend DOLPHIN by replacing its soft mask estimation block with a DNN based

estimation block, referred to as DNN-SME, to estimate the speech spectral features more accurately. We show that the extended DOLPHIN, referred to as DOLPHIN-DNN in this paper, outperforms DOLPHIN in terms of the word error rate (WER) reduction using the CHiME-3 challenge dataset.

2. RELATED WORK

Multi-channel linear filtering has been extensively studied as a technique to enhance noisy speech signals for the cases where multi-channel signals are available. In particular, dereverberation and beamforming based on linear filtering have been shown to be very effective to improve the robustness of ASR with multi-condition DNN-AMs [12, 13]. However, even after performing such filtering, certain amount of residual noise inevitably remains, and substantially limits the improvement of the ASR performance. For reducing such noise, nonlinear noise reduction needs to be applied, which generally degrade the performance of ASR with multi-condition DNN-AMs. In contrast, as shown in this paper, DOLPHIN-DNN can improve the performance of ASR with multi-condition DNN-AMs by nonlinear noise reduction, and thus could be complementary with the multi-channel linear filtering techniques.

Denoising autoencoder (DAE) could also be an alternative to DOLPHIN-DNN as a method that performs nonlinear noise reduction for single and multi-channel audio signals [14, 15]. However, based on our best knowledge, it has been reported that the ASR performance improvement by DAE is very limited when we use ASR with multi-condition DNN AMs [9].

3. IMPROVING DOLPHIN USING A GENERATIVE-DISCRIMINATIVE HYBRID APPROACH

One crucial step of the DOLPHIN algorithm is given by the estimation of spectral masks, referred to as posteriors of dominant source indices (DSI) in [11], which is—in its original form—based on a generative model approach, using spatial and spectral features. Hence, the overall performance of DOLPHIN is naturally defined by the accurateness of the estimated spectral masks.

As discriminative model approaches have proven to be superior for various class discrimination problems as compared to generative approaches (cf. [9]), we propose to replace DOLPHIN's generative mask estimation algorithm by a DNN, in order to make ASR more robust under severe noise conditions when used in combination with DOLPHIN as a front-end method.

3.1. System Overview

Figure 1 provides a fundamental overview of the DOLPHIN system. Due to the space limitations of this paper, we refer

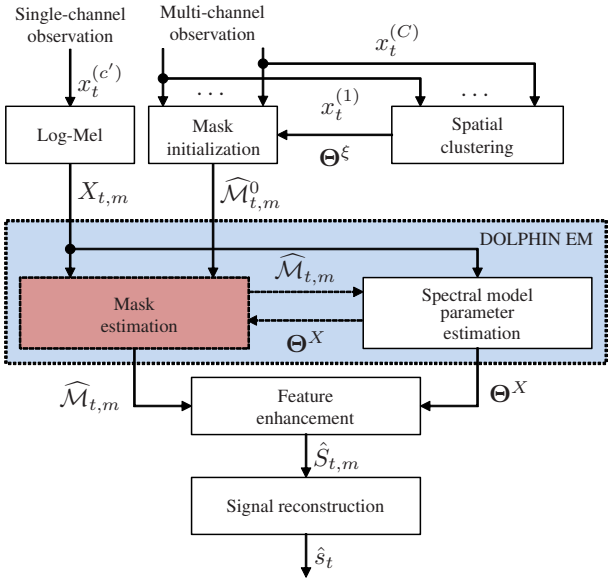

Fig. 1: Overview of the original DOLPHIN system.

the interested reader to [11] for a detailed discussion of the original DOLPHIN approach.

The single-channel signal $x_t^{(c')}$ represents the noisy input to DOLPHIN that shall be enhanced, with t denoting the frame time, and it is simply given by the observation signal of a single (pre-selected) microphone channel c. $x_t^{(c')}$ is then transformed into the log-Mel spectral domain, yielding $X_{t,m}$, where m is the index of the Mel frequency band[1].

At the same time, a spatial clustering is applied to the multi-channel noisy observation $x_t^{(c)}$ that yields the parameters of a generative spatial model $\boldsymbol{\Theta}^{\xi}$ as well as an initialization for the Mel spectral soft mask $\widehat{\mathcal{M}}^0_{t,m}$.

The soft mask $\widehat{\mathcal{M}}_{t,m}$ and the parameters of the spectral model $\boldsymbol{\Theta}^{X}$ are then iteratively updated by means of the expectation-maximization (EM) algorithm and used for enhancing the noisy observation $X_{t,m}$. The enhanced features $\hat{S}_{t,m}$ are then used for reconstructing the single-channel signal $\hat{s}_t$, which represents the noise reduced version of $x_t^{(c')}$. A detailed description of the parameter estimation procedure of the DOLPHIN algorithm is given in [11].

3.2. DOLPHIN-DNN

Several methods for estimating time-frequency masks by means of DNNs have been reported (e.g., [16, 17]). Inspired by these methods, we propose DOLPHIN-DNN that represents an improvement of the existing DOLPHIN approach.

[1]It should be noted that DOLPHIN operates in a lower-dimensional feature space for efficiency reasons. However, our investigations do not indicate any significant performance difference for DOLPHIN when using Mel features instead of an uncompressed linear frequency transform.

3.2.1. Soft Mask Estimation by DNN-SME

Let $\mathbf{x}_t$ denote the vector representation of all M log-Mel coefficients of the noisy observation at time t, i.e.,

$$\mathbf{x}_t = \begin{bmatrix} X_{t,1} & X_{t,2} & \dots & X_{t,M} \end{bmatrix}, \quad (1)$$

and defining further the vector representation of the masked version of $X_{t,m}$

$$\tilde{\mathbf{x}}_t = \begin{bmatrix} \tilde{X}_{t,1} & \tilde{X}_{t,2} & \dots & \tilde{X}_{t,M} \end{bmatrix}, \quad (2)$$

where,

$$\tilde{X}_{t,m} = X_{t,m} + \log\left(\widehat{\mathcal{M}}^0_{t,m}\right), \quad (3)$$

the extended feature that makes use of spectral and spatial information can then be conveniently defined as

$$\bar{\bar{\mathbf{x}}}_t = \mathbf{x}_t \,||\, \tilde{\mathbf{x}}_t, \quad (4)$$

where $||$ denotes vector concatenation. Here, $\tilde{\mathbf{x}}_t$ can be interpreted as an auxiliary feature, whose effect has been confirmed in e.g., [15, 18]. Furthermore, it is worth to note that using the masked version of $X_{t,m}$ is equivalent to using $\widehat{\mathcal{M}}^0_{t,m}$ directly, but we have found that the DNN training is more stable when using the above approach. Using Eq. (4), the input layer of the DNN reads

$$\mathbf{z}^{\text{IN}} = \begin{bmatrix} \bar{\bar{\mathbf{x}}}_{t-W} & \cdots & \bar{\bar{\mathbf{x}}}_t & \cdots & \bar{\bar{\mathbf{x}}}_{t+W} \end{bmatrix}, \quad (5)$$

with W being the context window length. The hidden layers are defined by

$$\mathbf{h}_l(\mathbf{z}) = \sigma\left(\mathbf{W}_l \mathbf{h}_{(l-1)}(\mathbf{z}^{\text{IN}}) + \mathbf{b}_l\right), \quad (6)$$

with l denoting the layer index and $\mathbf{h}_0$ is the identity function. $\sigma(\cdot)$ represents the Sigmoid function and $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and the bias vector. The output layer reads

$$\mathbf{z}^{\text{OUT}} = \sigma\left(\mathbf{W}_L \mathbf{h}_{(L-1)}(\mathbf{z}^{\text{IN}}) + \mathbf{b}_L\right), \quad (7)$$

which represents the Mel frequencies of the desired soft mask estimate $\widehat{\mathcal{M}}$ at time t.

3.2.2. Parameter Estimation

For training the DNN we use the minimum mean square error criterion (MMSE) where the loss function is given by the difference of the values of the output layer and the values of a precomputed ideal binary mask (IBM), i.e.,

$$\mathbb{E}^{\text{MMSE}} = \sum_t \sum_m \left(\widehat{\mathcal{M}}_{t,m} - \widetilde{\mathcal{M}}_{t,m}\right)^2. \quad (8)$$

The IBM is computed by comparing the power of the clean speech $S_{t,m}$ with the power of the isolated background noise $N_{t,m}$ for each time frame t and each Mel frequency band m

$$\widetilde{\mathcal{M}}_{t,m} = \begin{cases} 1-\epsilon, & \text{if } S_{t,m} \geq N_{t,m}, \\ \epsilon, & \text{otherwise}, \end{cases} \quad (9)$$

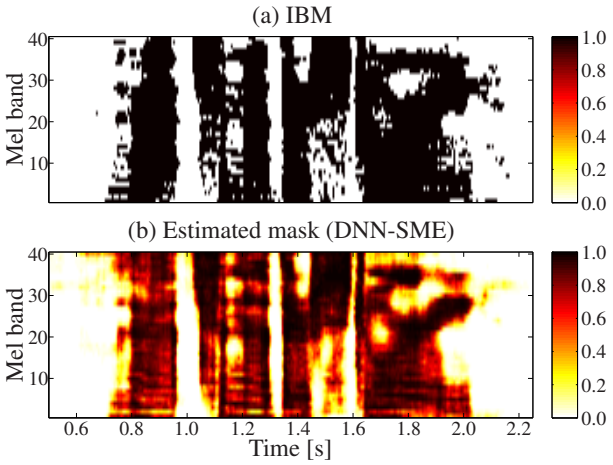

Fig. 2: Comparison of spectral masks.

where ϵ represents a scalar positive constant that allows to adjust the mask suppression level and we set $\epsilon = 0.01$. The parameters of the DNN are then estimated by means of the well-known backpropagation algorithm using the gradient of Eq. (8). Figure 2 shows an example of the IBM (a) and the soft mask estimated by DNN-SME (b).

4. EVALUATION

The proposed approach is evaluated in terms of the ASR performance achieved on the challenge data of the 3rd CHiME Speech Separation and Recognition Challenge [4].

4.1. CHiME-3 Corpus Description

The CHiME-3 corpus comprises real speech recordings that were created by using a 6-channel microphone array attached to a tablet device. The recordings were obtained within four different noisy everyday environments, i.e., public transport (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR), and they feature several male and female speakers, uttering the Wall Street Journal (WSJ) [19] sentences.

The corpus consists of real and simulated data, abbreviated as `real` and `simu` in the ensuing discussion, where the latter has been generated by artificially mixing the clean WSJ utterances with the environmental noise recordings.

The corpus is divided into 3 individual subsets, i.e., a training set (`tr_s`), containing 8738 noisy utterances (1600 `real` + 7138 `simu`), a development set (`dt_05`), containing 3280 noisy utterances (1640 `real` + 1640 `simu`), and an evaluation set (`et_05`), containing 2640 noisy utterances (1320 `real` + 1320 `simu`). Each of these subsets contains the same number of utterances for each individual environment (BUS, CAF, PED, and STR).

Table 1: Word error rates in percent for the Noisy signal, for DOLPHIN using a conventional generative SME (GSME), for DOLPHIN using the proposed discriminative SME (DSME), and for DOLPHIN using ideal binary masks (IBM). The results are shown for the development set (dt_05) and the evaluation set (et_05).

		simu					real					
Type	Set	BUS	CAF	PED	STR	Av.	BUS	CAF	PED	STR	Av.	Av. (all)
Noisy	dt_05	8.48	10.56	6.42	7.51	8.24	14.00	7.94	6.03	8.05	9.01	8.62
IBM	dt_05	5.18	6.19	3.97	5.29	5.16	-	-	-	-	-	5.16
GSME	dt_05	8.48	9.57	6.53	7.88	8.12	12.13	8.79	6.15	8.02	8.77	8.44
DSME	dt_05	7.99	7.74	5.90	7.27	7.23	11.05	6.37	5.49	6.83	7.44	7.33
Noisy	et_05	8.37	11.69	9.86	10.78	10.17	22.55	16.21	12.89	10.74	15.60	12.89
IBM	et_05	5.38	7.15	7.00	7.49	6.76	-	-	-	-	-	6.76
GSME	et_05	7.83	9.69	12.50	14.18	11.05	16.21	13.45	13.45	11.64	13.69	12.37
DSME	et_05	6.65	8.93	11.69	12.18	9.86	15.31	11.36	11.83	10.59	12.27	11.07

4.2. Speech Recognition Backend

The speech recognizer that is used to evaluate the proposed approach is based on a multi-condition DNN-AM and a recurrent neural network (RNN) language model in addition to a trigram language model. A detailed description of the system is given as that for the 1-pass SI system in [20].

4.3. Experimental Setup

All spectral input quantities, i.e., the noisy observation X and the initial mask $\widehat{\mathcal{M}}^0$, are computed by using a 40-dimensional Mel filterbank, where we utilize a frame length of 25 ms and a frame shift of 6.25 ms during signal analysis. For the single-channel noisy input signal $x^{(c')}$ we use $c' = 5$. The initial mask $\widehat{\mathcal{M}}^0$ is estimated by using a complex GMM based clustering approach [20].

The DNN-SME uses a context window size of $W = 5$, which results in an overall number of 880 units (11 frames $\times$ 40 log-Mel coefficients $\times$ 2 feature types) of the input layer for the given filterbank dimension. We train the DNN-SME by using the simu data of the tr_s set, where the simu data of the dt_05 set is used for cross-validation. As we found that there is no performance improvement for more than one hidden layer (i.e., setting $L > 2$), we present our results for $L = 1$. The number of units in the hidden layer is set to 1024, and the output layer consists of 40 dimensions, which is the number of Mel frequencies of the soft mask.

5. RESULTS

The speech recognition performance is measured in terms of the word error rate (WER).

Table 1 shows the results for the case of no additional signal enhancement (Noisy) as well as for the case when DOLPHIN makes use of the IBM, i.e., when no model for mask estimation is used[2]. The results for DOLPHIN using the conventional generative soft mask estimation are indicated by GSME and the proposed generative-discriminative hybrid approach using the DNN-SME is shown in the last row (DSME).

The conventional DOLPHIN approach (GSME) reduces the word error rate for most of the environments, where the relative average word error rate reduction is 2 % for dt_05 and 4 % for et_05, when comparing the results to the noisy case. The effectiveness of the DOLPHIN approach for ASR can be seen when DOLPHIN makes use of IBMs, as the relative average word error rate reduction is more than 40 % for both data sets.

For the proposed discriminative SME (DSME), the performance of DOLPHIN improves as compared to GSME for each presented case. The relative average word error rate reduction is 15 % and 14 % for dt_05 and et_05, respectively, when comparing DSME with Noisy.

6. CONCLUSIONS

In this work we have shown that a generative-discriminative hybrid approach that incorporates a DNN-SME into DOLPHIN, is beneficial for a multi-condition noise reduction task. Comparing the performance with the conventional approach, which is solely based on generative models, the proposed approach yields a relative word error rate reduction of 15 % on the CHiME-3 challenge dataset.

It was shown that dereverberation and beamforming are also effective to reduce the WER for CHiME-3 [20], and thus the future work will cover the impact of the combination of these methods with DOLPHIN-DNN.

[2]The IBM can only be computed for the simu data using the provided annotation files that contain the information about the temporal placement of the WSJ utterances in the background noise recordings.

7. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. 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] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?," in *Proc. Interspeech'13*, August 2013, pp. 2992–2996.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

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

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. ICASSP'96*, May 1996, vol. II, pp. 733–736.

[6] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP 2000*, October 2000, pp. 806–809.

[7] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Interspeech'03*, September 2003, pp. 1009–1012.

[8] M. Fujimoto, K. Ishizuka, and T. Nakatani, "Study of integration of statistical model-based voice activity detection and noise suppression," in *Proc. Interspeech'08*, September 2008, pp. 2008–2011.

[9] M. Fujimoto and T. Nakatani, "Feature enhancement based on generative-discriminative hybrid approach with GMMs and DNNs for noise robust speech recognition," in *Proc. ICASSP'15*, April 2015, pp. 5019–5023.

[10] S. Liu and K. C. Sim, "Joint adaptation and adaptive training of TVWR for robust automatic speech recognition," in *Proc. Interspeech'14*, September 2014, pp. 636–640.

[11] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2516–2531, December 2013.

[12] S. Renals and P. Swietojansk, "Neural networks for distant speech recognition," in *Proc. Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 172–176.

[13] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognition in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, pp. 1–15, July 2015.

[14] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. ICASSP'14*, May 2014, pp. 4623–2627.

[15] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. ICASSP'15*, May 2015, pp. 116–120.

[16] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition.," in *Proc. ICASSP'13*, September 2013, pp. 7092–7096.

[17] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation.," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, July 2013.

[18] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP'13*, September 2013, pp. 7398–7402.

[19] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, Stroudsburg, PA, USA, 1992, HLT'91, pp. 357–362, Association for Computational Linguistics.

[20] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU'15*, December 2015.