INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Far-Field Speech Enhancement using Heteroscedastic Autoencoder for Improved Speech Recognition

Shashi Kumar, Shakti P. Rath

Samsung Research Institute India - Bangalore
{sk.kumar, shakti.rath}@samsung.com

Abstract

Automatic speech recognition (ASR) systems trained on clean speech do not perform well in far-field scenario. Degradation in word error rate (WER) can be as large as 40% in this mismatched scenerio. Typically, speech enhancement is applied to map speech from far-field condition to clean condition using a neural network, commonly known as denoising autoencoder (DA). Such speech enhancement technique has shown significant improvement in ASR accuracy. It is a common pratice to use mean-square error (MSE) loss to train DA which is based on regression model with residual noise modeled by zero-mean and constant co-variance Gaussian distribution. However, both these assumptions are not optimal, especially in highly non-stationary noisy and far-field scenario. Here, we propose a more generalized loss based on non-zero mean and heteroscedastic co-variance distribution for the residual variables. On the top, we present several novel DA architectures that are more suitable for the heteroscedastic loss. It is shown that the proposed methods outperform the conventional DA and MSE loss by a large margin. We observe relative improvement of 7.31% in WER compared to conventional DA and overall, a relative improvement of 14.4% compared to mismatched train and test scenerio.

Index Terms: distant speech recognition, parallel data, speech enhancement, autoencoder, homoscedastic, heteroscedastic

1. Introduction

Recent breakthrough in deep learning based acoustic model for automatic speech recognition (ASR) has shown substantial improvement in accuracy. However, in practical real-world scenario, noise can be a detrimental factor that can degrade performance considerably. Noise can arise either due to interfering background events (such as babble or street noise), or reverberation due to non-ideal room acoustics that can cause multiple reflected replicas of speech to arrive at the microphone distorting the speech characteristics that are considered useful for ASR. When ASR is trained in clean condition and tested on far-field speech, the performance degrades substantially due to mis-match in train and test conditions. A common technique is to enhance the distant speech such that the mis-match is reduced, which is known as speech enhancement [1, 2, 3, 4].

Many approaches have been explored for speech enhancement. Spectral magnitude based approaches include estimating inverse-filter to cancel the effect of the late reverberation [2, 5] and non-negative matrix factorization based methods [6]. Domain adaptation is another such important work in this direction where acoustic model is trained on data from a source distribution and testing is done on data from a target distribution [7, 8, 9]. The most dominant works in speech enhancement include mapping characteristics of speech from source domain to target domain which are commonly known as denoising autoencoder (DA). Some of the works in this direction are spectral mask estimation [10, 11] and feature domain transformation [9, 3, 4, 12]. Phase based mask estimation has also been explored for dereverberation [13, 11, 14] as phase based mask is shown to improve perceptual quality of noisy speech [15]. Similarly, speech enhancement methods have also been explored for speaker recognition in far-field scenario [16]. In [17], it has been explored for a large vocabulary ASR task which uses mapping of bottleneck features from source to target domains.

In this paper, we focus on DA in the feature domain. It is a common practice to use mean-square error (MSE) to train DA whose foundation is based on the statistical regression problem. This idea of using MSE can be dated back to [1]. In this setting, the residual noise involved in the regression function is modeled by a zero-mean and constant co-variance Gaussian distribution. The zero-mean assumption is based on the hypothesis that the regression function is sufficiently powerful to model the "true" generative process and the constant co-variance assumption signifies that the residual noise is homoscedastic. In more challenging far-field scenario, the mapping from reverberant to clean speech can be highly non-linear and time-variant that can not be optimally represented by a stationary DA whose parameters are fixed (after convergence). Thus, it is more appropriate to introduce a time-varying component in the prediction model that can handle the non-stationarity better than the conventional approach. With this aim, in this paper, we alleviate the homoscedastic assumption and propose a heteroscedastic (meaning co-variance of acoustic features is dependent on time) noise distribution model with non-zero mean. The mathematical details, including the formulation for maximum likelihood estimation and optimal prediction model are presented. On the top several novel architectures are presented that we found to be more suitable for proposed heteroscedastic loss. Heteroscedastic loss has been explored earlier in other non-speech domains [18, 19]. We show that the proposed heteroscedastic loss and DA architectures outperforms the conventional DA and MSE loss by a large margin. We observe relative improvement of 7.31% in WER compared to conventional DA and overall, relative improvement of 14.4% compared to mismatched scenerio.

The rest of the paper is organized as follows. In Section 2 we review conventional denoising autoencoder giving out the detailed mathematical formulation. In Section 3, various proposed heteroscedastic losses are described. Section 4 describes the neural network architectures that are developed in this paper keeping the proposed loss in mind. Experimental set-up and results are presented in Section 5. Finally conclusions are presented in Section 6.

2. Conventional Denoising Autoencoder

In this section, we formulate the denoising autoencoder (DA) approach to speech enhancement for distant speech recognition. In the DA, it is assumed that acoustic feature vectors in close-

http://dx.doi.org/10.21437/Interspeech.2019-2032

talk (target) domain are generated by an unknown stochastic and non-linear mapping of the features from the far-field (source) domain. Mathematically, $y_n = \mathcal{F}(\mathbf{x}_n)$, where $\mathbf{x}_n$ and y_n denote the acoustic features in the far-field and close-talk domains at time instance n, respectively. Since $\mathcal{F}$ is not known, we recourse to regression framework, where the true generative function $\mathcal{F}$ is approximated by a parameterized function f, typically a neural network, which is a deterministic mapping between source domain to target domain. In more details,

$$y_n = f(\mathbf{x}_n, \mathbf{w}) + \epsilon_n \tag{1}$$

where $\mathbf{w}$ denotes the weights of the neural network. In Eq. 1, ϵ_n represents an additive residual term that models the inaccuracies in the regression function, f.

2.1. Homoscedastic Loss Formulation

Typically in conventional DA, the residual term is assumed to be

$$\epsilon_n \sim \mathcal{N}(0, \beta) \tag{2}$$

where β denotes the co-variance. Note that there are two approximations here. Firstly the co-variance, β, is assumed to be constant for all time instances, which is known as homoscedastic condition, and secondly the zero-mean assumption, which is based on the hypothesis that the regression model is powerful enough to closely approximate $\mathcal{F}$. Under this setting, the likelihood function for y_n for a single frame is given by

$$p(y_n|\mathbf{x}_n, \mathbf{w}) = \mathcal{N}\left(y_n|(f(\mathbf{x}_n, \mathbf{w})), \beta\right) \tag{3}$$

Now considering a data set of matched inputs and outputs $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ with corresponding target values $\mathbf{y} = \{y_1, \cdots, y_N\}$ and making use of the standard form for the univariate Gaussian with β co-variance and assuming the data points are drawn independently, the likelihood function after taking log becomes

$$\begin{aligned}\ln(p(\mathbf{y}|\mathbf{X}, \mathbf{w})) &= \sum_{n=1}^{N} \ln(\mathcal{N}(y_n|(f(\mathbf{x}_n, \mathbf{w})), \beta)) \\ &= -\frac{N}{2}\ln 2\pi - \sum_{n=1}^{N}\frac{1}{2}\ln\beta - \sum_{n=1}^{N}\frac{E_D(\mathbf{w})}{2\beta}\end{aligned} \tag{4}$$

where the squared error is defined by

$$E_D(\mathbf{w}) = |y_n - (f(\mathbf{x}_n, \mathbf{w}))|^2 \tag{5}$$

Maximizing log-likelihood is equivalent to minimizing negative of log-likelihood. After removing the terms that are independent of $\mathbf{w}$, the final loss we optimize is given by

$$\mathcal{L}oss^{Homo} = \frac{1}{N}\sum_{n=1}^{N}|y_n - f(\mathbf{x}_n, \mathbf{w})|^2 \tag{6}$$

which leads to the the standard MSE loss typically optimized in conventional DA. It is also noted that the scaling due to co-variance term is absorbed into the constant term and do not appear in the loss we actually optmize.

Now, coming to the decision function for regression problem, when the loss to be optimized is the expected loss [20], the optimal prediction for a new vector $\mathbf{x}$ is given by the conditional mean of the target variable, i.e.,

$$\mathrm{E}[y|\mathbf{x}] = \int yp(y|\mathbf{x})dy \tag{7}$$

In the case of homoscedastic assumption of ϵ (Eq. 2), the predicted output, given in Eq. 7, simplifies to

$$\mathrm{E}[y|\mathbf{x}] = f(\mathbf{x}, \mathbf{w}) \tag{8}$$

which gives rise to the conventional "forward-pass" typically used in the case of neural-network based DA. It processes far-field features and generates the "cleaned" version frame-by-frame. To train the DA, it is a common pratice to assume that parallel data is available, i.e., for every $\mathbf{x}_n$ we know the time synchronous y_n, then the MSE loss defined in Eq. 6 is minimized to obtain the optimal parameters $\mathbf{w}^*$. The maximum likelihood estimate for co-variance β is given by

$$\beta_{ML} = \frac{1}{N}\sum_{n=1}^{N}|y_n - f(\mathbf{x}_n, \mathbf{w}^*)|^2 \tag{9}$$

3. Proposed Heteroschedastic Loss

In this paper we explore a simple yet more general model for the residual term, given by

$$\epsilon_n \sim \mathcal{N}(\mu_n, \beta_n) \tag{10}$$

which uses time-dependent (non-zero) mean and co-variance. This representation is referred to as heteroscedastic regression model. Under this condition, the log-likelihood function is given by

$$\ln(p(\mathbf{y}|\mathbf{X})) = -\frac{N}{2}\ln 2\pi - \sum_{n=1}^{N}\frac{E_D(\mathbf{w})}{2\beta_n} - \sum_{n=1}^{N}\frac{1}{2}\ln\beta_n \tag{11}$$

where the squared error is now defined by

$$E_D(\mathbf{w}) = |y_n - (f(\mathbf{x}_n, \mathbf{w}) + \mu_n)|^2 \tag{12}$$

The negative of log-likelihood function is given by

$$\mathcal{L}oss^{Hetero} = \frac{1}{N}\sum_{n=1}^{N}\frac{E_D(\mathbf{w})}{\beta_n} + \frac{1}{N}\sum_{n=1}^{N}\ln\beta_n \tag{13}$$

In this most general setting, this is the form of loss we optimize in this paper, where the mean and co-variance are also predicted by neural networks along with the cleaned features.

Following similar analysis as with the case of homoscedastic loss for decision function, the prediction model (forward-pass) for heteroscedastic loss is given by

$$\mathrm{E}[y_n|\mathbf{x}_n] = f(\mathbf{x}_n, \mathbf{w}) + \mu_n \tag{14}$$

which includes a mean compensation term besides the forward-pass. As with MSE loss, parallel data is used for training. In the following section we present a number of novel neural network architectures which are more suitable for proposed generalized loss. Common in all, these networks take far-field features as input and predict mean, variance and "cleaned" features simultaneously.

It may be noted from Eq. 13 that the term μ_n can possibly introduce a bias in the predicted outputs that may lead to non-unique solutions while training the neural-network. To account for this issue, we introduce a regularization term in loss, which is given by

$$\mathcal{L}oss^{Hetero} = \frac{1}{N}\sum_{n=1}^{N}\frac{E_D(\mathbf{w})}{\beta_n} + \frac{1}{N}\sum_{n=1}^{N}\ln\beta_n + \lambda\frac{1}{N}\sum_{n=1}^{N}\mu_n^2 \tag{15}$$

A weight λ is applied to the regularization term. The idea is to force the magnitude of μ_n to be as small as possible, while ensuring that the nature of heteroscedasticity remains intact.

4. Proposed Deep Neural Network Architectures

In this Section we present several novel neural network architectures that take far-field features as input and predict mean, variance and close-talk feats.

4.1. Basic Architecture

The first architecture consists of a deep neural network (DNN) with shared lower layers and three parallel task-dependent layers to predict close-talk features as well as the mean and variance given far-field features as input. Similar architectures have also been explored previously [21, 22]. The shared layers use ReLU as activation function whereas activation used for mean and close-talk layers is linear and variance is softplus. The network is trained using regularized heteroscedastic loss defined in Eq. 15. Unfortunately, the results yielded with this basic architecture were not promising, so they are not reported. However, for the matter completeness of the paper, we decided to describe the architecture here.

4.2. ParallelNet

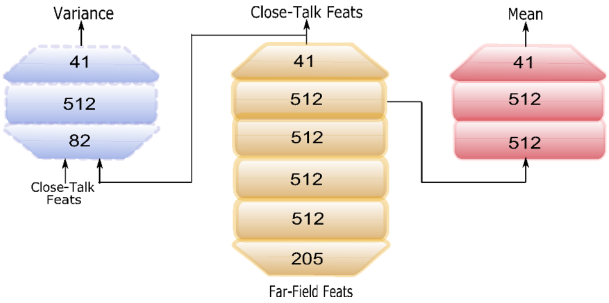

Figure 1: *ParallelNet architecture. Block shown in dotted line is not used at test time*

In this proposed architecture (shown in Fig. 1), separate neural networks are used for prediction of close-talk features, mean and variance. As shown, the block that predicts the variance takes concatenated close-talk and predicted close-talk features as input. It may be noted that the prediction model described in Eq. 14 does not include the variance, β_n, so at the test time the block that predicts the variance becomes inoperational and thus can be removed. The activations for individual blocks are same as previous architecture. We use loss function deduced in Eq. 15 to train this network.

4.3. ParallelNet: with only variance

It is noted that ParallelNet described above introduces additional computation in the forward-pass as it involves mean μ_n in the prediction model. The architecture shown in Fig. 2 is based on the case when the distribution of the residual term is $\mathcal{N}(0, \beta_n)$, i.e., zero-mean with heteroscedastic co-variance. As a result, the prediction model reduces to Eq. 8, therefore no mean-compensation happens at test time. Additional benefit of this network is that variance prediction model could be made more powerful without increasing the computational overhead in test time as the variance prediction network becomes inactive during testing. Consequently, better variance prediction may encourage a more generalized prediction model for close-talk features. This network is trained using loss defined in Eq. 13 where squared error $E_D(\mathbf{w})$ is defined by Eq. 5.

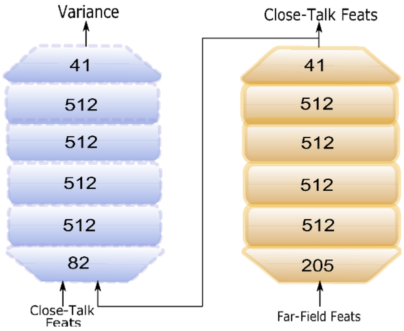

Figure 2: *ParallelNet with only variance. Block shown in dotted line is not used at test time*

5. Experiments and Results

5.1. Experimental setup

To evaluate the proposed loss functions and architectures, experiments were performed on AMI corpus [23, 24]. The AMI data set is meeting speech corpus which contains 100 hours of conversational non-native English speech. This corpus is recorded in specially equipped meeting rooms with individual head microphones (IHM, close-talk) lapel microphones and one or more distant microphones. Acoustic signals from different microphone sources are time aligned with beamforming. For the purpose of the work described here, we take aligned recordings from IHM and first distant microphone, referred to as single distant microphone (SDM). We assume there is an unknown deterministic function $\mathcal{F} : SDM \rightarrow IHM$. We estimate this function using a neural network which is trained by optimizing various losses formulated in this paper using parallel data available in the AMI dataset. In general, this is a general mapping from SDM features to IHM features using deep neural network without posing any further assumptions about reverberation condition, room noise type, room acoustics. Once this mapping model is trained, at test time, we first enhance SDM test audios by forward passing the far-field features through this speech enhancing model and then decode using ASR trained on IHM data. We report the word error rate (%WER) on standard dev set which is created by following Kaldi standard recipe for AMI corpus, it labels around 80 hours of data as training corpus and around 8 hours of data as standard dev set. For the extent of this work, we use Kaldi [25] for GMM-HMM training and pytorch for training our proposed novel architectures by optimizing proposed loss functions. We followed standard Kaldi recipe for AMI corpus to train LDA-MLLT-SAT GMM-HMM model using IHM data. We used this model to generate senone alignment for acoustic model training.

Table 1: *Baseline results in close-talk and far-field scenerio*

System	**Scenerio**	**WER**(%)
LSTM-HMM	IHM	29.4
LSTM-HMM	SDM	70.03

Table 2: *ParallelNet enhancement results using different prediction models*

Front-End	**Prediction**	**WER**(%)
Homoscedastic	$f(\mathbf{x}_n, \mathbf{w})$	64.67
ParallelNet	$f(\mathbf{x}_n, \mathbf{w}) + \mu_n$	**59.94**
ParallelNet	$f(\mathbf{x}_n, \mathbf{w})$	59.98
ParallelNet with only Variance	$f(\mathbf{x}_n, \mathbf{w})$	60.73

The ASR acoustic model is a Long Short Term Memory Hidden Markov Model (LSTM-HMM) system, which is trained using 41-dimensional log mel-filter bank features with ±2 splicing. The model consists of three LSTM layers with 512 cells each. Left context of 20 frames is used for better cell state initialization before processing a sequence of 30 frames. We used Adam with learning rate varying from 0.001 to 0.0001 following an exponential learning rate scheduler and trained the model for 20 epochs and then for further 5 epochs with fixed final learning rate. We achieved 29.4% WER on IHM standard dev set. When close-talk ASR model is decoded using SDM dev set, the WER increases from 29.4% to 70.03%, a significant drop. The results are shown in Table 1.

5.2. Homoscedastic Baseline

For homoscedastic DA baseline we trained a 6-layer DNN using SDM features as input with ±2 splicing and parallel IHM features as output without splicing. We use stochastic gradient descent (SGD) with learning rate of 0.001 for first 30 epochs and then with 0.0001 for another 20 epochs. At test time, SDM dev set is first passed through the DA, thus enhancing distant speech, then decoded using LSTM-HMM system. The WER on the enhanced SDM is significantly reduced from 70.03% to 64.67%, an absolute improvement of 5.36%. The results are reported in first row of Table 2.

5.3. ParallelNet Training

ParallelNet is trained in two proposed flavours, first one is the generalized model with mean and variance (Fig. 1) and the second one is only with variance (Fig. 2). Both the networks are trained in similar fashion as homoscedastic DA baseline, however, with some modifications. It was observed that variance prediction was going out of bound which can be accounted by the exponential component of softplus function, so the output is clipped before applying softplus. Also, the learning rate of the block in ParallelNet, which predicts close-talk features, was kept 5 times smaller than default learning rate. We fix LSTM-HMM model trained on IHM as our recognition system and SDM dev set as our standard test scenerio to report final results.

The WER with ParallelNet architecture as shown in Fig. 1 and trained using the heteroscedastic loss defined in Eq. 15 is shown in the second row of Table 2. It gives an absolute improvement of 4.73% in WER on the top of DA trained using homoscedastic loss. In order to examine the variance of residual noise learned by homoscedastic autoencoder and ParallelNet, we computed these terms on a subset of training set. It was found that the maximum likelihood estimate of β in the former case (Eq. 9) was 0.325, whereas in the later case the mean of variance (β_n) was 0.348, which are similar. The standard deviation of β_n turned out to be 0.81, which justifies the efficacy of using a time-varying co-variance term for residual noise. We also compared the MSE loss yielded by homoscedastic system (Eq. 5 which is 0.32) and heteroscedastic system (Eq. 12 which is 0.35). The higher MSE in the later case gives an indication that it is possibly learning a more generalized regression function which avoids over-fitting by not learning subtle variations in the training set and possibly outliers.

Table 3: *ParallelNet enhancement results on shallow networks and low-resource scenerio*

Front-End	**Data**(hours)	**No. of Layers**	**WER**(%)
Homoscedastic	20	3	66.21
Heteroscedastic	20	3	**63.22**
Homoscedastic	20	6	65.96
Heteroscedastic	20	6	**62.36**
Homoscedastic	80	3	64.92
Heteroscedastic	80	3	**60.83**
Homoscedastic	80	6	64.67
Heteroscedastic	80	6	**60.73**

We also experimented with excluding the μ_n from the prediction model in test time. From the third row of Table 2, we note that the WER does not drop significantly when μ_n is ignored (during testing). This can be attributed to the fact that the regularization term included in the final loss (Eq. 15) forces the mean to take very small values, thus not affecting the predicted close-talk features too much.

Getting large amount of parallel data can be difficult and time-consuming for other enhancement tasks. Moreover, deeper front-end networks can increase latency in practical scenerio. In the next set of experiments we target low-resource scenerio when availability of parallel data is less. The results are shown in Table 3. We experimented with ParallelNet with only variance excluding mean in the loss (as μ_n introduces extra parameters and is a part of prediction model). From the results, it can be seen that an absolute improvement of around 3% in WER is obtained in all possible cases.

To summarize, in distant speech enhancement scenerio, we achieved a significant reduction in WER from 64.67% to 59.94%, a relative improvement of **7.31**% using proposed heteroscedastic loss with our novel architecture. Overall, WER is reduced from 70.03% to 59.94%, which amounts to a relative improvement of **14.4**% in mismatch scenerio.

6. Conclusions

In this paper we propose a more generalized loss to train heteroscedastic denoising autoencoder to map speech features from far-field to close-talk domains. The proposed loss allevaiates the homoscedastic and zero-mean assumptions made in conventional residual noise model employed in standard denoising autoencoder. In addition, we also propose several novel neural network architectures that are akin to the proposed loss, where besides enhanced features, mean and co-variance are also predicted. Overall, we achieve a relative improvement of 14.4% in mismatched scenerio. We also show that the proposed loss and architecture give significant improvement in low-resource and low-computation scenerio.

7. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[4] A. Maas, Q. V. Le, T. M. O'neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," 2012.

[5] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation," in *Proc. Interspeech 2017*, 2017, pp. 384–388. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-733

[6] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 2, pp. 276–289, 2016.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, pp. 151–175, 2010. [Online]. Available: http://www.springerlink.com/content/q6qk230685577n52/

[8] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[9] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[10] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[11] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[12] R. Hsiao, J. Ma, W. Hartmann, M. Karafiát, F. Grézl, L. Burget, I. Szöke, J. H. Černockỳ, S. Watanabe, Z. Chen *et al.*, "Robust speech recognition in unknown reverberant and noisy conditions," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 533–538.

[13] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5590–5594.

[14] M. Krawczyk and T. Gerkmann, "Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[15] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[16] O. Novotny, O. Plchot, P. Matejka, and O. Glembek, "On the use of dnn autoencoder for robust speaker recognition," *arXiv preprint arXiv:1811.02938*, 2018.

[17] I. Himawan, P. Motlicek, D. Imseng, B. Potard, N. Kim, and J. Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4540–4544.

[18] N. Ng, R. A. Gabriel, J. McAuley, C. Elkan, and Z. C. Lipton, "Predicting surgery duration with neural heteroscedastic regression," *arXiv preprint arXiv:1702.05386*, 2017.

[19] Q. Hu, S. Zhang, M. Yu, and Z. Xie, "Short-term wind speed or power forecasting with heteroscedastic support vector regression," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 1, pp. 241–249, 2016.

[20] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[21] Y. Qian, T. Tan, and D. Yu, "An investigation into using parallel data for far-field speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5725–5729.

[22] Y. Qian, T. Tan, D. Yu, and Y. Zhang, "Integrated adaptation with multi-factor joint-learning for far-field speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5770–5774.

[23] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[24] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005, p. 100.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.