DEEP MIXTURE DENSITY NETWORK FOR STATISTICAL MODEL-BASED FEATURE ENHANCEMENT

Keisuke Kinoshita, Marc Delcroix, Atsunori Ogawa, Takuya Higuchi, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
{kinoshita.k, marc.delcroix, ogawa.atsunori, higuchi.takuya, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

We propose a novel framework designed to extend conventional deep neural network (DNN)-based feature enhancement approaches. In general, the conventional DNN-based feature enhancement framework aims to map input noisy observation to clean speech or a binary/soft mask in a deterministic way, assuming that there is one-to-one mapping between the input and the output without any uncertainty. However, when we consider that the general feature enhancement problem to be an ill-posed inverse problem where the mapping cannot be uniquely determined given an input signal, the assumption in the conventional approaches is not theoretically correct and potentially limits the performance of DNN-based feature enhancement. To overcome this problem, this paper proposes utilizing a mixture density network (MDN), which is a neural network that maps an input feature to a set of Gaussian mixture model (GMM) parameters representing the distribution of a target variable. By estimating the distribution of clean speech feature based on MDN, we are now able to explicitly consider the uncertainty in the parameter estimation. Then, we further utilizes the estimated GMM to obtain a refined clean speech estimate in the framework of statistical model-based feature enhancement. In this paper, after detailing the proposed framework and the MDN, we show mathematically and experimentally how MDN appropriately models the uncertainty information. We also show that the proposed method can outperform a conventional DNN-based feature enhancement method.

Index Terms— Mixture density network, model-based feature enhancement, conditional density

1. INTRODUCTION

Speech signals captured with distant microphones inevitably contain acoustic interferences such as background noise and reverberation, which are known to severely degrade the audible speech quality of captured signals [1] and the performance of automatic speech recognition (ASR) [2, 3]. To cope with such acoustic interferences, it is essential to establish effective speech/feature enhancement technologies. Thus a considerable amount of speech/feature enhancement research has already been undertaken from various perspectives [4].

Traditionally, before the deep learning era, many effective enhancement technologies were based on statistical model-based signal processing, such as the vector Taylor series (VTS) approach [5, 6]. These approaches were formulated in mathematically rigorous ways based on a model of a clean speech signal often represented as a Gaussian mixture model (GMM) [5, 6], a hidden Markov model (HMM) [7], or non-negative matrix factorization (NMF)-based model [8–10]. Thanks to the given models and the formulation, these statistical model-based approaches can appropriately handle uncertainty in parameter estimation and generate a final enhanced signal in an optimal manner, which clearly is a strong advantage of these methods. However, these methods became obsolete in the deep learning era, since they performed somewhat poorly compared with deep neural network (DNN)-based approaches.

Recently, deep learning was successfully applied to speech/feature enhancement problems [11–13], and has outperformed the conventional model-based approaches. These approaches perform enhancement by using learned DNN-based mapping between corrupted speech signals and clean speech signals or soft/binary masks. Although the conventional DNN-based enhancement approaches are very powerful and can work very well for many tasks, we argue that there is a fundamental problem that potentially limits their performance. With these approaches, it is explicitly assumed that there is deterministic 1-to-1 mapping between a given input signal (i.e., an observed signal) and a target signal (i.e., an output clean speech signal) without any uncertainty. However, this assumption is clearly incorrect, since the enhancement problem, which is a typical inverse problem, should be theoretically viewed as an ill-posed problem, where the mapping to be learned cannot be uniquely determined, namely it is a 1-to-many mapping problem. In fact, theoretically, given an input observed signal, we have an infinite number of combinations of clean speech and noise, which can form the particular input signal. If we solve the ill-posed problem as if it were a 1-to-1 mapping problem, we may end up obtaining simply an average of several potential candidate values, which is not necessarily itself a correct value [14]. A traditional way of avoiding such an unwanted average value is to first consider the distribution of the target variable and then obtain an optimal estimate by appropriately handling the estimation uncertainty.

In this study, we extend the conventional DNN-based enhancement approaches and formulate the enhancement problem as a 1-to-many mapping problem based on a DNN. We explicitly consider uncertainty in the parameter estimation in the proposed framework unlike conventional methods [11–13]. The key to dealing with this issue is to incorporate i) a mechanism for modeling the 1-to-many mapping based on a DNN and ii) a mechanism for forming a most likely answer based on multiple candidates. To handle the 1-to-many mapping problem while taking the uncertainty in the parameter estimation into account, we utilize a neural network model called mixture density network (MDN) [14], which, given a noisy input signal $\mathbf{y}$, can output a conditional distribution of a clean speech signal $\mathbf{s}$, $p(\mathbf{s}|\mathbf{y})$, in the form of a GMM. Then, using the distribution of clean speech $p(\mathbf{s}|\mathbf{y})$ estimated by the MDN as a basis, we obtain a final optimal estimate in a minimum mean square error (MMSE) sense based on the VTS approach, taking advantage of the statistical model-based feature enhancement.

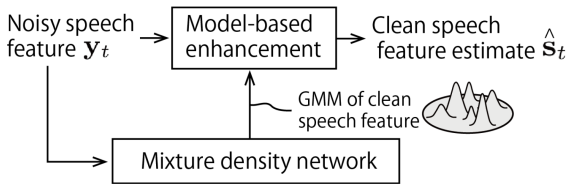

Fig. 1. Overview of proposed framework

2. PROPOSED FRAMEWORK

In this section, we first provide a brief overview of the proposed framework in Section 2.1, and then introduce its essential components in Sections 2.2. and 2.3.

2.1. Overview of proposed framework

Figure 1 shows the concept of the proposed framework. At each time frame t, given an input noisy speech feature $\mathbf{y}_t$, we first obtain the distribution of a clean speech feature $\mathbf{s}_t$ in the form of GMM by using an MDN. Then, based on the estimated distribution of the clean speech feature and the input noisy speech feature $\mathbf{y}_t$, we obtain an optimal estimate of $\mathbf{s}_t$ in the MMSE sense by using a model-based approach that explicitly considers the physical interaction between the clean speech feature, observed speech feature and the noise feature. In this study, for simplicity, we use the standard VTS approach for the model-based enhancement. In the proposed framework, we can take advantage of both the DNN and model-based feature enhancement, since the clean speech distribution is estimated in the well known DNN framework, and the final output is formed with the model-based feature enhancement approach whose performance and behavior are easy for us to analyze and improve.

In the following section, we describe the essential components of the proposed framework, namely an MDN and model-based enhancement, and show how it handles the overall enhancement process.

2.2. Mixture density network (MDN)

In this subsection, we briefly review the concept and formulation of the MDN, and show how it predicts the conditional distribution of clean speech given noisy input speech. An MDN is a neural network that maps the input observed noisy log Mel filterbank (hereafter, FBANK) feature $\mathbf{y}_t$ at time frame t to a set of GMM parameters representing the distribution of a clean speech FBANK feature $\mathbf{s}_t$. The GMM parameters to be estimated include mean vectors $\boldsymbol{\mu}_i(\mathbf{y}_t)$, variances $\sigma_i(\mathbf{y}_t)$, and weights $\alpha_i(\mathbf{y}_t)$ as shown in Fig. 2. i is a mixture component index.

In the actual process, first, the MDN converts the input vector $\mathbf{y}_t$ using a multi-layer perceptron (MLP) with an output layer of linear units, and obtain outputs $\mathbf{z}_t$ as:

$$\mathbf{z}_t = f_\theta(\mathbf{y}_t) \tag{1}$$

where $f_\theta(\cdot)$ corresponds to a set of transformations in the MLP. The total number of network outputs, i.e., the dimension of $\mathbf{z}_t$, is $(c+2)\times M$ where c corresponds to the dimension of the clean speech feature to be obtained. M corresponds to the number of mixture components in the GMM estimated by the MDN. Then, $\mathbf{z}_t$ is partitioned into three subsets $\mathbf{z}_{t,i}^{(\boldsymbol{\mu})}$, $\mathbf{z}_{t,i}^{(\sigma)}$, and $\mathbf{z}_{t,i}^{(\alpha)}$, which correspond to the outputs used to calculate the GMM mean vectors, variances and weights, respectively.

$$\mathbf{z}_t = [\mathbf{z}_{t,1}^{(\boldsymbol{\mu})}, \ldots, \mathbf{z}_{t,M}^{(\boldsymbol{\mu})}, \mathbf{z}_{t,1}^{(\sigma)}, \ldots, \mathbf{z}_{t,M}^{(\sigma)}, \mathbf{z}_{t,1}^{(\alpha)}, \ldots, \mathbf{z}_{t,M}^{(\alpha)}], \tag{2}$$

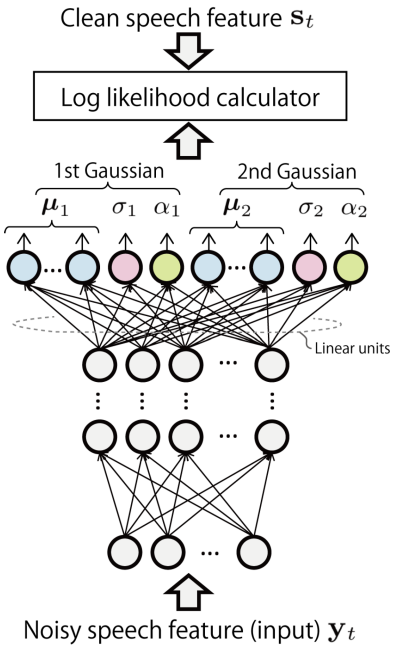

Fig. 2. Training of mixture density network (with 2 mixture components (i.e. $M = 2$)). All the output parameters, namely $\sigma_1(\mathbf{y}_t)$, $\sigma_2(\mathbf{y}_t)$,$\alpha_1(\mathbf{y}_t)$, $\alpha_2(\mathbf{y}_t)$, $\boldsymbol{\mu}_1(\mathbf{y}_t)$, and $\boldsymbol{\mu}_2(\mathbf{y}_t)$ are functions of $\mathbf{y}_t$, but $(\mathbf{y}_t)$ was omitted from the figure for the sake of simplicity.

After the partitioning, each subset is passed through a set of specific transformations for conversion to the GMM mean vectors, variances and weights as:

$$\boldsymbol{\mu}_i(\mathbf{y}_t) = \mathbf{z}_{t,i}^{(\boldsymbol{\mu})}, \tag{3}$$

$$\sigma_i(\mathbf{y}_t) = \exp(\mathbf{z}_{t,i}^{(\sigma)}), \tag{4}$$

$$\alpha_i(\mathbf{y}_t) = \frac{\exp(\mathbf{z}_{t,i}^{(\alpha)})}{\sum_{j=1}^{M}\exp(\mathbf{z}_{t,j}^{(\alpha)})}, \tag{5}$$

In the training stage of the MDN, these GMM parameters are passed to a log likelihood calculator, which calculates the likelihood of the clean speech FBANK feature $\mathbf{s}_t$ as:

$$p(\mathbf{s}_t|\mathbf{y}_t) = \sum_{i=1}^{M}\alpha_i(\mathbf{y}_t)\phi_i(\mathbf{s}_t|\mathbf{y}_t), \tag{6}$$

$$\phi_i(\mathbf{s}_t|\mathbf{y}_t) = \frac{1}{(2\pi)^{c/2}\sigma_i(\mathbf{y}_t)}\exp\left\{-\frac{||\mathbf{s}_t - \boldsymbol{\mu}_i(\mathbf{y}_t)||^2}{2\sigma_i(\mathbf{y}_t)^2}\right\} \tag{7}$$

The calculated likelihood is directly used to define an error function for the neural network by taking the negative logarithm of the likelihood as:

$$E = -\ln\left\{\sum_{i=1}^{M}\alpha_i(\mathbf{y}_t)\phi_i(\mathbf{s}_t|\mathbf{y}_t)\right\}, \tag{8}$$

The parameters of the MLP in the MDN will be optimized such that the negative log likelihood in eq. (8) will be minimized. In the test stage, using the optimized MDN, we estimate the distribution of clean speech $\mathbf{s}_t$, i.e., the GMM parameters at each time frame t, given $\mathbf{y}_t$.

Note that, if we limit the structure of the MDN so that it outputs only a single Gaussian component, and takes the mean vector $\boldsymbol{\mu}_1(\mathbf{y}_t)$ as an estimate of clean speech feature, it becomes theoretically equivalent to the conventional DNN-based feature enhancement optimized with MMSE criterion [14]. However, by outputting not only the mean vector $\boldsymbol{\mu}_1(\mathbf{y}_t)$ but also $\sigma_1(\mathbf{y}_t)$ representing estimation uncertainty, we may be able to further refine the clean speech estimate based on the distribution, and obtain a better estimate than the conventional method. Furthermore, by incorporating the idea of the *mixture* model, we can appropriately handle a complex clean speech distribution and get to know how confident DNN is about the current clean speech estimate.

2.3. Model-based feature enhancement based on MDN

In this study, we used the standard VTS approach [5, 6] to obtain an optimal clean speech estimate based on the clean speech GMM generated by the MDN.

In the VTS approach, the clean speech FBANK/MFCC feature is modeled as a GMM. Conventionally, its parameters are trained in advance with the maximum likelihood criterion by using training data. The FBANK/MFCC feature of the background noise is represented as a single Gaussian and its parameters are estimated blindly for each test utterance. In the test phase, the clean speech GMM and the noise Gaussian are combined by using a VTS approximation to form the probability density function of the observed signal. After iteratively updating the parameters of the noise Gaussian and the posterior probability of the clean speech GMM in the maximum likelihood sense, we can obtain an optimal estimate of the clean speech feature in the MMSE sense.

With the proposed method, we simply replace the pretrained GMM with the one estimated with the MDN. For example, with the 1st order VTS handling FBANK features, we can compose the mean vector of the i-th GMM component of the observed siganal at time frame t, $\boldsymbol{\mu}_{i,t}^{(\mathrm{o})}$, as:

$$\begin{aligned}\boldsymbol{\mu}_{i,t}^{(\mathrm{o})} &= h(\boldsymbol{\mu}_i(\mathbf{y}_t), \boldsymbol{\mu}^{(\mathrm{n},0)}) + \mathbf{H}_i(\boldsymbol{\mu}^{(\mathrm{n},l)} - \boldsymbol{\mu}^{(\mathrm{n},0)}),\\ h(\boldsymbol{\mu}_i(\mathbf{y}_t), \boldsymbol{\mu}^{(\mathrm{n},0)}) &= \boldsymbol{\mu}_i(\mathbf{y}_t) + \log(\mathbf{1} + \exp(\boldsymbol{\mu}^{(\mathrm{n},0)} - \boldsymbol{\mu}_i(\mathbf{y}_t)),\end{aligned}$$

where $\boldsymbol{\mu}^{(\mathrm{n},0)}$ and $\boldsymbol{\mu}^{(\mathrm{n},l)}$ correspond to the initial estimate of the mean vector of the Gaussian representing noise, and its current estimate. $\mathbf{H}_i(\boldsymbol{\mu}^{(\mathrm{n},l)} - \boldsymbol{\mu}^{(\mathrm{n},0)})$ is a Gaussian dependent Jacobian matrix [5, 6]. In contrast to the conventional VTS, here we must compose a different GMM for the observed signal at each time frame t, since the clean speech GMM is time-varying as the term $\boldsymbol{\mu}_i(\mathbf{y}_t)$ is explicitly dependent on the time frame index t. the variance of the GMM can be similarly updated.

3. EXPERIMENTS

In this section, we first analyze the behavior of the MDN to see how appropriately it predicts clean speech distribution. Then, we evaluate the proposed framework in comparison with a conventional DNN-based feature enhancement method [11] where the aim is to map the input signal to the clean speech. This conventional method is hereafter referred to as conv-DNN.

3.1. Experimental conditions

3.1.1. Database and test acoustic environments

We used the Aurora-4 database to analyze the MDN behavior and evaluate the proposed framework. Aurora-4 is based on the Wall Street Journal 5k task, whose training data set comprises about 14 hours of speech including 83 speakers. To generate the training data, 6 different types of noise (street traffic, train stations, cars, babble, restaurants, airports) were artificially added to clean speech at randomly selected signal-to-noise ratios (SNR) of between 10 and 20 dB. To train the conv-DNN and MDN, we used noisy training data and the corresponding clean speech data. Note that for monitoring the convergence of the conv-DNN and MDN learning process, we randomly extracted 5% of the training data as a validation set, and used the remaining 95% of the data for the actual training.

To test the algorithms, we used half of the development set of Aurora-4 (*.wv1 set), which contains the same types of noise as the training data but with SNRs of between 5 and 15 dB. The sampling frequency of the data was 16 kHz.

3.1.2. Network structure and other related configurations

For both the conv-DNN and the MDN that we used in our proposed approach, we employed a standard feed-forward DNN structure with 5 hidden layers of hyperbolic tangent activation functions. Although any type of network structure could be employed, we chose this network configuration for the sake of simplicity. We used a 40-order FBANK feature and its 1st and 2nd derivatives as the input feature of the conv-DNN and the MDN. Following the common practice in speech processing, the feature of the current frame was spliced with features within 5 left and 5 right context frames to form an input feature vector consisting of 11 frames. The number of nodes in the hidden layers was set at 2048. The output node of the conv-DNN was set at 40, which corresponds to the vector dimension of the static clean speech FBANK feature at the current frame. Similarly, the dimension of the mean vector $\boldsymbol{\mu}_i(\mathbf{y}_t)$ in the output of MDN was set to 40. The number of mixture components M for the MDN was set at 1 for behavior analysis in Section 3.2, and at 1 and 2 in Section 3.3. We first discriminatively pretrained the network [15], where we trained the whole network every time we added one more hidden layer. And then, it was fine-tuned with back-propagation. The conv-DNN was trained with the MMSE criterion to directly estimate the clean speech FBANK feature. The MDN was optimized by using an Adam optimizer with an initial learning rate of 0.0005. We obtained the initial noise statistics for the VTS by subtracting the initial MMSE clean speech estimates from the observed signal in the power spectral domain.

3.2. Behavior of MDN

We first analyzed the behavior of the MDN by setting the number of mixture components at 1, and confirming whether it can appropriately generate uncertainty information in the parameter estimation, i.e. $\sigma_1(\mathbf{y}_t)$. The upper panel in Fig. 3 shows the contour of the clean speech FBANK feature, the observed speech FBANK feature and the mean of a Gaussian $\boldsymbol{\mu}_1(\mathbf{y}_t)$ estimated by the MDN, at the 5th Mel filterbank. We can see that the contour of $\boldsymbol{\mu}_1(\mathbf{y}_t)$ closely follows the clean speech FBANK feature contour in most cases but occasionally deviates from them. We confirmed that this behavior of $\boldsymbol{\mu}_1(\mathbf{y}_t)$ closely matches the FBANK feature contour estimated by the conv-DNN.

The lower panel in Fig. 3 shows the contour of $\boldsymbol{\mu}_1(\mathbf{y}_t)$ along with the variance information $\sigma_1(\mathbf{y}_t)$. The pink regions correspond to $\pm 2\sigma$. By comparing the upper and lower panels, we can see that the variance becomes appropriately large when the mean vector of the Gaussian $\boldsymbol{\mu}_1(\mathbf{y}_t)$ deviates from the true clean speech FBANK feature contour. The results of this experiment enabled us to confirm that the MDN can appropriately output uncertainty in the parameter estimation, which can be further utilized to determine an optimal

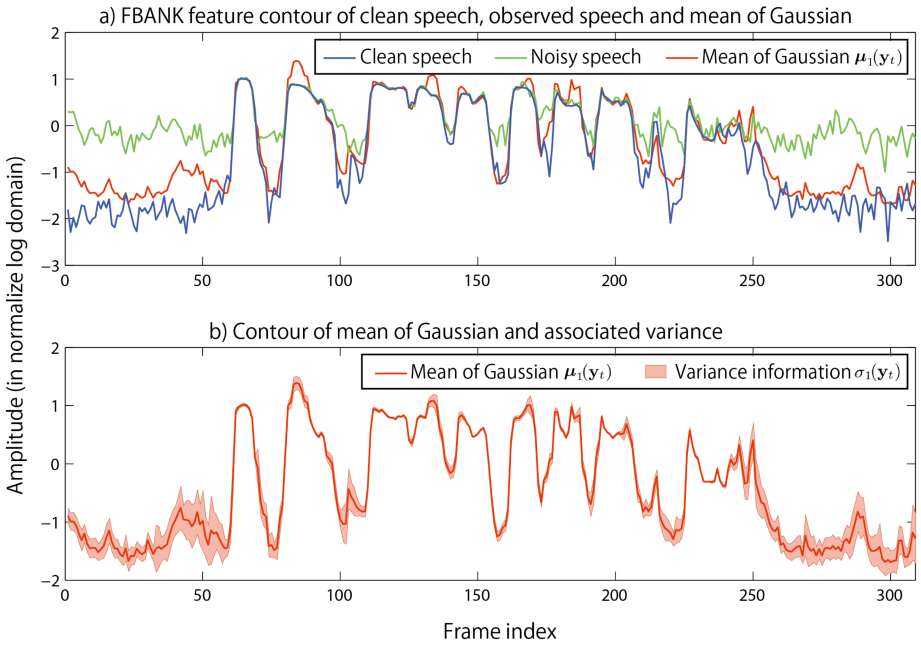

Fig. 3. Behavior of MDN: (upper panel) Contour of clean speech, observed speech and mean of a Gaussian estimated by MDN (5th Mel filterbank), (Lower panel) Contour of mean of Gaussian estimated by MDN and associated (time-varying) $\pm 2\sigma$ taken from $\sigma_1(\mathbf{y}_t)$.

Table 1. Log spectral distance between a clean FBANK feature and an unprocessed (i.e., observed) FBANK feature, an FBANK feature enhanced by the conventional DNN-based feature enhancement method and an FBANK feature enhanced by the proposed method (Average over all 1981 utterances).

Unproc.	Conv.	Prop. ($M = 1$)	Prop. ($M = 2$)
0.96	0.53	0.44	0.42

clean speech estimate.

3.3. Results of model-based enhancement based on MDN

We also evaluated the degree to which we could improve the clean speech estimate by performing VTS based on the GMM estimated by the MDN. Table 1 shows the log spectral distance for each evaluation target excluding non-speech segments. As the table shows, the proposed framework (MDN-VTS) successfully improved the accuracy of the clean speech estimate and outperformed the conv-DNN for both $M = 1$ and $M = 2$, where M is the number of mixture components. We also tested $M > 2$ cases, but found that the performance remained largely unchanged. This issue will be revisited in more detail in future work.

4. RELATION TO PRIOR WORKS

Conventional DNN-based feature enhancement can be roughly categorized into two trends. One trend is to solve the enhancement problem as a regression problem between an observed feature and a clean feature or a soft mask to obtain clean speech [11–13, 16]. Typically these networks are trained with the MMSE criterion. A characteristic of MMSE training is that uncertainty in the DNN-based regression is not explicitly taken into account [14], as described in the introduction of this paper. The other trend is to treat the enhancement problem as a classification problem [17]. In such studies, the DNN is typically trained with the cross entropy criterion, on the assumption that each time frequency bin must belong solely to either speech or noise, i.e. a binary mask. In this case, the classification uncertainty is considered in the DNN output as a form of soft mask. However, it is not clear how the value of the output soft mask is related to the physical properties of the input observed signal such as the SNR, and thus the network output is not easily interpreted. Consequently, it becomes difficult to obtain an optimal clean speech estimate based on the network output. In contrast to these conventional methods, the proposed approach explicitly takes the estimation uncertainty into account, while the network output is easily interpreted and can be further utilized to refine the clean speech estimate. An experimental comparison with other approaches [16, 17] will be done in future work.

5. CONCLUSIONS

In this paper, we extended conventional DNN-based feature enhancement to appropriately handle estimation uncertainty and obtain an optimal clean speech estimate. To deal with the estimation uncertainty, we employed an MDN, which can estimate the distribution of clean speech based on input noisy speech. To obtain an optimal MMSE estimate of the clean speech based on the estimated distribution of the clean speech, we performed model-based feature enhancement, i.e. VTS. In the experiments, we first confirmed the behavior of the MDN and found that the variance of the estimated clean speech GMM became appropriately large when the estimated mean value deviated from the true target value. We also confirmed that by employing model-based feature enhancement based on the GMM estimated with the MDN, we could obtain a better clean speech estimate than with the conventional DNN-based feature enhancement approach.

6. REFERENCES

[1] I. Tashev, *Sound Capture and Processing*, Wiley, New Jersey, 2009.

[2] X. Huang, A. Acero, and H.-W. Hong, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, New Jersey, 2001.

[3] M. Wölfel and J. McDonough, *Distant Speech Recognition*, Wiley, New Jersey, 2009.

[4] X. Huang, A. Acero, and H-W. Hon, *Spoken language processing*, Prentice Hall, Upper Saddle River, NJ, 2001.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 733–736.

[6] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, vol. 24(1), pp. 39–49, 1998.

[7] T. Kristjansson, J. Hershey, P. Olsen, and R. Gopinath, "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," in *Proc. Int'l Conf. Speech and Language Process. (ICSLP)*, 2006, pp. 97–100.

[8] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech de-noising using nonnegative matrix factorization with priors," in *ICASSP*, 2008, pp. 4029 – 4032.

[9] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15(1), pp. 1 – 12, 2007.

[10] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21(10), pp. 2140–2151, 2013.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Letters*, vol. 21(1), pp. 65 – 68, 2014.

[12] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *ICASSP*, 2014, pp. 4656–4659.

[13] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Interspech*, 2012.

[14] C. Bishop, *Mixture density networks*, Ph.D. thesis, Aston University, Tech. Rep. NCRG/94/004, 1994.

[15] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011, pp. 24–29.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*, 2015, pp. 708–712.

[17] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016, pp. 196–200.