



# Integration of DNN based Speech Enhancement and ASR

Ramón F. Astudillo, Joana Correia, Isabel Trancoso

INESC-ID / Instituto Superior Técnico, Universidade de Lisboa, Portugal

{ramon.astudillo, joanac, isabel.trancoso}@inesc.id.pt

## Abstract

Speech enhancement employing Deep Neural Networks (DNNs) is gaining strength as a data-driven alternative to classical Minimum Mean Square Error (MMSE) enhancement approaches. In the past, Observation Uncertainty approaches to integrate MMSE speech enhancement with Automatic Speech Recognition (ASR) have yielded good results as a lightweight alternative for robust ASR. In this paper we thus explore the integration of DNN-based speech enhancement with ASR by employing Observation Uncertainty techniques. For this purpose, we explore various techniques and approximations that allow propagating the uncertainty of inference of the DNN into feature domain. This uncertainty can then be used to dynamically compensate the ASR model utilizing techniques like uncertainty decoding. We test the proposed techniques on the AU-RORA4 corpus and show that notable improvements can be attained over the already effective DNN enhancement.

**Index Terms:** ASR, Observation Uncertainty, Uncertainty Propagation

## 1. Introduction

As many other fields, speech enhancement has been recently struck by the newest wave of neural network based research. Deep Learning techniques have shown outstanding ability to learn the mapping from noisy speech to clean speech providing a straightforward data-driven method for speech enhancement when sufficient data is available.

Incorporating speech enhancement as a pre-processing stage to an Automatic Speech Recognition (ASR) system is a simple way to attain robustness. For clean speech trained ASR systems, such pre-processing compensates the environmental distortions present in real-world speech applications. For ASR systems trained with real-world data, enhancement reduces the variability of the acoustic space to be learnt. Another advantage of pre-processing is the possibility of utilizing the expertise of the active speech enhancement community. This is particularly important in Short-Time Fourier Transform (STFT) domain speech enhancement, where source independence, spatial filtering or signal sparsity can be exploited.

Conventional speech enhancement as e.g. Wiener [1] or Ephraim-Malah [2] filters is based on Minimum Mean Square Error (MMSE) estimation. MMSE methods are well understood and allow easy incorporation of prior knowledge or additional cues as e.g. spatial information in multi-channel settings [3]. On the negative side, they depend on proper estimation of the a priori parameters i.e. sufficiently good voice activity detection. An additional advantage of MMSE methods is the possibility

of integrating them with the ASR system via Observation Uncertainty techniques. Techniques like Uncertainty Decoding [4] or Modified Imputation [5] can be used to dynamically compensate the acoustic models for the residual uncertainty after enhancement. Similar techniques are also applicable to deep learning based acoustic models [6].

Speech enhancement using deep learning is a purely data-based approach. It exploits the ability of neural networks to learn complex mappings from observed speech features e.g. noisy speech to clean speech features or ideal filter coefficients. Many approaches utilize larger versions of the old Multi-Layer Perceptron (MLP) model [7], often incorporating newly discovered initialization and regularization techniques. The so called Deep Neural Networks (DNN) [8] have been used to estimate Ideal Binary Mask (IBMs) or Ideal Ratio Masks (IRMs) [9]. Denoising Auto-Encoders (DAE) are another related option. These estimate directly a mapping from noisy to clean speech features. In [10], the log-amplitude of the clean STFT of speech is estimated. Other DAE approaches account for the time structure of speech through Recurrent Neural Networks (RNNs), [11] or Long-Short Time Memory (LSTM) networks [12]. These two last works estimate speech in the MFCC domain.

The objective of this work, is to attain integration of DNN-based enhancement with ASR systems by means of Observation Uncertainty techniques. This would provide this methods with the same advantage as classical MMSE methods including dynamic compensation of both Statistical and DNN-based acoustic models. This work is therefore related to previous works integrating MMSE estimators and ASR systems [13] as well as recent investigations on the residual uncertainty in DNN/MLP inference. We use the Gaussian Marginalization model developed in [14], to obtain a measure of estimate uncertainty in the context of speech enhancement. Furthermore, since this estimate concerns a matrix of binary random variables, we apply the Sparsity-based Uncertainty model introduced in [15] to attain uncertainty propagation into MFCC domain for IRM-based DNN estimation.

The reminder of the paper is as follows. Section 2 briefly reviews speech enhancement with DNNs. Section 3 deals with the modeling of uncertainty when using DNNs for speech enhancement with IRMs and how to propagate this into MFCC domain. Section 4 presents the experimental setup used to validate the presented approaches and finally Section 5 introduces the conclusions.

## 2. Speech Enhancement with DNNs

This work focuses on DNN speech enhancement based on mask estimation in STFT domain. In practice, for the purpose of ASR, mask estimation can be done in other domains as e.g. in Mel-domain [9]. The STFT domain is chosen for various rea-

10.21437/Interspeech.2015-709

This work has been partially supported by the FCT through the grant numbers SFRH/BPD/68428/2010, SFRH/BD/103402/2014 and projects UID/CEC/50021/2013 and CMUP-ERI/HCI/0051/2013.

sons. In the STFT-domain, properties like signal sparsity are known to hold well. STFT it is also the domain utilized for speech enhancement for human use e.g. hearing aids. Working in the STFT domain allows thus to benefit from progress in DNN speech enhancement techniques aimed for this purpose. The choice of STFT domain is also consistent with parallel works integrating MMSE speech enhancement with ASR [13].

Let  $\mathbf{X} \in \mathbb{C}^{K \times L}$  and  $\mathbf{Y} \in \mathbb{C}^{K \times L}$  denote the clean and noisy STFTs of a signal respectively.  $L$  is the number of analysis frames and  $K$  the number of frequency bins under half the Nyquist frequency. Under the assumption of signal sparsity, each Time-Frequency (T-F) element of  $\mathbf{Y}$  is going to hold either speech or noise. We can thus find an IBM that separates both signals [16]. In practice, rather than a binary decision, estimating mask of ratios between 0 and 1 yields often better results [9]. An IRM can also be interpreted as a probabilistic or uncertain estimate of an IBM. In other words, an estimate of the probability that a given T-F bin is occupied by speech. Under this view, an estimate of the amplitude of each clean T-F bin can be attained through

$$|\widehat{X}|_{kl} = E\{|X_{kl}|\} = p(\Lambda_{kl}|\mathbf{Y}) \cdot |Y_{kl}| = G_{kl} \cdot |Y_{kl}| \quad (1)$$

where  $(k, l)$  index frequency bins and frames respectively and  $\Lambda_{kl}$  is a Bernoulli variable indicating speech presence.

When used for enhancement, a DNN must be trained to map features of  $\mathbf{Y}$  to  $\mathbf{G}$ . The DNN model can be described as

$$\hat{G}_{kl} = f_k^N \circ \mathbf{f}^{N-1} \circ \dots \circ \mathbf{f}^1(\log(|\mathbf{Y}|)) \quad (2)$$

where absolute and logarithm non-linearities act here element-wise. In practice, other features of  $\mathbf{Y}$  are often appended and only a context of frames is used to estimate each  $\hat{G}_{kl}$ . This is however omitted for notation simplicity. Each function in (2) corresponds to the operations involved in a layer and can be expressed as

$$\tilde{z}_{jl}^n = f_j(\tilde{\mathbf{z}}_l^{n-1}) = \frac{1}{1 + \exp(-\sum_{i=1}^I W_{ij}^{n-1} \tilde{z}_{il}^{n-1})} \quad (3)$$

where  $\mathbf{W}^{n-1}$  are the weights of layer  $n-1$ . The context of frames at the input is usually stacked into a single vector.

Given a set of  $L^{\text{tr}}$  example frames of noisy speech  $\mathbf{Y}_l$  and corresponding target gains  $\mathbf{G}_l$ , training is performed employing Stochastic Gradient Descent (SGD) [7] or any of its variants. Additional pre-training or regularization is also sometimes applied. The typical cost functions used for learning are either cross-entropy (CE)

$$\mathcal{F}^{\text{CE}} = \frac{1}{L^{\text{tr}}} \sum_{l=1}^{L^{\text{tr}}} \sum_{k=1}^K \hat{G}_{kl} \cdot \log(G_{kl}) + (1 - \hat{G}_{kl}) \cdot \log(1 - G_{kl}) \quad (4)$$

or Mean Square Error (MSE)

$$\mathcal{F}^{\text{MSE}} = \frac{1}{L^{\text{tr}}} \sum_{l=1}^{L^{\text{tr}}} \sum_{k=1}^K (\hat{G}_{kl} - G_{kl})^2. \quad (5)$$

Strictly speaking, the probabilistic interpretation of the estimated gain can only be employed when using the CE training criterion. In this case we have that

$$\hat{G}_{kl}^{\text{CE}} \equiv p(\Lambda_{kl}|\mathbf{Y}). \quad (6)$$

In practice, MSE training also produces estimates in the same domain. Values close to 0.5 can also be interpreted as uncertain although not in the same probabilistic sense.

### 3. Accounting for DNN Enhancement Uncertainty

#### 3.1. The Bernoulli Observation Uncertainty Model

For conventional MMSE enhancement methods, it is easy to see that an uncertain description of the signal after enhancement is already available. Any estimate performed with a Wiener filter has an associated variance equal to the residual Bayesian MSE [13]. This uncertainty can be ignored during training and used at testing time for dynamic compensation with e.g. Uncertainty Decoding [4].

Similarly, the speech enhancement estimate coming from the DNN, when interpreted probabilistically, has an associated variance

$$\lambda_{kl} = \text{Var}\{|X_{kl}|\} = p(\Lambda_{kl}|\mathbf{Y}) \cdot (1 - p(\Lambda_{kl}|\mathbf{Y})) \cdot |Y_{kl}|^2. \quad (7)$$

This corresponds to the scaled variance of the Bernoulli variable implied in the DNN. A propagation approximation for the Bernoulli uncertainty model was proposed at the second CHiME challenge [15]. This was used to propagate the uncertainty of a Wiener estimate interpreted as a binary mask under the sparsity assumption. Unlike in the approach presented here, in this work and in [17] the sparsity uncertainty is obtained from conventional MMSE methods. Here it is used to directly account for the natural uncertainty of DNNs as probabilistic model. This approximation, thus, allows to treat DNN speech enhancement in the same way as in MMSE methods in [13]. In other words, the uncertainty of enhancement is ignored during training, and applied during recognition for compensation.

#### 3.2. Accounting for the internal Uncertainty of Inference

The output layer of a DNN for IRM-based enhancement is still a sigmoid layer. It is thus not different from any intermediate layer of the DNN. In the same way that the output of the DNN is here considered as uncertain, any output of an internal layer can be considered as such as well.

In fact, it is well known that a DNN with sigmoid layers can be interpreted as a concatenation of multivariate Bernoulli models [8]. Exact Inference for such a model would be attained by marginalizing over all possible activations of each node of each layer. Inference can thus be expressed as

$$p(\Lambda_{kl}|\mathbf{Y}) = \sum_{\mathbf{h}^{N-1} \in \mathbf{H}^{N-1}} \dots \sum_{\mathbf{h}^1 \in \mathbf{H}^1} p(\Lambda_{kl}, \mathbf{h}^{N-1}, \dots, \mathbf{h}^1|\mathbf{Y}). \quad (8)$$

This is, however, computationally prohibitive, since for each layer of the network we need to marginalize over the  $2^{|\mathbf{H}^n|}$  possible states of the Bernoulli variables, where  $|\mathbf{H}^n|$  is the number of nodes in that layer.

Conventional inference for a DNN, the so called forward pass, can then be seen as a coarse approximation of inference for this model, where the uncertainty at the output of each layer is

neglected. The marginalization can be then computed by propagating only the mean through the network, yielding a probabilistic interpretation for (3)

$$E\{\mathbf{h}^n|\mathbf{Y}\} = \frac{1}{1 + \exp(-\sum_{i=1}^I W_{ij}^{n-1} E\{\mathbf{h}^{n-1}|\mathbf{Y}\})}. \quad (9)$$

Since we are choosing to consider the uncertainty at the output of the DNN, for enhancement purposes, it is also worth to consider the uncertainty in the intermediate layers. In [14] a closed form solution was presented to take into account the uncertainty in the intermediate layers.

The Gaussian Marginalization Multi-Layer Perceptron (GM-MLP) [14] assumes the large weighted sum of conditionally independent Bernoulli variables at the input of each layer to converge to a Gaussian distribution. It then utilizes the Piecewise Exponential (PIE) approximation [6] to attain a closed form for the Marginalization. The resulting forward pass formulas propagate not only the mean, but also the variance through the network. These correspond to

$$\begin{aligned} E\{h_j^n|\mathbf{Y}\} &\approx 2(\mu_j^n + \frac{1}{2} \log(2)\Sigma_j^n - 1) \\ &\cdot \Phi\left(-\frac{\mu_j^n}{\sqrt{\Sigma_j^n}} - \log(2)\sqrt{\Sigma_j^n}\right) \\ &- 2(-\mu_j^n + \frac{1}{2} \log(2)\Sigma_j^n - 1) \\ &\cdot \Phi\left(\frac{\mu_j^n}{\sqrt{\Sigma_j^n}} - \log(2)\sqrt{\Sigma_j^n}\right) \\ &+ \Phi\left(\frac{\mu_j^n}{\sqrt{\Sigma_j^n}}\right) \end{aligned} \quad (10)$$

where  $\mu_j^n, \Sigma_j^n$  are the first and second order statistics after each linear transformation of each layer, defined in [6, Eqs. 8,17]. The frame index  $l$  has also been removed for simplicity of notation.

Finally the variance, is again that of a scaled Bernoulli variable

$$\text{Var}\{h_j^n|\mathbf{Y}\} = (1 - E\{h_j^n|\mathbf{Y}\})E\{h_j^n|\mathbf{Y}\}. \quad (11)$$

To obtain the mean and variance at the output of the network, we just need to take into account the equivalence  $\Lambda_{kl} \equiv h_{kl}^N$

## 4. Propagation of DNN Uncertainty into MFCC domain

### 4.1. Uncertainty Propagation for the Bernoulli Model

In the case of MMSE speech enhancement methods, the uncertainty after enhancement is characterized by a circularly symmetric complex Gaussian distribution. In order to propagate this uncertainty into MFCC domain, the log-normal assumption can be used, leading to a closed form solution for a MMSE-MFCC estimator [13]. The log-normal assumption also guarantees that the uncertain features in MFCC domain are Gaussian distributed with a known variance. This allows the use of techniques like Uncertainty Decoding [4].

As seen in previous sections, in the case of DNN-based IRM enhancement, the uncertainty model is a scaled Bernoulli distribution. For this observation uncertainty model, an approximate closed-form solution for propagation was derived in [15]. This is based on assuming that the weighted sum of Bernoulli random variables at the Mel-filterbank converges to a

log-normal distribution. Upon this premise, the variance of the uncertain features in MFCC domain can be obtained as

$$\begin{aligned} \Sigma_{il}^{\text{MFCC}} &\approx \sum_{j=1}^J \sum_{j'=1}^J T_{ij} T_{ij'} \\ &\log\left(\frac{\sum_{k=1}^K W_{jk} W_{j'k} \lambda_{kl}}{\mu_{jl}^{\text{Mel}} \mu_{j'l}^{\text{Mel}}} + 1\right) \end{aligned} \quad (12)$$

where  $i$  is the DCT index,  $W_{jk}, T_{ij}$  are the Mel-filterbank and DCT coefficients respectively and

$$\mu_{jl}^{\text{Mel}} = \sum_{k=1}^K W_{jk} \cdot p(\Lambda_{kl}|\mathbf{Y}) \cdot |Y_{kl}|. \quad (13)$$

However, unlike for the circularly symmetric complex Gaussian model, the log-normal assumption is a coarse approximation. This leads to inaccuracies in the computation of the mean of the propagated variable. As in [15], it was empirically observed that neglecting the uncertainty at the Mel-filterbank stage, led to better results. The formula thus corresponds to

$$\mu_{il}^{\text{MFCC}} \approx \sum_{j=1}^J T_{ij} \cdot \log(\mu_{jl}^{\text{Mel}}). \quad (14)$$

### 4.2. Dynamic Compensation of Acoustic Models

Once the mean and variance have been propagated into MFCC domain, they can be used to compensate the acoustic models of the ASR system. Both closed form solutions for compensation of GMM-HMM [4] and DNN-HMM [6] acoustic models exist. In this work we limit ourselves however to GMM models. To evaluate the  $q$ -th mixture of a GMM model  $p(x^{\text{MFCC}}|q)$  when the features are uncertain, we can replace the conventional likelihood by the expected likelihood obtaining

$$E\{p(x^{\text{MFCC}}|q)\} = \mathcal{N}\left(\boldsymbol{\mu}^{\text{MFCC}}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}^{\text{MFCC}}\right). \quad (15)$$

## 5. Experimental Results

For reproducibility purposes, the AURORA4 [18] database was used both for the training of the DNN enhancement front-end and the ASR system. AURORA4 is a noisy version of the well known Wall-Street Journal medium vocabulary task of 5K words featuring read Journal news.

### 5.1. DNN Enhancement Training

In order to train a DNN enhancement system, parallel sets of corrupted and clean utterances are needed. For this purpose, the utterances corrupted only by additive noise in the multi-style and multi-noise training sets were used, amounting to a total of 5352 sentences. Compared to other DNN enhancement experiments on AURORA4, [9] only employed the multi-style set noises while [19] utilized both additive and channel distortions.

The DNNs were trained with log amplitude features utilizing both the noisy spectrum  $\mathbf{Y}$  and a noise estimate attained through IMCRA [20]. The geometry was similar to that of [19] employing 3 hidden layers of 2000 nodes. The DNN was trained using SGD [7] with mini-batches of 400 examples and a learning rate of 0.05. A simple random initialization following

Enhancement	01	02	03	04	05	06	07	av.	08	09	10	11	12	13	14	av.	t.av.
No Method	8.9	19.1	33.8	43.5	41.1	33.7	44.0	32.0	27.1	39.5	52.9	54.6	60.9	49.8	58.1	49.0	40.5
AFE	8.9	19.9	23.6	30.8	28.3	27.9	26.4	23.7	31.0	37.4	42.4	46.5	46.2	46.8	42.8	41.9	32.8
DNN (CE)	8.8	13.7	24.5	23.4	24.5	22.2	24.9	20.3	26.1	32.4	46.1	46.2	47.0	43.5	42.7	40.6	30.4
GM-DNN	8.5	13.6	22.6	22.1	22.9	20.7	25.4	19.4	25.6	32.6	42.3	42.2	45.3	41.6	41.8	38.8	29.1
DNN+UD	8.7	<b>12.4</b>	<b>20.6</b>	21.6	22.3	19.5	<b>21.9</b>	<b>18.1</b>	24.4	<b>31.8</b>	42.9	42.0	42.1	41.0	39.3	37.6	27.9
GM-DNN+UD	<b>8.1</b>	12.5	22.2	22.1	23.4	21.0	25.6	19.3	<b>24.2</b>	32.3	<b>41.4</b>	40.4	43.4	40.0	42.1	37.7	28.5
DNN (MSE)	8.4	15.4	25.1	23.4	24.5	21.7	25.2	20.5	28.0	35.6	45.8	45.5	44.7	43.9	43.8	41.0	30.8
DNN+UD	8.2	14.0	21.9	<b>20.7</b>	<b>21.5</b>	<b>18.8</b>	22.6	18.2	25.5	32.7	42.9	<b>40.0</b>	<b>41.2</b>	<b>39.1</b>	<b>38.9</b>	<b>37.2</b>	<b>27.7</b>

Table 1: AURORA4 enhanced clean speech training results in terms of WER. DNN enhancement trained with additive noises of multi-condition and multi-noise datasets. Best results highlighted in bold.

Enhancement	01	02	03	04	05	06	07	av.	08	09	10	11	12	13	14	av.	t.av.
No Method	13.7	12.0	17.2	21.9	19.6	17.5	22.2	17.7	24.1	25.8	33.9	35.3	37.5	32.7	36.4	32.2	25.0
AFE	<b>10.1</b>	13.0	18.2	21.8	20.2	18.0	20.3	17.4	23.1	26.6	32.3	34.8	36.1	32.9	34.0	31.4	24.4
DNN (CE)	10.6	11.4	15.8	19.2	17.4	15.5	<b>16.3</b>	15.2	22.3	<b>24.0</b>	35.7	40.0	38.3	35.2	35.1	32.9	24.1
GM-DNN	11.8	11.0	<b>13.3</b>	16.6	<b>16.5</b>	14.6	<b>16.3</b>	<b>14.3</b>	22.9	24.9	32.2	33.8	<b>34.5</b>	31.8	33.0	30.5	22.4
DNN+UD	10.4	10.9	14.9	18.3	18.4	14.6	17.6	15.0	21.5	25.3	33.2	36.2	36.4	33.4	34.0	31.4	23.2
GM-DNN+UD	11.5	10.9	<b>13.3</b>	<b>16.2</b>	<b>16.5</b>	<b>14.5</b>	17.1	<b>14.3</b>	22.0	25.6	<b>31.4</b>	<b>33.6</b>	<b>34.5</b>	32.1	33.5	30.4	<b>22.3</b>
DNN (MSE)	10.8	<b>10.7</b>	14.9	18.3	17.0	15.1	16.8	14.8	22.3	<b>24.0</b>	33.9	37.4	35.2	33.3	34.3	31.5	23.1
DNN+UD	10.4	10.8	14.3	17.9	16.8	14.7	16.4	14.5	<b>21.3</b>	25.2	32.4	35.1	34.3	<b>30.5</b>	<b>33.1</b>	<b>30.3</b>	22.4

Table 2: AURORA4 enhanced multi-style training results in terms of WER. DNN enhancement trained with additive noises of multi-condition and multi-noise datasets. Best results highlighted in bold.

[21] was used. To prevent over-fitting, a 20% of all the training data was held out for validation and a fixed number of 100 iteration was used.

As baselines, a MFCC front-end with no enhancement using Cepstral Mean Subtraction (CMS) per sentence, the Advanced Front End (AFE) [22] and a DNN speech enhancement front-end were provided. The DNNs were trained as previously described, using the CS and MSE criteria. The baselines were compared to the same models when propagating the uncertainty and compensating with uncertainty decoding (UD). Additionally for the CS-trained model, experiments accounting for the internal uncertainty of inference described in Section 3.2 were carried out. These are termed Gaussian Marginalization DNN (GM-DNN).

The methods tested can thus be summarized as per-sentence batch, due to the use of CMS and single pass. As mentioned by other authors, the DNNs have in-domain knowledge, a unfair advantage over non data-driven methods, such as the AFE. Results for these method are however provided for comparability purposes with other works. The main objective of the tests is to assess if the use of uncertainty can improve the already efficient DNN enhancement.

## 5.2. GMM-HMM ASR Training

The ASR system training was carried out using the Hidden Markov Toolkit (HTK) [23] and Vertannen’s recipe for WSJ0 [24]. Both clean training and multi-style training sets of 7138 sentences were used. The sampling frequency was 16KHz and speech enhancement was used during training. The test set used is AURORA4’s sennheiser microphone 166x14 sentence set. This includes 166 clean speech sentences and six corrupted versions using different additive noise. From this set of noises, the first can be regarded as relatively stationary whereas the rest are non-stationary. A second set recorded from a distant microphone is also available. A version of the HTK toolkit modified to optionally perform uncertainty decoding was used.

## 5.3. Analysis of the Results

Tables 1 and 2 contain the results for clean and mixed training, respectively. Each table is divided into two vertical blocks containing the baselines, No Method, AFE, DNN (CE) and DNN (MSE) respectively and the proposed observation uncertainty approaches. As it can be seen from the results, the use of UD consistently improves the overall performance of the DNN by around a 10% relative for the clean training case. This improvement drops to around 3.5% in the case of the multi-condition set. It has to be taken into account, however, that both the DNN and the trained GMM-HMM system are noise-matched in this case. In similar conditions, sources of uncertainty other than DNN ones as e.g. [13, 15] perform worse or offer no improvement at all.

One unexpected result is that the MSE-trained DNN, despite performing worse than the CE-trained one, attains larger performances when combined with UD in clean training conditions. The opposite happens in multi-style conditions.

Another interesting result concerns the internal DNN uncertainty approach (GM-DNN). The GM-DNN is consistently better than the normal DNN. However, as in the MSE case, it attains smaller improvements when combined with UD in the clean training setup.

## 6. Conclusions

We have explored methods to account for the residual uncertainty of inference when using IRM-based DNN enhancement. Experiments show that consistent improvements can be attained over the already efficient DNNs when accounting for this uncertainty. Improvements are achieved both when Uncertainty Decoding is used for dynamic compensation and when the Gaussian Marginalization approximation is used for enhancement. Future work could explore the extension of this approach to auto-encoding neural network enhancement schemes as well as situations of higher noise mismatch.

## 7. References

- [1] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.
- [2] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.
- [3] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proceedings of the Sensor Array and Multichannel Signal Processing Workshop*, Aug. 2002, pp. 209–213.
- [4] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 1–57–1–60 vol.1.
- [5] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.
- [6] R. F. Astudillo and J. P. Neto, "Propagation of Uncertainty through Multilayer Perceptrons for Robust Automatic Speech Recognition," in *Interspeech*, October 2011, pp. 461–464.
- [7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.
- [8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.
- [9] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.
- [10] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai, and C.-H. Lee, "Robust Speech Recognition with Speech Enhanced Deep Neural Networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [11] A. L. 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," in *INTERSPEECH*. Citeseer, 2012.
- [12] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks," in *Proceedings of the 2nd CHiME workshop on machine listening in multi-source environments*, 2013, pp. 86–90.
- [13] R. F. Astudillo and R. Orglmeister, "Computing MMSE Estimates and Residual Uncertainty directly in the Feature Domain of ASR using STFT Domain Speech Distortion Models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1023–1034, May 2013.
- [14] R. F. Astudillo, A. Abad, and I. Trancoso, "Accounting for the Residual Uncertainty of Multi-Layer Perceptron based Features," in *Proc. ICASSP 2014*, 2014, pp. 6909–6913.
- [15] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *2nd International Workshop on Machine Listening in Multisource Environments CHiME*, June 2013, pp. 33–38. [Online]. Available: 'http://something'
- [16] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.
- [17] D. Tran, E. Vincen, and D. Jouvét, "Fusion of multiple uncertainty estimators and propagators for noise robust ASR," in *Proceedings of the ICASSP 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing*, apr 2014, pp. 5512–5516.
- [18] G. Hirsch, *Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task*, Niederrhein University of Applied Sciences, November 2002.
- [19] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 65–68, 2014.
- [20] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.
- [21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.
- [22] *ETSI Standard document, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 050 v1.1.5 (2007-01)*, ETSI, January 2007.
- [23] S. Young, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department., 2006.
- [24] K. Vertanen, "HTK Wall Street Journal Training Recipe," 2006.