Interspeech 2018
2-6 September 2018, Hyderabad

# Global SNR Estimation of Speech Signals using Entropy and Uncertainty Estimates from Dropout Networks

*Rohith Aralikatti, Dilip Kumar Margam, Tanay Sharma, Abhinav Thanda, Shankar Venkatesan*

Samsung R&D Institute Bangalore, India

r.aralikatti@samsung.com, dilip.margam@samsung.com, tanay.sharma@samsung.com, abhinav.t89@samsung.com, s.venkatesan@samsung.com

## Abstract

This paper demonstrates two novel methods to estimate the global SNR of speech signals. In both methods, Deep Neural Network-Hidden Markov Model (DNN-HMM) acoustic model used in speech recognition systems is leveraged for the additional task of SNR estimation. In the first method, SNR is estimated using the entropy of the posterior distribution obtained from DNN of an ASR system. Recent work on bayesian deep learning has shown that a DNN-HMM trained with dropout can be used to estimate model uncertainty by approximating it as a deep Gaussian process. In the second method, this approximation is used to obtain model uncertainty estimates. Noise specific regressors are used to predict the SNR from the entropy and model uncertainty. The DNN-HMM is trained on GRID corpus and tested on different noise profiles from the DEMAND noise database at SNR levels ranging from -10 dB to 30 dB.

**Index Terms**: SNR estimation, Entropy, Bayesian Uncertainty, Dropout, Neural Networks

## 1. Introduction

Signal-to-noise ratio (SNR) estimation of a signal is an important step in many speech processing techniques such as robust automatic speech recognition (ASR) [1, 2], speech enhancement [3, 4], noise suppression and speech detection.

The global signal-to-noise ratio (SNR) of a signal $x(t)$ in dB is defined as follows.

$$SNR_{dB}(x) = 10 \log_{10} \frac{Power(s)}{Power(n)} \tag{1}$$

The signal $x(t) = s(t) + n(t)$ where $s(t)$ represents the clean signal and $n(t)$ is the noise component.

State-of-the-art ASR has achieved very low error rates with the advent of deep learning. However, performance of ASR systems can still be improved in noisy conditions. Robust ASR techniques such as noise-aware training [1] and related methods [5, 2] require an estimate of the noise present in the speech signal.

Recently, it has been shown that incorporating visual features (extracted from lip movements during speech) can lead to improved word error rates (WER) during noisy environment [6, 7]. In [8], both audio and visual modalities are used for speech enhancement. With the proliferation of voice assistants and front facing cameras in smartphones, using visual features to improve ASR seems feasible. This raises the crucial question - when should the camera be turned on to make use of features from the visual modality? In such scenarios, we can benefit from accurate SNR estimation by turning on the camera in noisy environments.

In this paper, we present two novel methods to estimate the global SNR (at an utterance level) of a speech signal. Both methods require training a DNN based speech classifier on noise free audio using alignments generated from a GMM-HMM model trained for ASR. The first method estimates SNR by computing the entropy of the DNN's output. The second method uses model uncertainty estimates obtained by using dropout during inference as shown in [9]. In section 2, we present related work that has been done. Section 3 describes the entropy based SNR estimator. Section 4 describes the dropout based SNR estimator. Section 5 describes the architecture of the network, the training procedure and the experiments done. Section 6 presents the results of the paper. The final section 7 has the conclusion.

## 2. Related Work

SNR estimation has been an active area of research. In [10], the authors use specific handcrafted features such as signal energy, signal variability, pitch and voicing probability to train noise specific regressors that compute SNR of an input signal. In [11], the amplitude of clean speech is modelled by a gamma distribution and noise is assumed to be normally distributed. SNR is estimated by observing changes to the parameters of the gamma distribution upon addition of noise.

The NIST-SNR measurement tool uses a sequential GMM to model the speech and non-speech parts of a signal to estimate the SNR. In [12], a voice activity detector (VAD) is used to classify frames as either voiced, unvoiced or silence and the noise spectrum is estimated from this information. After subtracting the noise spectrum from the input signal to obtain the clean signal, SNR is estimated. In [13], computational auditory scene analysis is used to estimate speech dominated and noise dominated portions of the signal in order to obtain SNR.

Estimation of instantaneous SNR is also a subtask in many speech enhancement methods [8, 14, 15, 16]. In [17], a neural network is trained to output the SNR in each frequency channel using amplitude modulation spectrogram (AMS) features which are obtained from the input signal. In [18], the peaks and valleys of the smoothened short time power estimate of a signal are used to estimate the noise power and instantaneous SNR.

In [19], entropy of the softmax output of the DNN-HMM classifier has been used as an uncertainty measure for stream selection before decoding is done. In [20], entropy is used as a weighting factor to combine predictions from multiple sub-band HMM outputs. In [21], inverse entropy weighting is used to combine the output of a multi-stream DNN-HMM system. However, they have not studied the estimation of SNR using entropy and dropout uncertainty estimates.

 10.21437/Interspeech.2018-1884

# 3. Entropy Based SNR Estimation

In this method, a neural network which is trained as a part of ASR system to predict the posterior distribution of HMM states is used. The Shannon entropy of the posterior distribution is computed. In information theory, Shannon entropy is realization of the average uncertainty of encoding machine. Similarly, in our case the posterior distribution obtained from DNN which is trained as a part of ASR system, acts as an encoding distribution for encoding machine. Whenever the feature vector of clean signal is forwarded through DNN it is expected to give meaningful posterior distribution. But when a feature vector of a noisy signal is forwarded through the neural network, the posteriors are expected to be arbitrary, which in most cases lead to higher entropy value. This comes from the assumption that addition of noise to the speech signal results in arbitrary features.

Let $\mathbf{F}_i$ denote the $i^{th}$ input frame of utterance $\mathbf{U}$ (consisting of $m$ audio frames) and $\mathbf{Y}$ (of dimension $d$) denote the output of DNN. The entropy for given input $\mathbf{F}_i$ is computed as shown in equation 2.

$$H(\mathbf{F}_i) = -\sum_{i=0}^{d} P[\mathbf{Y}_i] \log P[\mathbf{Y}_i] \tag{2}$$

$$Entropy(\mathbf{U}) = \sum_{i=0}^{m} \frac{H(\mathbf{F}_i)}{m} \tag{3}$$

$$SNR(\mathbf{U}) = f_1(Entropy(\mathbf{U})) \tag{4}$$

Where $P[.]$ denotes softmax activation, $\mathbf{Y}_i$ is $i^{th}$ dimension of $\mathbf{Y}$. The average entropy of all input frames for a given utterance is used as a measure of the entropy for an utterance. A polynomial regressor $f_1(.)$ is trained on utterance level entropy values to predict the SNR of speech signal. The advantage of this method is that it can work on any kind of noise which can randomize the speech signal. The DNN-HMM based ASR systems which are sensitive to noisy conditions, can take advantage of entropy values to estimate the SNR with low computational overhead. In Figure 1, it is clearly seen that with increase in noise, the average entropy increases.

# 4. SNR Estimation Using Dropout Uncertainty

## 4.1. Bayesian uncertainty using dropout

Gal and Ghahramani showed in [9] that the use of dropout while training DNNs can be thought of as a bayesian approximation of a deep Gaussian process (GP). Using the above GP approximation, estimates for the model uncertainty of DNNs trained using dropout are derived. More specifically, it is shown that uncertainty of the DNN output for a given input can be approximated by computing the variance of multiple output samples obtained by using dropout during inference. The use of dropout during inference, results in different output every time the forward pass is done, for a given input. The variance of these output samples is the uncertainty for the given input.

The above method is used to obtain uncertainty estimates for the DNN that was trained as a part of DNN-HMM based ASR system as explained in section 6. This DNN is referred as dropout network throughout this paper. If the input is corrupted by noise, it is expected that the model uncertainty derived from dropout will be higher. The model uncertainty for given input $\mathbf{F}_i$ is computed as shown in equation 5.

$$MU(\mathbf{F_i}) = \sum_{i=0}^{d} Var[\mathbf{Y}_i] \tag{5}$$

$$uncertainty(\mathbf{U}) = \sum_{i=0}^{m} \frac{MU(\mathbf{F}_i)}{m} \tag{6}$$

$$SNR(\mathbf{U}) = f_2(uncertainty(\mathbf{U})) \tag{7}$$

$$SNR(\mathbf{U}) = f_3(uncertainty(\mathbf{U}), Entropy(\mathbf{U})) \tag{8}$$

Where $MU$ stands for model uncertainty per frame. The average variance over all input frames is used as a measure of uncertanity for an utterance. The SNR of the utterance is estimated as shown in equation 7, where $f_2(.)$ is polynomial regressor trained to predict SNR from uncertainty value. The regressor $f_3(.)$ is trained on both uncertainty and entropy of utterance to output SNR value. We have compared the performance of all three regressors in table 1.

The assumption made here is that test data that is corrupted by noise is similar to the train data. Since the uncertainty computed above comprises of both model uncertainty (confidence of the model regarding the prediction) as well as the input uncertainty, it is possible to get high uncertainty values for clean speech signals that are dissimilar to train data. However, training the DNN-HMM classifier on large amounts of varied speech data should solve this problem.

## 4.2. Fast dropout uncertainty estimation

It may not always be feasible to run the forward pass multiple times per input frame in order to obtain output samples. Given the input frame and the weights of the dropout network, it should be possible to algebraically derive the variance and expectation of the output layer.

The uncertainty of the model is the consequence of uncertainty added because of dropout in each layer of network. Following equation depicts how the uncertainty of model can be computed mathematically. For mathematical simplicity let us consider the neural network with one layer. The output of the one layer network (with weight $\mathbf{W}$ and bias $\mathbf{b}$) with ReLU activation function is: $\mathbf{Y} = ReLU(\mathbf{W} \cdot (\mathbf{D} \circ \mathbf{F}) + \mathbf{b})$. Where $\circ$ denotes hadamard product, $\mathbf{D}$ denotes the dropout mask. The variance of $i^{th}$ dimension of output is given as shown in equation 9.

$$\begin{aligned} Var[\mathbf{Y}_i] &= Var[ReLU(\mathbf{W}_i^T(\mathbf{D} \circ \mathbf{F}) + \mathbf{b}))] \\ &= Var[ReLU(\sum_{j=0}^{m-1} \mathbf{W}_{ij}\mathbf{D}_j\mathbf{F}_j)] \\ &= Var[ReLU(A_i)] \end{aligned} \tag{9}$$

Where $A_i = \sum_{j=0}^{m-1} \mathbf{W}_{ij}\mathbf{D}_j\mathbf{F}_j$. $\mathbf{W}_i$ denote $i^{th}$ row of matrix $\mathbf{W}$, $m$ is the dimension of $F$. The dropout variable $\mathbf{D}_i$ being a bernoulli variable with probability of success $p$, $Var[D_i] = p(1-p)$.

$$\begin{aligned} Var[A_i] &= \sum_{j=0}^{m-1} \mathbf{W}_{ij}^2\mathbf{F}_j^2 Var[D_j] \\ &= p(1-p)\sum_{j=0}^{m-1} \mathbf{W}_{ij}^2\mathbf{F}_j^2 \end{aligned} \tag{10}$$

Since all the dropout bernoulli random variables are independent of one another, the equation 10 follows. The difficulty comes in computing the $Var[\mathbf{Y}_i]$ because it involves a non-linear ($Relu$) activation function. To compute the $Var[\mathbf{Y}_i]$ one has to integrate the $\mathbf{Y}_i$s over all possible dropout distributions ($2^m$ possibilities), which will increase the computational complexity. One can proceed from here using the Taylor first order approximation of $m$ variables. In [22] it is assumed that sum of activation values follows normal distribution following the central limit theorem, but this assumption did not hold good empirically in our case because of multiple layers in network.

However, the variance of the output is some complex non-linear function of the input and the dropout network weights. Therefore it must be possible to train another DNN to learn this non-linear relationship so that the uncertainty can be estimated by a single forward pass of this second network. This second neural network from now on will be referred to as the *variance network* in this paper. The variance network explained in section 5.2 was able to successfully learn the mapping from the input frame to the output (dropout uncertainty), as shown in the Figure 3.

# 5. Experiments

A DNN-HMM based ASR system is trained on the Grid corpus [23] (95% of it is used for training, 5% for testing), which has 34 speakers and 1000 utterances per speaker. The Mel scale filter-bank features of 40 dimension, with 5 contextual frames on both sides are used as input features. The duration of 25 ms and shift of 10 ms is used in feature extraction process. The activation function used is ReLU, along with dropout with $p = 0.2$ ($p$ is probability of dropping a neuron) is used in all hidden layers. The output of DNN is of dimension 1415 corresponding to number of HMM states. There are six hidden layers with 1024 neurons in each layer. This DNN which is also referred to as *dropout network* in this paper is used for estimating entropy and variance in all our experiments, except for section 5.2.

## 5.1. Entropy method and dropout uncertainty method

We experimented on 16 different noise types from the DEMAND noise dataset, where noise is added to the test set of utterances. We observe that there is a strong correlation between average entropy and SNR as shown in Figure 1. Similar kind of results for average dropout uncertainty estimates versus the SNR are obtained, where model uncertainty increases with increase in noise as shown in the Figure 2. The uncertainty/entropy values shown in Figure 1 and 2 were obtained by averaging 100 test utterances per noise type per dB level.

The variance has been computed by taking 100 output samples per input frame, but we obtained similar results when we reduced the number of samples to 20 per input frame. Figure 2 shows the variation in model uncertainty with respect to SNR for same six arbitrarily chosen noises as in Figure 1.

The variance computation was done on the output samples obtained from the DNN before the application of softmax to obtain probabilities. This gave better results, since the softmax function tends to exponentially squash the outputs to lie between 0 and 1 and this causes the variance along many of the dimensions of the output to be ignored.

Using the ReLU non-linearity also gave better results as compared to the sigmoid and tanh non-linearity. This is expected, as both the sigmoid and tanh tend to saturate and this does not allow the variance (or model uncertainty) to be propagated to the output layer.

## 5.2. Variance network (fast dropout uncertainty estimation)

This is the network used for fast dropout uncertainty estimation. The variance network is trained on uncertainty estimates obtained from the dropout network. The training is done on utterances from the GRID corpus mixed with noise from the DEMAND [24] dataset using the previously trained dropout network. The training is done on utterances mixed with 12 types of noise at 40 different SNR levels (from -10 dB to 30 dB). The testing is done on different utterances from the GRID corpus mixed with noise samples not exposed to the network during training.

Variance network is able to successfully learn the mapping from the input frame to the output uncertainty. The plots shown in the Figure 3 shows the variation of output uncertainty for the four types of noise (CAR, PARK, KITCHEN, MEETING) which were not used during training.

## 5.3. Online SNR estimation

The methods discussed in this paper are for the global SNR estimation of an entire speech utterance. In an online ASR setting, we will not have access to all the frames of the utterance. Entropy averaging over limited number of available frames also shows a similar trend for entropy vs SNR. As a result of this, sudden increase in noise can be identified in real time speech to text systems. In this section, we plot the average entropy value (averaged over 50 consecutive frames, with a shift of 25 frames) during online decoding as a function of time. In the chosen audio file, the same noise free utterance is mixed with noise at SNR values ranging from 20 dB to -10 dB and these files are concatenated together to form one long audio file. The local SNR value starts from 20 dB at the start of the audio file and reaches -10 dB at the end of the file. Thus, entropy averaging in an online setting also closely mirrors the local SNR of the signal as seen in Figure 4.

# 6. Results

To obtain the SNR of an input signal, we have trained noise specific regressors to obtain the SNR value given the uncertainty obtained from variance network and/or entropy. The mean-absolute-error (MAE) for three different types of noise at different SNR levels are shown in Table 1.

We have compared the result of the three regressors ($f_1$,$f_2$ and $f_3$) described previously with well known SNR estimation methods, namely the NIST STNR estimation tool and the WADA SNR estimation method described in [12]. It is observed that the regressor trained on dropout uncertainty performed better than the entropy based regressor. Indeed, it is observed that the regressor trained on both the dropout uncertainty and entropy perfomed worse than just regressing on the network uncertainty. However, all three regressors have produced better SNR estimates than either WADA or NIST, particularly at low SNR levels.

Though we clearly see a correlation between the entropy/dropout uncertainty and the noise in the signal, to finally obtain the SNR value of the signal we have to train a noise specific regressor on top of the entropy/dropout uncertainty values. The possibility of directly predicting SNR independent of the background noise is something that needs further research. In [10], the authors propose using a DNN to find out which of the

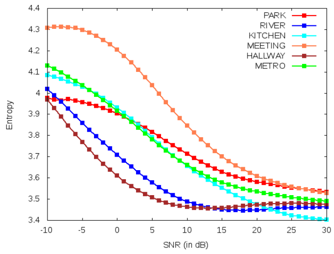


Figure 1: *Plot depicts the relationship between averaged entropy of utterance (defined in equation 3) with SNR value of utterance for test utterances for six arbitrarily chosen noise types.*

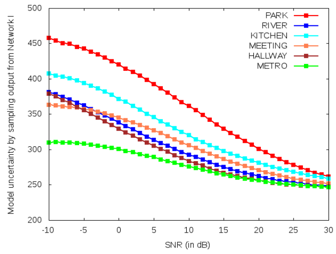


Figure 2: *Figure shows the relationship between averaged uncertainty of utterance (as in equation 6) and SNR value of utterance for test utterances for six arbitrarily chosen noise types.*

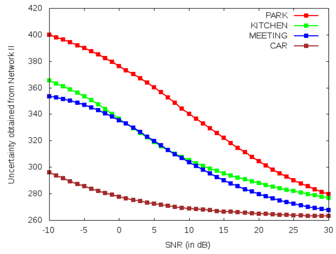


Figure 3: *Figure shows the relationship between output of variance network and noisy input speech with different SNR values for four unseen (not used in training) noises.*

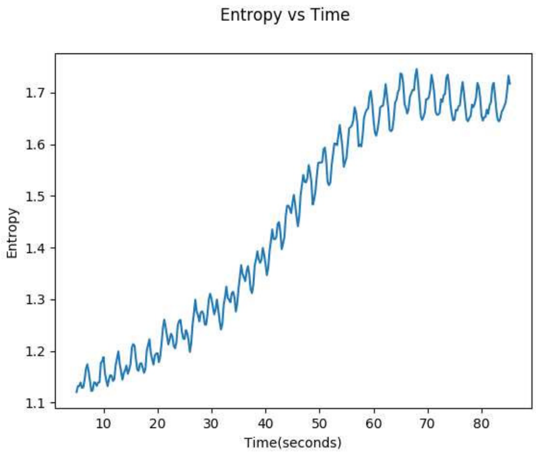


Figure 4: *Plot depicts the relationship between averaged entropy (over 50 consecutive frames) with time for an audio file whose local SNR varies from 20 dB (start of the audio file) to -10 dB (end of the audio file).*

Table 1: *The Mean Absolute Error (MAE) of our SNR estimation methods is compared against pre-existing methods*

| Noise type | Method | SNR (dB) -10 | -5 | 0 | 5 | 10 |
|---|---|---|---|---|---|---|
| DKITCHEN | NIST | 15.55 | 10.58 | 6.66 | 5.08 | 4.73 |
| | WADA | 9.34 | 5.35 | **1.31** | **0.93** | **0.67** |
| | $f_1$ | 2.12 | 3.20 | 3.21 | 2.76 | 2.46 |
| | $f_2$ | **1.31** | **1.72** | 1.82 | 1.99 | 0.91 |
| | $f_3$ | 3.08 | 2.85 | 3.57 | 4.34 | 4.02 |
| NPARK | NIST | 17.32 | 12.64 | 8.71 | 6.94 | 6.91 |
| | WADA | 7.83 | 4.13 | 2.31 | 1.89 | 2.25 |
| | $f_1$ | **1.97** | 2.37 | 2.9 | 2.3 | 1.5 |
| | $f_2$ | 2.34 | **2.01** | **1.86** | 1.62 | **1.28** |
| | $f_3$ | 2.43 | 2.11 | 1.9 | **1.61** | 1.35 |
| OMEETING | NIST | 17.25 | 12.97 | 10.46 | 9.26 | 11.3 |
| | WADA | 12.11 | 8.44 | 6.61 | 6.08 | 6.39 |
| | $f_1$ | **1.67** | 1.48 | **1.63** | **1.54** | **1.45** |
| | $f_2$ | 1.98 | 1.46 | 1.79 | 1.98 | 1.95 |
| | $f_3$ | 2.32 | **1.24** | 1.68 | 2.12 | 2.12 |

noise types most closely resemble the input and use the corresponding regressor to estimate SNR.

However, since dropout network is trained on clean audio, irrespective of the type of noise in the speech signal, the trend of increasing uncertainty with increasing noise did hold even in unseen noise conditions. The variance network, which is trained on specific noise types in order to avoid the computational costs of taking samples during inference, clearly maintained this trend even in unseen noise conditions as shown in figure 3

## 7. Conclusion

In this paper, we have shown that it is possible to extract useful information from the uncertainty (either from entropy or from bayesian estimates) and predict the SNR of a speech signal. Using the above uncertainty information to better design and improve the performance of current ASR and speech enhancement algorithms will be possible future directions of research. Another possible improvement that can be done is to investigate the possibility of predicting instantaneous SNR instead of global SNR. The methods proposed in this paper for SNR estimation do not impose any conditions on the type of noise corrupting the signal. This leaves open the possibility of applying similar noise estimation techniques to non-speech signals.

## 8. References


[1] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[2] K. H. Lee, S. J. Kang, W. H. Kang, and N. S. Kim, "Two-stage noise aware training using asymmetric deep denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5765–5769.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] E. Nemer, R. Goubran, and S. Mahmoud, "Snr estimation of

speech signals using subbands and fourth-order statistics," *IEEE Signal Processing Letters*, vol. 6, no. 7, pp. 171–174, 1999.

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] A. Thanda and S. M. Venkatesan, "Multi-task learning of deep neural networks for audio visual automatic speech recognition," *arXiv preprint arXiv:1701.02477*, 2017.

[7] ——, "Audio visual speech recognition using deep recurrent neural networks," in *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*. Springer, 2016, pp. 98–109.

[8] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 629–632.

[9] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[10] P. Papadopoulos, A. Tsiartas, and S. Narayanan, "Long-term snr estimation of speech signals in known and unknown channel conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2495–2506, 2016.

[11] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[12] J. A. Morales-Cordovilla, N. Ma, V. Sánchez, J. L. Carmona, A. M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4808–4811.

[13] A. Narayanan and D. Wang, "A casa-based system for long-term snr estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2518–2527, 2012.

[14] J. Lee, K. Kim, T. Shabestary, and H.-G. Kang, "Deep bidirectional long short-term memory based speech enhancement for wind noise reduction," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 41–45.

[15] I. Cohen, "Relaxed statistical model for speech enhancement and a priori snr estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[16] Y. Ren and M. T. Johnson, "An improved snr estimator for speech enhancement," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4901–4904.

[17] J. Tchorz and B. Kollmeier, "Snr estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 184–192, 2003.

[18] R. Martin, "An efficient algorithm to estimate the instantaneous snr of speech signals." in *Eurospeech*, vol. 93, 1993, pp. 1093–1096.

[19] O. Malladi and Hermansky, "Uncertainty estimates of dnn classifiers," in *Automatic Speech Recognition and understanding (ASRU)*, 2015.

[20] B. Okawa and Potamionas, "Multi-band speech recognition in noisy environments," in *Proceedings of the 1998 IEEE International Conference on Acoustic, Speech and Signal Processing*, 1998.

[21] B. Misra and Tyagi, "New entropy based based combination in hmm/ann multi stream asr," in *Proceedings of the 2003 IEEE International Conference on Acoustic, Speech and Signal Processing*, 2003.

[22] S. Wang and C. Manning, "Fast dropout training," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 118–126.

[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[24] J. Thiemann, N. Ito, and E. Vincent, "Demand: Diverse environments multichannel acoustic noise database (2013)."