A PROPAGATION APPROACH TO MODELLING THE JOINT DISTRIBUTIONS OF CLEAN AND CORRUPTED SPEECH IN THE MEL-CEPSTRAL DOMAIN

Ramón Fernadez Astudillo

Spoken Language Systems Laboratory, INESC-ID-Lisboa, Lisboa, Portugal

ramon@astudillo.com

ABSTRACT

This paper presents a closed form solution relating the joint distributions of corrupted and clean speech in the short-time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficient (MFCC) domains. This makes possible a tighter integration of STFT domain speech enhancement and feature and model-compensation techniques for robust automatic speech recognition. The approach directly utilizes the conventional speech distortion model for STFT speech enhancement, allowing for low cost, single pass, causal implementations. Compared to similar uncertainty propagation approaches, it provides the full joint distribution, rather than just the posterior distribution, which provides additional model compensation possibilities. The method is exemplified by deriving an MMSE-MFCC estimator from the propagated joint distribution. It is shown that similar performance to that of STFT uncertainty propagation (STFT-UP) can be obtained on the AURORA4, while deriving the full joint distribution.

Index Terms: Speech Enhancement, Uncertainty Propagation, Uncertainty Decoding, Modified Imputation

1. INTRODUCTION

A majority of the single and multi-channel speech enhancement methods operate in the short-time Fourier transform (STFT) domain. This is so because in this domain assumptions like source additivity or signal sparsity hold well, and phenomena like late reverberation are relatively easy to model. Automatic speech recognition (ASR) systems work on domains that are non-linear transformations of the STFT domain, such as the Mel-Frequency Cepstral Coefficients (MFCCs). Such domains provide compact representations of the acoustic space, and thus lead to smaller and more accurate models than the STFT domain. They are, however, ill suited for speech enhancement, due to their non-linear nature.

Throughout the years, various approaches have been proposed to simultaneously exploit the properties of STFT and non-linear feature domains used in ASR such as MFCCs. The best known approach nowadays is the approximation of mismatch functions relating channel and additive distortions in the STFT and MFCC domains [1]. For example Vector Taylor Series (VTS) compensation [2] uses truncated Taylor series to approximate a mismatch function that ignores the effect of speech and noise phases. ALGONQUIN [3] builds on this approach by taking into account the phase information, and modeling it as a residual uncertainty. Other methods like MBFE [4], use similar approximations for model-based feature enhancement.

An alternative to approximating non-linear mismatch functions is approximating the propagation of uncertainty through the non-linearity [5]. In this case, rather than considering a deterministic relation between speech and noise in the STFT domain, a probabilistic model relating both is used. By approximating the random variable change from the STFT to the non-linear domain, speech distortion in both domains can be related. The probabilistic model in the STFT domain can be obtained from standard STFT domain speech enhancement [6, 7, 8], or missing feature frameworks in the STFT domain [9].

One of the limitations of uncertainty propagation approaches is that they either provide a point-estimate in the feature domain [7, 10], or a measure of estimation uncertainty in the form of a posterior distribution of the features [8]. On the contrary, methods based on mismatch functions often allow more complete models in the form of likelihoods [3], or the full joint distribution [11] relating corrupted and clean speech.

This paper proposes a propagation technique that estimates the joint distribution of corrupted and clean speech in the MFCC domain from the conventional model used for speech enhancement in the STFT domain. The proposed method follows the same approach as STFT Uncertainty Propagation (STFT-UP) [8], but propagates the joint distribution of corrupted and clean speech, rather than a posterior distribution of the clean features. This allows to derive not only a posterior as in STFT-UP, but also a likelihood, thus making possible the use of Front-End Joint Uncertainty Decoding (FE-JUD) [11], and similar techniques. The approach is therefore termed STFT Joint Uncertainty Propagation (STFT-JUP), following the analogy between STFT-UP

This work was supported by the Portuguese Foundation for Science and Technology through grant number SFRH/BPD/68428/2010 and project PEst-OE/EEI/LA0021/2013.

and uncertainty decoding (UD) [12].

To exemplify the method, the posterior attained from STFT-JUP is compared with that of STFT-UP in robust ASR experiments on the AURORA4 task. The paper is structured as follows: Section 2 details the conventional speech distortion model in the STFT domain used for speech enhancement; Section 3 discusses the assumption of the jointly-Gaussian distortion model in the MFCC domain, and introduces the STFT-JUP method to estimate its parameters; Section 4 introduces the experimental setup, and finally Section 5 provides the conclusions.

2. MODELING SPEECH DISTORTION IN THE STFT DOMAIN

The majority of STFT domain techniques aimed at improving the acoustic quality of corrupted speech are based in the complex Gaussian model of speech distortion. Let $y(n)$ and $x(n)$ denote corrupted and clean speech signals, respectively, and $\mathbf{Y}$ and $\mathbf{X}$ their respective complex valued STFT matrices. Let also k and l denote frequency and analysis frame indices, respectively. The complex Gaussian model for speech distortion in the STFT domain implies the following assumptions. First, each Fourier coefficient of the observable corrupted speech signal Y_{kl} corresponds to the sum

$$Y_{kl} = X_{kl} + D_{kl}, \tag{1}$$

where X_{kl} is the hidden Fourier coefficient of the clean speech, and D_{kl} is a hidden distortion. Second, all Fourier coefficients are statistically independent, and their a priori distributions are circular symmetric complex Gaussian

$$X_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{kl}^{X}\right), \tag{2}$$

$$D_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{kl}^{D}\right). \tag{3}$$

The distortion D_{kl} can be used to model interfering phenomena that are independent of X_{kl}, such as background noises [13] or late reverberation [14]. It is however also used in beamforming and blind source separation post-processing (see e.g. [15]). Given the noisy signal $y(n)$, the complex Gaussian model is completely determined once the variances of each Fourier coefficient λ_{kl}^{X} and λ_{kl}^{D} are computed. These variances are estimated with different models, depending on the type of distortion. For additive noises, a voice activity detector is typically used. Other more complex set-ups may use spatial information, such as beamforming or late reverberation models. It should be noted that these methods operate separately from the ASR systems and imply comparatively low computational costs. They are also often causal, meaning that they can output the result, as each frame is processed.

Once the variances of the model have been determined, MMSE estimators like Wiener e.g. [16], amplitude (MMSE-STSA) [17] and log-amplitude (MMSE-LSA) [18] can be employed to estimate the clean STFT. In a conventional robust ASR architecture with speech enhancement pre-processing, such point-estimate would be directly fed to the feature extraction stage of the ASR system. The objective of this work is, however, to propagate the statistical relation implied by equations (1), (2) and (3) into the MFCC domain, prior to performing any estimation. In what follows we will then consider the joint distribution of each corrupted and clean Fourier coefficient. For the complex Gaussian model here presented is straightforward to see that this is given by

$$\begin{bmatrix} Y_{kl} \\ X_{kl} \end{bmatrix} \sim \mathcal{N}_{\mathbb{C}^2}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_{kl}^{X} + \lambda_{kl}^{D} & \lambda_{kl}^{X} \\ \lambda_{kl}^{X} & \lambda_{kl}^{X} \end{bmatrix}\right). \tag{4}$$

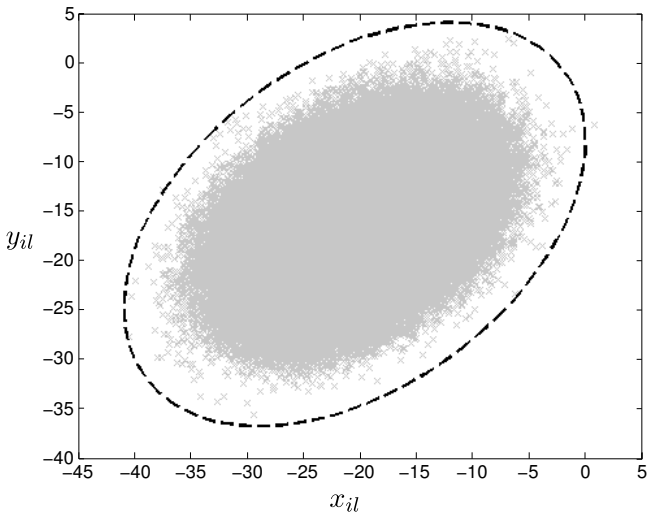

Fig. 1. *Joint probability distribution of corrupted and clean speech in the MFCC domain at 20dB segmental SNR. Monte Carlo approximation for STFT complex Gaussian models (Grey crosses). Scaled covariance contour for STFT-JUP estimated parameters (dashed ellipse).*

3. JOINT UNCERTAINTY PROPAGATION INTO THE MFCC DOMAIN

3.1. On the Gaussian distortion model in the MFCC Domain

Let the power based MFCCs of a clean signal be defined by

$$x_{il} = \sum_{j=1}^{J} C_{ij} \log\left(\sum_{k=1}^{K} W_{jk}|X_{kl}|^2\right), \tag{5}$$

where x_{il} is the i^{th} MFCC of the l^{th} analysis frame, W_{jk} are the weights of the Mel-filterbank channel, and C_{ij} are the weights of the discrete cosine transform (DCT), truncated to the first 13 coefficients. Delta and acceleration coefficients are also usually computed from x_{il} and appended. The MFCC of the corrupted speech y_{il} is computed in analogous form from Y_{kl}. Since according to the joint distribution (4), X_{kl}

and Y_{kl} are correlated, x_{il} and y_{il} will also be statistically dependent. The main premise of the approach presented here is that this joint distribution can be modeled by the Gaussian distribution

$$\begin{bmatrix} y_{il} \\ x_{il} \end{bmatrix} \sim \mathcal{N}_{\mathbb{R}^2} \left(\begin{bmatrix} \mu_{il}^y \\ \mu_{il}^x \end{bmatrix}, \begin{bmatrix} \Sigma_{il}^y & \Sigma_{il}^{yx} \\ \Sigma_{il}^{yx} & \Sigma_{il}^x \cdot \end{bmatrix} \right). \quad (6)$$

Jointly-Gaussian distortion models in the MFCC domain have been used in the past (see e.g. SPLICE [12] or JUD [11]). In particular, the Gaussianity of the joint distribution has been questioned in [11, Sec. 5.1], but it should be clear that this is a different case than the one considered here. On the one hand, the experiment in [11] assumes uniform and Gaussian distributions in the MFCC domain for speech and distortion respectively, and utilizes a simplified mismatch function. The case here presented departures from the assumption of complex Gaussian distributions for speech and noise in the STFT domain, as given by (1), (2) and (3). As shown in Fig. 1, Monte Carlo simulation experiments under these assumptions and the hypothesized joint Gaussian distribution computed from STFT-JUP match quite well.

On the other hand, it is important to differentiate between modeling the relation of corrupted and clean speech for an entire region of the acoustic space, as SPLICE or JUD does, and modeling this relation for a particular STFT frame. The latter case models the uncertainty over the value of a particular clean Fourier coefficient X_{kl} conditioned on the available information about that coefficient, i.e. the a priori parameters λ_{kl}^X, λ_{kl}^D and the observed Y_{kl}. For these conditions, the complex Gaussian assumption is considered to yield an acceptable trade-off between modeling accuracy and simplicity [19]. It can be also expected that the joint distribution in the MFCC domain with respect to that particular prior information has a simpler form than a joint distribution conditioned on e.g. the state of an HMM, which refers to a multiplicity of frames and can be often multi-modal.

3.2. Estimating the parameters of the Joint distribution with STFT-JUP

The objective is thus to derive the parameters of the joint distribution in the MFCC domain μ_{il}^y, Σ_{il}^y, μ_{il}^x, Σ_{il}^x, Σ_{il}^{yx} from the parameters of the joint distribution in the STFT domain λ_{kl}^X λ_{kl}^D. This is equivalent to solving the random variable change of (4) through the MFCC transformation (5) for X_{kl} and Y_{kl}.

A closed form solution for this variable change can be obtained based on only two assumptions.

1. The STFT complex Gaussian model presented in Section 2 holds.

2. The corrupted and clean Mel-filterbank features are jointly log-normal, which leads to joint-Gaussianity in the MFCC domain.

Under these assumptions, obtaining the parameters of the marginal distributions $p(y_{il})$, $p(x_{il})$ from $p(\mathbf{Y}_l)$, $p(\mathbf{X}_l)$ is just a particular case of the posterior propagation. These parameters can thus be derived by using the STFT-UP formulas in [8], for a zero mean posterior and $p = 2$. For the clean MFCC x_{il}, the variance is given by

$$\begin{aligned} \Sigma_{il}^x \approx & \sum_{j=1}^{J} \sum_{j'=1}^{J} C_{ij} C_{ij'} \log \left(\sum_{k=1}^{K} W_{jk} W_{j'k} \left(\lambda_{kl}^X \right)^2 \right) \\ & - \sum_{j=1}^{J} \sum_{j'=1}^{J} C_{ij} C_{ij'} \\ & \cdot \log \left(\sum_{k=1}^{K} \sum_{k'=1}^{K} W_{j'k} W_{jk'} \lambda_{kl}^X \lambda_{k'l}^X \right) \end{aligned} \quad (7)$$

and the mean by

$$\mu_{il}^x \approx \sum_{j=1}^{J} C_{ij} \log \left(\sum_{k=1}^{K} W_{jk} \lambda_{kl}^X \right) - \frac{1}{2} \Sigma_{il}^x. \quad (8)$$

The formulas for the corrupted MFCCs y_{il} may be computed in an analogous form.

The only missing parameters are the covariances Σ_{il}^{yx}. These can also be computed in a similar fashion by first noting that the covariance between STFT squared amplitudes can be derived from [20, Eq. 2.7] as

$$\Sigma_{kl}^{|Y|^2 |X|^2} = \left(\lambda_{kl}^X \right)^2, \quad (9)$$

and, due to the joint-log normality assumption, we have

$$\begin{aligned} \Sigma_{il}^{yx} \approx & \sum_{j=1}^{J} \sum_{j'=1}^{J} C_{ij} C_{ij'} \log \left(\sum_{k=1}^{K} W_{jk} W_{j'k} \left(\lambda_{kl}^X \right)^2 \right) \\ & - \sum_{j=1}^{J} \sum_{j'=1}^{J} C_{ij} C_{ij'} \\ & \cdot \log \left(\sum_{k=1}^{K} \sum_{k'=1}^{K} W_{j'k} W_{jk'} \lambda_{kl}^Y \lambda_{k'l}^X \right). \end{aligned} \quad (10)$$

Although the resulting matrix $\mathbf{\Sigma}_l^{yx}$ is full, only the diagonal for each frame is used, in order to simplify computations. Note also that to reduce the computational load, the full covariance after the Mel-filterbank transformation can be ignored, which reduces the double summatories over J and K to single ones. This case will be considered in the experimental setup.

The MFCCs are usually complemented with delta and acceleration coefficients. This poses no additional difficulty for the computation of the joint parameters, since these are linear transformations. Furthermore, the Gaussianity of the joint distribution is maintained. This also makes possible the application of other linear transformations like cepstral mean subtraction, or in general infinite or finite response filters like e.g. RASTA.

3.3. JUP based Feature and Model Compensation

After all the parameters of the joint distribution $p(y_{il}, x_{il})$ have been computed, there are various techniques applicable to improve ASR robustness. As described in [11, 5.2.1], the likelihood $p(y_{il}|x_{il})$ of the corrupted features can be easily derived from $p(y_{il}, x_{il})$. This makes possible model-compensation by approximating the distribution for the corrupted speech.

Another possible approach is feature compensation through the computation of a posterior distribution $p(x_{il}|y_{il})$, as in ALGONQUIN [3], STFT-UP [8] or MBFE [4]. According to the definition of MMSE estimator, we have

$$\hat{x}_{il}^{\text{MMSE-MFCC}} = E\{x_{il}|y_{il}\} = \mu_{il}^{x} + \frac{\Sigma_{il}^{yx}}{\Sigma_{il}^{y}}(y_{il} - \mu_{il}^{y}). \quad (11)$$

Furthermore the residual mean square error (MSE) can be computed as

$$\text{MSE} = \text{Var}\{x_{il}|y_{il}\} = \Sigma_{il}^{x} - \frac{\left(\Sigma_{il}^{yx}\right)^2}{\Sigma_{il}^{y}}. \quad (12)$$

This also allows the optional use of model-compensation by using techniques like uncertainty decoding [12]. Modified Imputation (MI) [6] often delivers a better performance when combined with STFT-UP, and it is therefore used here. MI can be described as a model-based feature compensation scheme, equivalent to a soft version of classical imputation. In MI, the features are re-estimated for each mixture of the acoustic model of the ASR system as

$$\hat{x}_{qil}^{\text{MI}} = \frac{\Sigma_{qii}}{\Sigma_{qii} + \text{MSE}} x_{il}^{\text{MMSE}} + \frac{\text{MSE}}{\Sigma_{qii} + \text{MSE}} \mu_{qi}, \quad (13)$$

where μ_{qi} and Σ_{qii} are the mean and variance for each mixture of the ASR model.

3.4. Computational Costs

The costs of the algorithm depend on the speech enhancement method, and the feature or model-compensation technique used. The only fixed costs are the computation of the JUP parameters in (8), (7), (10). These costs are however relatively small compared to ASR costs. In the experimental setup used here, noise is estimated with IMCRA [13] and a MMSE-MFCC with MI compensation is derived from the JUP for feature and model compensation. Such configuration yields a fast, single pass, causal implementation. The only limitation is a small buffer of 8 frames needed for the delta and acceleration computations.

3.5. Comparison with Related Methods

JUP can be compared with other approaches to derive a relation between speech and noise in the MFCC domain. As commented in the introduction, most methods use missmatch functions based on the Taylor series truncation e.g. [2, 4, 3]. Unlike such approaches, in JUP the phase difference between X_{kl} and D_{kl} is implicitly considered in the marginalization of the phase of the posterior (see [5]), and there is no linearization involved. The only assumption therefore needed is the log-normality of the Mel-filterbank features for a complex Gaussian model, which seems to hold quite well in practice. Other approaches using the log-normal assumption such as PMC [21] could be also used to derive the joint distribution. PMC assumes however that the sum of log-normal distributions is log-normal, which does not always hold in practice. The model implicit in PMC is also different from the one presented here as commented in Sec. 3.1. Finally, compared to STFT-UP [5, 8], STFT-JUP provides not only a posterior for the clean features but the full joint distribution relating corrupted and clean speech. This allows the use of additional compensation techniques discussed in this section. STFT-JUP has however a slightly higher computational cost than STFT-UP, since it needs to propagate the two marginals for corrupted and clean speech and compute the covariance. It also needs of more restrictive assumptions like joint-log-normality. Finally it should also be noted the the JUP approach has been also successfully applied for the purpose of speech enhancement. In this case, a closed-form solution was found for the cepstrum non-linearity [22].

Table 1. Word Error Rates (WER) [%] for no compensation baseline (top), feature compensation (middle) and model-based feature compensation (MI, bottom). Best results per block are displayed in bold.

	A	B1	B2	Av.
Baseline	9.1	41.7	57.2	48.1
MMSE-STSA	**9.4**	24.0	47.5	38.7
MMSE-LSA	9.7	26.0	45.4	37.5
MMSE-MFCC (UP)[8]	9.5	21.6	44.5	36.2
ETSI-AFE	9.5	22.6	**28.9**	**25.2**
MMSE-MFCC (JUP, F)	9.6	23.1	46.4	37.8
MMSE-MFCC (JUP, D)	**9.4**	**19.6**	47.1	37.8
MMSE-MFCC+MI (UP, F)[8]	9.5	15.5	37.5	30.4
MMSE-MFCC+MI (JUP, F)	9.6	15.2	37.6	30.4
MMSE-MFCC+MI (JUP, D)	**9.3**	**14.4**	**36.9**	**29.7**

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

To test the proposed method, the same AURORA4 task [23] scenario as in [8] was used[1]. It uses Vertannen's recipe [24] to train a word internal triphone model using the Hidden Markov

[1] See http://www.astudillo.com/ramon/research/stft-up/ for the latest configuration and available code.

Toolkit (HTK) [25]. AURORA4 is noisy version of the well known Wall-Street Journal medium vocabulary task of 5K words featuring read Journal news. The test was centered in robustness against additive noise for single pass causal methods with no prior knowledge of noise as in e.g. mobile voice search scenarios. Such methods are not as performant as multi-pass, non causal methods as e.g. VTS [26] or MBFE [4], but their computational cost is sensibly lower and their cost does not grow significantly with the size of the acoustic models.

The training was therefore carried out with the clean data set of 7138 sentences with a sampling frequency of 16KHz, incorporating speech enhancement into the feature extraction. The test set used is AURORA4's sennheiser microphone 166x7 sentence set, which includes 166 clean speech sentences and six corrupted versions using different additive noises, car, street, train, babble, restaurant and airport. From this set of noises car noise can be regarded as relatively stationary whereas the rest are non-stationary. Since there is only an instance of a relatively stationary noise, for which speech enhancements is usually much more efficient, this set was singled out. The non-stationary noises were averaged into one single coefficient. Clean, stationary and non-stationary noises were labeled A, B1 and B2 respectively.

The experiments compared the results obtained for a MMSE-MFCC estimator derived from STFT-UP [8], with the equivalent estimator obtained with the JUP approach presented here, corresponding to (11) and (12). These were labeled MMSE-MFCC (UP) and MMSE-MFCC (JUP) respectively. For the MMSE-MFCC (UP) estimator the version with full covariance after Mel-filterbank (F) was selected as it provided the best results. In the case of the MMSE-MFCC (JUP) here introduced, both diagonal (D) and full (F) covariance variants were considered. For completeness, results for amplitude (MMSE-STSA) [17] and log-amplitude (MMSE-LSA) [18] estimators were also included. All these methods shared the same estimator for the variances λ_{kl}^X and λ_{kl}^D based on IMCRA [13] and the decision directed method [17]. No parameter of the speech enhancement or propagation frameworks were adapted to the AURORA4 corpus. In addition to this, the ETSI Advance Front-End (ETSI-AFE) [27] was also included as reference.

Aside from pure feature-compensation methods, a modified version of HTK capable of performing Modified Imputation (13) was used for the MMSE-MFCC estimators.

5. ANALYSIS OF RESULTS

Word error rate (WER) results are displayed in Table 1 in three blocks. Feature compensation experiments, displayed in the middle of the table, show that MMSE-MFCC (JUP) performance falls behind MMSE-MFCC (UP) and even MMSE-LSA estimators, in terms of average performance. It achieves however better stationary noise suppression where a relative WER reduction above 50% is achieved. On the contrary, in the experiments employing modified imputation, displayed at the bottom of the table, MMSE-MFCC (JUP) does outperform MMSE-MFCC (UP) although by a relatively small margin. It should be noted, however, that from a theoretical point of view, both estimators should perform similarly. The main advantage of STFT-JUP is that it provides a joint distribution, which allows using a posterior distribution, as it is done in this experiment, or a likelihood for other model-base compensation schemes. Finally, it should also be noted that both UP and JUP fail to outperform the ETSI-AFE on non-stationary noises. It has to be taken into account that the ETSI-AFE was designed using the AURORA2 database, which also positively affects its performance on the AURORA4.

6. CONCLUSIONS

STFT joint uncertainty propagation (STFT-JUP) has been presented as a closed form solution relating the joint distributions of clean and corrupted speech in short-time Fourier transform (STFT) and Mel-Frequency Cepstral Coefficient (MFCC) domains. The method can be used to determine the parameters of a joint distribution usable by methods like FE-JUD [11] or produce feature posteriors such as ALGONQUIN or STFT-UP. Compared to STFT-UP, STFT-JUP attains comparable and even better performance in some scenarios. Further work will explore the use of the joint distribution with other forms of model-compensation as JUD.

7. REFERENCES

[1] M. J. F. Gales, "Model Based Approaches to Handling Uncertainty," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds., chapter 4, pp. 101–125. Springer, Berlin, Germany, 2011.

[2] P. J. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon university, 1996.

[3] L. Deng, "Front-End, Back-End, and Hybrid Techniques to Noise-Robust Speech Recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds., chapter 4, pp. 67–99. Springer, Berlin, Germany, 2011.

[4] V. Stouten, H. Van hamme, and W. Wambacq, "Model based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Communication.*, vol. 48(11), pp. 1502–1514, 2006.

[5] R. F. Astudillo and D. Kolossa, "Uncertainty propagation," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and

R. Haeb-Umbach, Eds., chapter 3, pp. 35–64. Springer, Berlin, Germany, 2011.

[6] 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.

[7] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A Minimum-Mean-Square-Error Noise Reduction Algorithm on Mel-Frequency Cepstra for Robust Speech Recognition," in *ICASSP 2008*, 2008, pp. 4041–4044.

[8] 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.

[9] S. Srinivasan and D. Wang, "Transforming Binary Uncertainties for Robust Speech Recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2130–2140, September 2007.

[10] Kuldip Paliwal Anthony Stark, "MMSE estimation of log-Filterbank energies for robust speech recognition," *Speech Communication*, vol. 53 (3), pp. 403–416, 2011.

[11] Hank Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambrigde, 2007.

[12] J. Droppo, A. Acero, and Li Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 2002, vol. 1, pp. I–57–I–60 vol.1.

[13] 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.

[14] Emanuël Anco Peter Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[15] F. Nesta and M. Matassoni, "Robust Automatic Speech Recognition through on-line Semi Blind Source Extraction," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, 2011, pp. 18–23.

[16] 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.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[18] 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.

[19] Y. Ephraim and I. Cohen, *Recent Advancements in Speech Enhancement*, pp. 1–22, CRC Press, May 17, 2004.

[20] Chrysostomos L. Nikias and Athina P. Petropulu, *Higher-order spectra analysis, a nonlinear signal processing framework*, Prentice Hall signal processing series, 1993.

[21] M. J. F. Gales, *Model-Based technique for noise robust speech recognition*, Ph.D. thesis, Gonville and Caius College, 1995.

[22] R. F. Astudillo and T. Gerkmann, "On the relation between speech corruption models in the spectral and the cepstral domain," in *ICASSP*, June 2013, pp. 7044–7048.

[23] Guenter Hirsch, *Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task*, Niederrhein University of Applied Sciences, November 2002.

[24] Keith Vertanen, "HTK Wall Street Journal Training Recipe," 2006.

[25] S. Young, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department., 2006.

[26] Yongqiang Wang and M. J. F. Gales, "Speaker and noise factorization for robust speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 7, pp. 2149–2158, 2012.

[27] ETSI, *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)*, January 2007.