ON THE RELATION BETWEEN SPEECH CORRUPTION MODELS IN THE SPECTRAL AND THE CEPSTRAL DOMAIN

*Ramón Fernandez Astudillo**

Spoken Language Systems Laboratory
INESC-ID-Lisboa, Lisboa, Portugal
ramon@astudillo.com

Timo Gerkmann

Speech Signal Processing Group
University of Oldenburg, Germany
timo.gerkmann@uni-oldenburg.de

ABSTRACT

The Gaussian distortion model in the short-time Fourier transform (STFT) domain is the basis of many of the modern speech enhancement algorithms. One of the reasons is that additive sources and late reverberation can be analyzed and processed quite efficiently in this domain. The STFT domain is however not well related to acoustic quality and is also not well suited for learning models due to the high variability of speech in this domain. On the other hand, the cepstral domain has proved to be very well suited for these last two purposes, however, at the cost of loosing the simple linear relation between desired source and additive interferences. In this paper we explore the relation between the Gaussian distortion models in the STFT and the cepstral domain. We show how the assumption of a jointly Gaussian distortion model in the cepstrum domain is fulfilled for well-known distortion models in STFT domain. We provide closed-form solutions relating the joint distributions of corrupted and clean speech in the STFT and the cepstrum domain. We also propose various ways in which this model can be used to enhance speech.

Index Terms— Speech Enhancement, Cepstrum Domain, Uncertainty Propagation

1. INTRODUCTION

The Short-time Fourier transform (STFT) domain provides a simple mean to obtain a time-frequency representation of speech signals with very desirable properties for signal processing purposes. It is a linear invertible transform and the modeling of speech corruption phenomena such as additive noise [1] or late reverberation [2] is easy compared to other domains. The STFT is however not short from drawbacks. The perceived quality of speech is better represented by non-linear features of the STFT [3]. Modeling of speech, e.g. phonetic units, is also very difficult due to its high variability in the STFT domain. This work shows that the conventional statistical distortion model in the STFT domain can be related to an equivalent model in the cepstral domain thus allowing to optimally exploit the properties of both domains simultaneously.

The work here presented is related to various pre-existing approaches that employ corruption models in the STFT domain while performing estimates or learning of models in non-linear domains. Approaches like VTS [4] or ALGONQUIN [5], for example, use Taylor series to approximate the effect of additive and convolutive STFT domain distortions in the MFCC domain for robust automatic speech recognition (ASR). Also applied to ASR, short-time Fourier transform uncertainty propagation (STFT-UP) [6] approximates the transformation of the statistical model resulting from the additive noise assumption in the STFT domain, attaining estimates in MFCC RASTA-LPCC or MLP domains. In the speech enhancement field, non-linear minimum mean square error (MMSE) estimators of speech that employ STFT domain distortion models are very extended. These include the well known Ephraim-Malah filters, which provide amplitude (MMSE-STSA) [1] and log-amplitude (MMSE-LSA) [3] domain estimators and MMSE estimators for other domains like MFCC [7, 8]. Ephraim and Rahim also derived a linear MMSE estimator in the cepstral domain which is directly related to the approach presented here [9]. Other techniques that combines spectral and cepstral processing for speech enhancement is that of [10], where it is shown that the estimation of the speech power spectral density (PSD) is more robust when selective cepstrum smoothing techniques are employed.

In this work we study the relation between the joint distributions of corrupted speech and noise in the STFT and cepstrum domains. We build on the work by Ephraim [9] which initially derived the means and variances of cepstral coefficients for complex Gaussian STFT models. We use the approach in [11] to derive formulas that consider the effect of tapered spectral analysis. We also show that both posterior and likelihood distributions are accurately described by Gaussian distributions and propose some possible ways of exploiting this fact for speech enhancement purposes.

2. THE GAUSSIAN MODEL OF SPEECH DISTORTION IN STFT DOMAIN

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 k and l denote frequency and analysis frame indices. In this work we employ the Gaussian model for speech in the STFT domain. This model assumes that each Fourier coefficient of the observable noisy 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 statistically independent of X_{kl}. The model also assumes following a priori circular symmetric complex Gaussian distributions

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

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

In order to determine this model, the variances of the hidden clean and corrupted speech Fourier coefficients, i.e. the PSDs $\lambda_{kl}^{X} =$

*Work supported by the Portuguese Foundation for Science and Technology, grant SFRH/BPD/68428/2010 and project PEst-OE/EEI/LA0021/2011.

$\mathrm{E}\{|X_{kl}|^2\}$ and $\lambda_{kl}^D = \mathrm{E}\{|D_{kl}|^2\}$ have to be estimated. There are multiple methods available depending on the type of distortion modeled. Within the scope of this work the additive noise case will be considered, the results can be however extended to late reverberance suppression [2] or other approaches exploiting the same model.

Once the a priori parameters have been determined MMSE estimators like the Wiener (e.g. [12]) or Ephraim-Malah filters [1, 3] can be employed to estimate the clean speech. The objective of this work is however to propagate the statistical relation implied by equations (1), (2), (3) to cepstrum domain prior to performing any estimation. In what follows we will then consider the joint distribution of each corrupted and clean Fourier coefficient, which is given by

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

3. DERIVING A MODEL OF SPEECH DISTORTION IN CEPSTRAL DOMAIN

3.1. Joint Uncertainty Propagation into Cepstrum Domain

Let the clean cepstrum be defined by

$$\mathbf{x}_l = \mathrm{IFFT}\left(\log\left(|\mathbf{X}_l|^2\right)\right), \tag{4}$$

where $\mathbf{X}_l$ is the the l^{th} frame of the clean STFT and magnitude, square and logarithm operations act element-wise. Let the cepstrum of the corrupted speech $\mathbf{y}_l$ be computed in analogous form from $\mathbf{Y}_l$. Since $\mathbf{X}_l$ and $\mathbf{Y}_l$ are correlated also the joint distribution of $\mathbf{x}_l$ and $\mathbf{y}_l$ will include a statistical dependence. The main premise of the approach presented here is that this distribution can be modeled by the jointly Gaussian distribution

$$\begin{bmatrix} y_{il} \\ x_{il} \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_{il}^y \\ \mu_{il}^x \end{bmatrix}, \begin{bmatrix} \Sigma_{il}^y & \Sigma_{il}^{yx} \\ \Sigma_{il}^{yx} & \Sigma_{il}^x \end{bmatrix} \right),$$

where we here consider the cepstral coefficients x_{il} and y_{il} to be independent of their adjacent coefficients. This is justified by the decorrelation properties of the IFFT in (4). The method proposed here relates the joint distributions of corrupted and clean speech in STFT and cepstrum domains, similarly to how uncertainty propagation (UP) [6] relates the posterior distributions of clean speech in both domains. For this reason the method is here termed joint uncertainty propagation (JUP).

3.2. JUP based Estimators and Inference

Once the parameters of the JUP have been determined, it is possible to perform various types of estimations. The most straightforward way is to derive the posterior distribution of the clean cepstrum given the observed cepstrum $p(x_{il}|y_{il})$. The mean of the posterior distribution is in fact the MMSE estimator of the cepstrum given the available information, thus

$$\hat{x}^{\mathrm{MMSE}} = \mu_{il}^{x|y} = \mu_{il}^x + \frac{\Sigma_{il}^{yx}}{\Sigma_{il}^y}(y_{il} - \mu_{il}^y). \tag{5}$$

An equivalent estimator to this was also derived in [9, eq. (32)], but in difference, our proposal includes the effect of the tapered spectral analysis windows. Further, in contrast to the result presented here, the equivalent solution in [9] was found as the linearly-constrained MMSE estimator without defining the joint distribution or posterior. Under the assumption of joint Gaussianity the associated posterior distribution can be propagated back into the amplitude and MFCC domains using the same formulas as for the RASTA filtered uncertain features in [13] and the residual mean square error (MSE)

$$\mathrm{MSE} = \Sigma_{il}^{x|y} = \Sigma_{il}^x - \frac{(\Sigma_{il}^{yx})^2}{\Sigma_{il}^y}. \tag{6}$$

as the source of uncertainty. Interestingly, the assumption of a jointly Gaussian distribution in the cepstral domain provides also the possibility of computing the likelihood $p(y_{il}|x_{il})$ with mean

$$\mu_{il}^{y|x} = \mu_{il}^y + \frac{\Sigma_{il}^{xy}}{\Sigma_{il}^x}(\mathbf{x}_{il} - \mu_{il}^x) \tag{7}$$

and variance

$$\Sigma_{il}^{y|x} = \Sigma_{il}^y - \frac{(\Sigma_{il}^{yx})^2}{\Sigma_{il}^x}. \tag{8}$$

This likelihood can be then combined with other models with richer a priori information obtained from pre-training, or the prediction step of recursive Bayesian estimators. This allows to exploit the properties of the cepstral domain, more appropriate for modeling of speech, while using conventional STFT domain distortion models. Let

$$x_{il}|m \sim \mathcal{N}\left(\mu_{il}^m, \lambda_{il}^m\right) \tag{9}$$

be a Gaussian prior obtained from this richer a priori information m. A new MMSE estimator and corresponding residual MSE can be attained from $p(y_{il}|x_{il})$ and $p(x_{il}|m)$ and the Bayes theorem as

$$p(x_{il}|y_{il}, m) = \frac{p(y_{il}|x_{il})p(x_{il}|m)}{\int_{\mathbb{R}} p(y_{il}|x_{il})p(x_{il}|m)dx_{il}}. \tag{10}$$

Since both prior and likelihood are Gaussian distribution this yields another Gaussian posterior with mean

$$\mu_{x|y,m} = \frac{\lambda_{il}^m}{a^2\lambda_{il}^m + \Sigma_{il}^{y|x}}u + \frac{\Sigma_{il}^{y|x}}{a^2\lambda_{il}^m + \Sigma_{il}^{y|x}}\mu_{il}^{y|x} \tag{11}$$

and variance

$$\Sigma_{x|y,m} = \frac{\lambda_{il}^m \Sigma_{il}^{y|x}}{a^2\lambda_{il}^m + \Sigma_{il}^{y|x}} \tag{12}$$

where

$$a = \frac{\Sigma_{il}^{yx}}{\Sigma_{il}^x}, \tag{13}$$

and

$$u = a \cdot y_{il} - a \cdot \mu_{il}^y + a^2 \cdot \mu_{il}^x. \tag{14}$$

Finally, the likelihood can be also used for robust inference in the cepstrum domain by using Joint Uncertainty Decoding (JUD) [14], similarly to how UP [6] is used with uncertainty decoding (UD) [15].

4. DERIVING THE CEPSTRAL MEAN, VARIANCE AND COVARIANCE

In order to compute the parameters of the JUP, for instance the means of the posterior and likelihood (5), (7), we need to model the means μ_{il}^x, μ_{il}^y, the variances $\Sigma_{il}^x, \Sigma_{il}^y$, as well as the covariance Σ_{il}^{yx}.

For complex Gaussian spectral coefficients where neighboring frequency bins are uncorrelated, the resulting means and variances of

cepstral coefficients are derived in [9]. In practice, tapered spectral analysis windows will be used when computing the STFT. The multiplication with this analysis window in time domain corresponds to a convolution in frequency domain. This convolution necessarily results in a correlation of neighboring frequency coefficients and results in cepstral variances that are not flat, but decaying from low to large cepstral coefficients. The general effect of a correlation of neighboring complex Gaussian distributed frequency coefficients on the cepstral variance is given in [16]. In [11] the results for the mean and variance of cepstral coefficients are generalized when the magnitude-square of complex spectral coefficients are Gamma (χ^2) distributed. This parameterizable distribution comprises the results for complex-Gaussian distributions but can also be used to model super-Gaussian distributed complex spectral coefficients and smoothed periodograms. Further, in [11] compact solutions are given for the effect of tapered spectral analysis windows on the cepstral variance. In this section we also derive the cross-covariance between the clean and corrupted speech in the cepstral domain, when tapered spectral analysis windows are employed that result in a correlation of neighboring frequency coefficients.

For a real-valued time domain signal, for large frame sizes the complex cepstral coefficients are asymptotically Gaussian distributed. In particular, the DC and Nyquist bin are real-valued and Gaussian distributed, while the remaining coefficients are complex with Gaussian distributed real and imaginary parts. As a result, we obtain for the mean of the cepstrum [11]

$$\begin{aligned}\mu_{il}^{x} &= \mathrm{IDFT}\left\{\mathrm{E}\left\{\log\left(|\mathbf{X}_l|^2\right)\right\}\right\} \\ &= \mathrm{IDFT}\left\{\log\left(\mathrm{E}\left\{|\mathbf{X}_l|^2\right\}\right)\right\} - \epsilon, \end{aligned} \tag{15}$$

with

$$\epsilon = \begin{cases} \boldsymbol{C} + \frac{2}{K}\log(2) & , i = 0 \\ \frac{2}{K}\log(2) & , i \text{ even} \\ 0 & , i \text{ odd}, \end{cases} \tag{16}$$

where we employed [17, Sec. 8.366] and $\boldsymbol{C} = 0.5772...$ is the Euler constant [17, Sec. 9.73]. The same results hold for μ_{il}^{y} correspondingly. For the large K usually employed in speech processing, the term $\frac{2}{K}\log(2)$ can be neglected.

For Gaussian distributed and spectrally uncorrelated coefficients the variance of cepstral coefficients is derived to be $\pi^2/(6K)$ and twice as high at the zeroth and $K/2$th cepstral coefficient [9]. However, the spectral correlation caused by a tapered spectral analysis window results in the cepstral variance [11]

$$\Sigma_{il}^{x} = \begin{cases} \frac{2}{K}\left(\frac{\pi^2}{6} + 2\sum_{m=1}^{M}\kappa_m \cos\left(m\frac{2\pi}{K}i\right)\right) & , i \in \left\{0, \frac{K}{2}\right\} \\ \frac{1}{K}\left(\frac{\pi^2}{6} + 2\sum_{m=1}^{M}\kappa_m \cos\left(m\frac{2\pi}{K}i\right)\right) & , \text{else}. \end{cases} \tag{17}$$

For the Hann spectral analysis window we have $M = 2$ and $\kappa_1 = 0.507$ and $\kappa_2 = 0.028$ [11]. The same results hold for Σ_{il}^{y} correspondingly.

To compute the covariance of cepstral coefficients, we will first consider the correlation in the spectrum domain. For neighboring frequency coefficients, the correlation coefficient is defined as

$$\rho_{XX}^2(m) = \frac{|\mathrm{E}\{X_k X_{k+m}^*\}|^2}{\mathrm{E}\{|X_k|^2\}\,\mathrm{E}\{|X_{k+m}|^2\}}. \tag{18}$$

Assuming speech and noise are uncorrelated, the correlation between the clean and corrupted speech can be computed as

$$\begin{aligned}\rho_{YX}^2(m) &= \frac{|\mathrm{E}\{X_k Y_{k+m}^*\}|^2}{\mathrm{E}\{|X_k|^2\}\,\mathrm{E}\{|Y_{k+m}|^2\}} = \frac{|\mathrm{E}\{X_k X_{k+m}^*\}|^2}{\mathrm{E}\{|X_k|^2\}\,\mathrm{E}\{|Y_{k+m}|^2\}} \\ &= \rho_{XX}^2(m)\frac{\mathrm{E}\left\{|X_{k+m}|^2\right\}}{\mathrm{E}\{|X_{k+m}|^2\} + \mathrm{E}\{|D_{k+m}|^2\}}. \end{aligned} \tag{19}$$

Thus, with the correlation introduced by a Hann window $\rho_{XX}(0) = 1, \rho_{XX}(1) = 2/3$, $\rho_{XX}(2) = 1/6$, we can determine correlation between clean and corrupted speech from the speech and noise PSDs $\lambda_{kl}^{X} = \mathrm{E}\left\{|X_{kl}|^2\right\}$ and $\lambda_{kl}^{D} = \mathrm{E}\left\{|D_{kl}|^2\right\}$. Similar to [9, 16], for complex Gaussian distributed spectral coefficients, we can obtain the covariance in the log-domain as

$$\mathrm{cov}(\log|X_k|^2, \log|Y_{k+m}|^2) = \sum_{n=1}^{\infty}\frac{1}{n^2}\rho_{YX}^{2n}(m). \tag{20}$$

A more general solution is obtained by modeling the periodograms $|X_k|^2, |Y_k|^2$ by the parameterizable Gamma (χ^2) distribution. The results for this generalized model are given in [11, Eq. (16)]. The cepstral covariance Σ_{il}^{yx} is finally obtained by taking a 2D inverse Fourier transform of (20).

5. EXPERIMENTS AND RESULTS

5.1. Monte Carlo Simulation Tests

In order to asses the accuracy of the proposed propagation algorithm a Monte Carlo simulation was used. Two stationary white signals of 1e6 samples where generated in the time domain simulating decorrelated clean speech and noise processes. Both the clean signal and the corrupted signal, the addition of both signals, were firstly transformed into the STFT domain using Hann spectral analysis windows where they were scaled with random variances to meet a given segmental signal to noise ratio [18]. The SNR ranged from -10dB to 40dB. These were later transformed into the cepstrum domain where the statistics were computed and compared with the JUP analytic solution derived in Section 4.

Figure 1 compares the Monte Carlo and JUP estimated parameters of the joint distribution of clean and corrupted speech in the cepstrum domain. The quefrency shown corresponds to the average case, although there is not much difference between different quefrencies. As it can be seen, the JUP estimates are very accurate for the whole range of SNR explored and the small variations observed are due to the intrinsic variability of Monte Carlo estimates.

Figure 2 compares the Monte Carlo and JUP estimated posterior and likelihood distributions for two selected SNRs of around 0dB and 20dB respectively. The JUP estimated distributions are assumed to be Gaussian with parameters given by (5) (6) (7) (8) while the Monte Carlo estimate is an histogram attained by selecting samples of a very thin interval of either x_{il} or y_{il}. As it can be seen the estimated JUP posteriors and likelihoods accurately match the empirical data and the Gaussian assumption under which this model was derived. A slight overestimation of the variance can be observed at high SNR, this is nevertheless small compared with the range of variation of the cepstra.

5.2. Speech Enhancement Tests

To provide some practical application of the presented technique, speech enhancement tests were carried out using the AURORA4

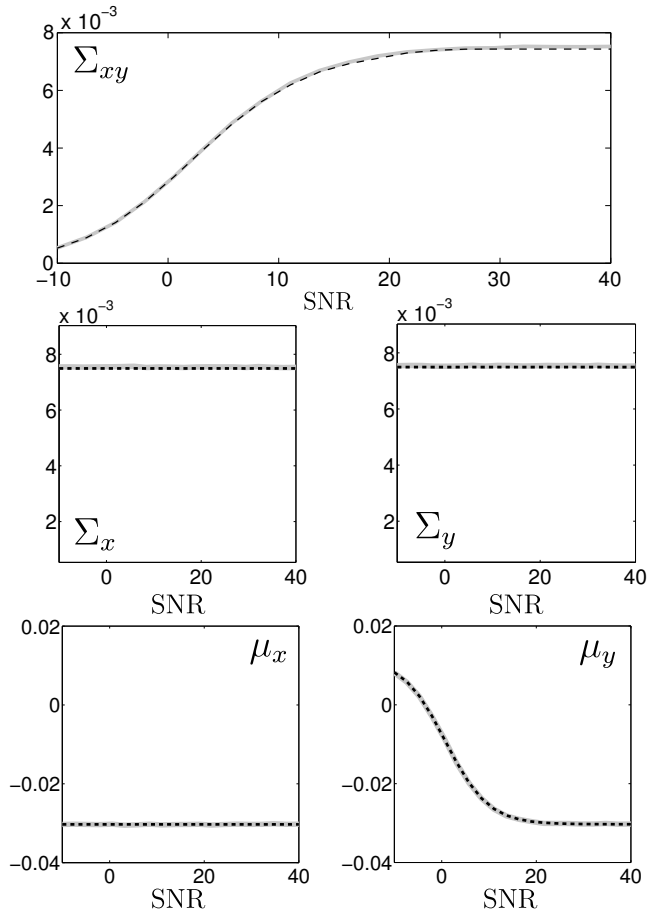

Fig. 1. Estimated parameters of the joint distribution of corrupted and clean speech in the cepstrum domain against SNR. Montecarlo (solid grey) versus JUP (dashed black).

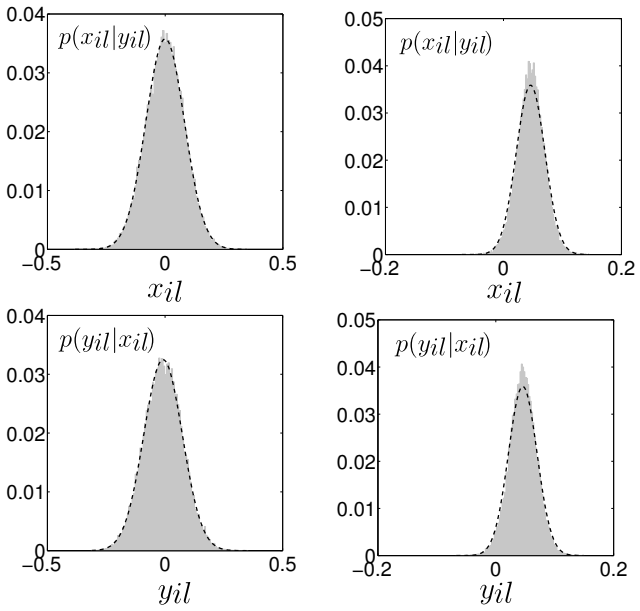

Fig. 2. Posterior and likelihood distributions of corrupted and clean speech in the cepstrum domain. 0dB SNR (left) compared to 20dB SNR (right). Montecarlo (solid grey) versus JUP (dashed black) estimates.

large vocabulary corpus [19]. This corpus provides speech artificially corrupted with a variety of noises. The 166 sentence car noise test set was selected for this experiment. The estimation of the a priori parameters λ_{kl}^X and λ_{kl}^D was carried out using [20] and [10] with the bias compensation from [11]. Two measures of acoustic quality were used, perceptual evaluation of speech quality (PESQ) [21] and MSE in the cepstral domain.

The test compared three conventional estimators derived from the Gaussian model of speech distortion, Wiener, MMSE-STSA [1] and MMSE-LSA [3] against two estimators derived from JUP. The first estimator was the JUP-MMSE-ceps estimator in (5), obtained from the posterior. The second was a proof-of-concept regarding the use of additional a prior information in the cepstrum domain. For this purpose an oracle prior, a Dirac delta centered on the clean cepstrum, was used. To artificially vary the amount of a priori information a distortion was added to the mean of the oracle prior at a given SNR. The variance of the prior was also modified accordingly to reflect the lack of information. This prior was then used together with the JUP likelihood defined by (7) and (8) to attain the estimate given by (11). This was termed JUP-MMSE-ceps-OP.

As shown in Table 1 the results for the three estimators are very similar in terms of PESQ performance when compared to the standard deviation. The performance of the JUP-MMSE-ceps remains between the Wiener and the MMSE-STSA and MMSE-LSA. In the case of cepstral MSE, however, the JUP based estimators outperform all other estimators. For the estimator with oracle prior JUP-MMSE-ceps-OP the behavior is as expected. For poor a priori information the performance is below that of the conventional estimators. As the a priori information increases the performance of the JUP-MMSE-ceps-OP estimator increases and outperforms all other estimators. Although not included here, results show that the gap between MMSE-STSA, MMSE-LSA and the JUP-MMSE-ceps is further reduced for ideal estimations of λ_{kl}^X and λ_{kl}^D. This indicates that JUP based estimators are more sensitive to errors in the a priori estimation of parameters, an aspect that will have to be studied in further works.

Table 1. PESQ and SNR Values

	PESQ		MSE	
	mean	std	mean	std
Noisy	2.62	0.24	5.87	2.29
MMSE-LSA	3.09	0.20	5.43	0.54
MMSE-STSA	3.08	0.20	4.73	0.45
Wiener	3.00	0.20	23.94	2.47
JUP-MMSE-ceps	3.03	0.22	3.77	1.33
JUP-MMSE-ceps (OP 0dB)	2.93	0.29	4.85	1.57
JUP-MMSE-ceps (OP 3dB)	3.12	0.25	3.97	1.26
JUP-MMSE-ceps (OP 5dB)	3.23	0.24	2.68	0.87

6. CONCLUSIONS

In this work we have shown that the joint distribution of clean and corrupted cepstrum can be accurately modeled by a jointly Gaussian distribution given the parameters of the conventional Gaussian model of speech distortion in STFT domain. We have derived closed form solutions to compute the parameters of this joint distribution when using tapered spectral analysis window and discussed different applications for enhancement purposes.

7. REFERENCES

[1] 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.

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

[3] 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.

[4] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, may 1996, vol. 2, pp. 733 –736 vol. 2.

[5] B. Frey, L. Deng, A. Acero, T. T., and Kristjansson, "Iterating laplaces method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[6] Ramon Fernandez Astudillo, *Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition*, Ph.D. thesis, Technische Universität Berlin, 2010.

[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 *Proc. ICASSP*, 2008, pp. 4041–4044.

[8] R. F. Astudillo and R. Orglmeister, "A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation," in *Proc. Interspeech*, 2010.

[9] Yariv Ephraim and Mazin Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 162–176, Mar. 1999.

[10] Colin Breithaupt, Timo Gerkmann, and Rainer Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 4897–4900.

[11] Timo Gerkmann and Rainer Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[12] 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.

[13] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Uncertainty propagation for speech recognition using rasta features in highly nonstationary noisy environments," in *ITG Workshop for Speech Communication*, 2008.

[14] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech*, 2005, pp. 3129–3132.

[15] J. Droppo, A. Acero, and Li Deng, "Uncertainty decoding with splice for noise robust speech recognition," in *Proc. ICASSP*, 2002, vol. 1, pp. I–57–I–60 vol.1.

[16] Yariv Ephraim and William J. J. Roberts, "On second-order statistics of log-periodogram with correlated components," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 625–628, Sept. 2005.

[17] I. S. Gradshteyn and I.M Ryzhik, *Table of Integrals, Series and Products*, Elsevier, 2007.

[18] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[19] 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.

[20] Timo Gerkmann and Richard C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[21] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.