MMSE Estimation of Speech Power Spectral Density Under Speech Presence Uncertainty for Automatic Speech Recognition

Jingang Liu, Yi Zhou, Yongbao Ma, Hongqing Liu
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
{jg_liu, yb_ma} @outlook.com, {zhouy, hongqingliu}@cqupt.edu.cn

***Abstract*—In order to improve the performance of automatic speech recognition (ASR) system in noisy environment, this paper first derives a minimum mean square error (MMSE) speech power spectral density (PSD) estimator. Besides, speech presence uncertainty (SPU) is explored to obtain further improvement. The speech spectral amplitude model assumed in this paper is Chi distribution instead of the traditional Rayleih distribution. Simulation experiments of ASR system demonstrate that the new proposed approach outperforms the traditional estimation of speech spectral amplitude and PSD with SPU under Rayleih distribution.**

Keywords—ASR; MMSE estimation ; speech PSD ;SPU;

I. Introduction

Due to noise corruption, the automatic speech recognition (ASR) system may suffer a performance loss. To solve this problem, a lot of speech enhancement algorithms have been developed to increase the robustness of the ASR system [1]. Over the past years, speech enhancement algorithms have been studied intensively based on statistical models in frequency domain. The minimum mean square error (MMSE) estimator is widely used in estimating the clean speech DFT coefficients or amplitudes. The well-known MMSE log spectral amplitude (MMSE-LSA) estimator was designed by Ephraim and Malah [2]. Then, Cohen proposed an optimal modified MMSE-LSA algorithm (OM-LSA) [3] which combines speech presence uncertainty (SPU) and MMSE-LSA so that improved robustness for ASR in noise environment is achieved [4]. Since the OM-LSA algorithm enhances the speech signal but not the extracted features to be used for ASR, it is sub-optimal [4]. [5], [6] proposed an optimal estimate of the mel-frequency cepstral coefficients (MFCC), which, however, could be only applied in the ASR system where the MFCC feature is utilized.

From [7] it can be concluded that compared to speech spectral amplitude estimation method, speech power spectral density (PSD) estimation may help ASR system achieve higher recognition accuracy when the speech PSD rather than its spectral amplitude is needed. So instead of estimating speech spectral amplitude or a specific optimal feature like MFCC, the speech PSD can be estimated for robust ASR system. In this paper, we work with the speech PSD estimator. Furthermore, taking the SPU into consideration, we derive a novel estimator incorporating the SPU with Chi *priori*. Both the speech PSD and the SPU estimator are based on the Chi spectral amplitude model, because [8] and [9] have proved speech spectral amplitude coefficients are better modeled by Chi distribution than Rayleigh distribution.

This paper is organized as follows: the basic assumptions and notations are introduced in Section II. In section III, the MMSE estimation of speech PSD with SPU under Chi *priori* for ASR system is derived. The proposed and other relevant algorithms are compared using simulations and tested for ASR system in Section IV. Finally, conclusions are drawn in Section V.

II. Basic Assumptions and Notations

Suppose the received noisy signal $y(n) = x(n) + d(n)$, where $x(n)$ and $d(n)$ denote clean speech and additive noise, respectively. Appling short-time Fourier transform (STFT) over consecutive frames to the above equation yields $Y(k) = X(k) + D(k)$ where $Y(k)$, $X(k)$ and $D(k)$ represent the k-th spectral components. They are assumed to be complex zero-mean random variables which are statistically independent across time and frequency. The polar-form expressions $Y(k) = Re^{j\phi}$, $X(k) = Ae^{j\theta}$ are also used where applicable. R and A respectively denote the spectral amplitudes of $Y(k)$ and $X(k)$, ϕ and θ are the corresponding phases. Furthermore, the *a priori* signal-to-noise ratio (*SNR*) and the *a posteriori SNR* are respectively defined as $\xi_k = \lambda_x / \lambda_d$ and $\gamma_k = R^2 / \lambda_d$, λ_d and λ_x respectively denote the variances of $D(k)$ and $X(k)$. It is assumed that the spectral amplitude observation obeys a Chi distribution given below:

$$p(A) = \frac{2\beta^a}{\Gamma(a)} A^{2a-1} \exp(-\beta A^2) \tag{1}$$

where $2a$ represents the degrees of freedom and $\beta = a / \lambda_x$ [10]. The Chi distribution can simplify to the Rayleigh distribution when $a = 1$. It is shown in [11] that speech amplitude coefficients obey Chi distribution more closely when $a < 1$.

Another assumption is that the DFT coefficients of the additive noise have a complex Gaussian distribution. Besides, speech and noise DFT coefficients are independent, which implies

$$p(Y(k) \mid A, \theta) = \frac{1}{\pi\lambda_d} \exp(-\mid Y(k) - A\exp(j\theta) \mid^2 / \lambda_d),$$
$$= \frac{1}{\pi\lambda_d} \exp(-\mid R^2 + A^2 - 2A\,\mathrm{Re}\{Y(k)\exp(-j\theta)\} \mid / \lambda_d) \text{ .(2)}$$

III. MMSE PSD Estimation Under Speech Presence Uncertainty

A. MMSE Speech PSD Estimation

Let $e_k = A^2$ denote the PSD of the clean speech signal, the estimation of e_k can be derived as follows.

The posterior probability distribution function (PDF) of clean speech spectral amplitude $p(A \mid Y(k))$ can be obtained from [2]:

$$p(A \mid Y(k)) = \frac{p(Y(k) \mid A)p(A)}{\int_0^{\infty} p(Y(k) \mid A)p(A)d\mathrm{A}} \tag{3}$$

where $p(Y(k) \mid A)$ can be computed as

$$p(Y(k) \mid A) = \int_0^{2\pi} p(Y(k) \mid A, \theta)p(\theta)d\theta \, . \tag{4}$$

θ is the phase of $Y(k)$ which obeys a uniform distribution over $[0, 2\pi]$. Substituting (2) into (4) yields

$$p(Y(k) \mid A) = \frac{1}{\pi\lambda_d} \mathrm{e}^{-\frac{R^2+A^2}{\lambda d}} \int_{-\pi}^{\pi} \exp(\frac{2A\,\mathrm{Re}(e^{-j\theta}Y(k))}{\lambda_d})\mathrm{d}\theta \, . \tag{5}$$

Using (23) in Appendix to solve the integral in (5) yields

$$p(Y(k) \mid A) = \tfrac{1}{\pi\lambda_d}\exp\left(-\tfrac{R^2+A^2}{\lambda_d}\right)\mathrm{I}_0(2A\sqrt{v_k / \lambda_k}) \tag{6}$$

where $v_k = \dfrac{\xi_k}{\xi_k+1}\gamma_k$, $\lambda_k = \dfrac{\lambda_x}{\xi_k+1}$. $\mathrm{I}_0(\cdot)$ is the modified first kind of Bessel function with zeroth order. Next, Substituting (6) and (1) into (3), $p(A \mid Y(k))$ can be measured by:

$$p(A \mid Y(k)) = \frac{A^{2a-1}\exp\{-(\beta+\frac{1}{\lambda_d})A^2\}\mathrm{I}_0(2A\sqrt{v_k / \lambda_k})}{\int_0^{\infty} A^{2a-1}\exp\{-(\beta+\frac{1}{\lambda_d})A^2\}\mathrm{I}_0(2A\sqrt{v_k / \lambda_k})\mathrm{d}A} \, . \tag{7}$$

Using eqn. (24) in Appendix to solve the integral in (7):

$$p(A \mid Y(k)) = \frac{A^{2a-1}\exp\{-(\beta+\frac{1}{\lambda_d})A^2\}\mathrm{I}_0(2A\sqrt{v_k / \lambda_k})}{\Gamma(a)\Phi(\mathrm{a},1;\frac{v_k}{\lambda_k(\beta+1/\lambda_d)})/2(\beta+1/\lambda_d)^a} \tag{8}$$

where $\Gamma(\cdot)$ is the gamma function and $\Phi(\cdot)$ is the confluent hypergeometric function.

The PDF of clean speech PSD $p(e_k \mid Y(k))$ is shown in [11] to be:

$$p(e_k \mid Y(k)) = p(A \mid Y(k))\frac{\mathrm{d}A}{\mathrm{d}e_k} = \frac{p(A \mid Y(k))}{2\sqrt{e_k}} \, . \tag{9}$$

Substituting (8) into (9) and using $A = \sqrt{e_k}$ yield the conditioned PDF of speech PSD:

$$p(e_k \mid Y(k)) = \frac{{e_k}^{a-1}\exp(-(\beta+\frac{1}{\lambda_d})e_k)\mathrm{I}_0(2e_k\sqrt{e_k v_k / \lambda_k})}{\Gamma(a)/(\beta+\frac{1}{\lambda_d})^a\Phi(\mathrm{a},1;\frac{v_k}{\lambda_k(\beta+1/\lambda_d)})} \, . \tag{10}$$

According to the MMSE criterion, the estimation of clean speech power spectrum can be calculated as:

$$\hat{e}_k = E[e_k \mid Y(k)] = \int_0^{\infty} e_k p(e_k \mid Y(k))\mathrm{d}e_k \, . \tag{11}$$

Finally, substituting (10) and $\beta = a / \lambda_x$ into (11) and computing the integral using eqns. (25),(26) in Appendix, the estimation of clean speech power spectrum can be represented in a closed form:

$$\hat{e}_k = \frac{z\lambda_k}{v_k}\exp(z)\frac{\Gamma(a+1)\Phi(-a,1;-z)}{\Gamma(a)\Phi(a,1;z)} \tag{12}$$

where $z = \dfrac{v_k}{\lambda_k(\beta+1/\lambda_d)} = \dfrac{v_k(a+\xi_k)}{(1+\xi_k)}$.

B. *Speech Presence Uncertainty*

To improve the performance of MMSE speech spectrum power estimation, we can combine $\hat{e}_k$ with the SPU framework [2] as follows:

$$\hat{e}_k^{\mathrm{spu}} = e_k \,|_{\xi_k=\xi_k'} \; p(H_1^k \mid Y(k)) \tag{13}$$

where H_1^k represents speech absence in the *k*-th frequency bin and $p(H_1^k \mid Y(k))$ denotes the *a posteriori* probability of speech presence, it can be computed as

$$p(H_1^k \mid Y(k)) = \frac{\Lambda_k}{1+\Lambda_k} \tag{14}$$

where Λ is defined as

$$\Lambda_k = \frac{1-q_k}{q_k}\frac{p(Y(k) \mid H_1^k)}{p(Y(k) \mid H_0^k)} \tag{15}$$

with $p(Y(k) \mid H_1^k)$ and $p(Y(k) \mid H_0^k)$ being the PDFs of Y under assumption of the speech presence and absence, respectively. q_k represents the probability of speech presence. In speech absence, the noisy signal DFT coefficients have a complex Gaussian distribution, $p(Y(k) \mid H_0^k)$ turns out to be

$$p(Y(k) \mid H_0^k) = p(Y(k) = D(k)) = \tfrac{1}{\pi\lambda_d}\exp(-R^2 / \lambda_d) \, . \tag{16}$$

While in speech absence case,

$$\begin{aligned} p(H_1^k \mid Y(k)) &= p(Y(k) = X(k) + D(k)) \\ &= p(Y(k) = X(k)) * p(Y(k) = D(k)) \end{aligned} \tag{17}$$

where $*$ denotes the convolution operation. According to (1) and the Appendix in [12], the PDF of the complex-valued variable $Y(k)$, $p(Y(k) = X(k))$ can be given by

$$p(Y(k) = X(k)) = \frac{1}{2\pi}\frac{2\beta^a}{\Gamma(a)}R^{2a-2}\exp(-\beta R^2) \, . \tag{18}$$

Next, substituting $\beta = a / \lambda_x$, (16) and (18) into (17) and using the method in [12] to solve the complex convolution, one can generate

$$\Lambda_k = \frac{1-q}{q}(\frac{1}{1+\xi_k' / a})\Phi(a,1,\frac{\gamma_k}{1+a/\xi_k'}) \, . \tag{19}$$

If the parameter $a = 1$, (19) can be simplified to $\Lambda_k = \dfrac{1-q}{q}\dfrac{\exp(v_k)}{1+\xi_k'}$ under Gaussian prior in [2].

Finally, substituting (12) and (14) into (13), the MMSE speech spectrum power estimation with SPU can be shown by:

$$\hat{e}_k^{spu} = \frac{\mathrm{z}\lambda_k}{v_k}\exp(z)\frac{\Gamma(a+1)\Phi(\text{-}a,1;-z)}{\Gamma(a)\Phi(a,1;z)}\frac{\Lambda_k}{1+\Lambda_k} \tag{20}$$

where $z=\dfrac{v_k(a+\xi'_k)}{(1+\xi'_k)}$, $\xi'_k=\dfrac{\xi_k}{1-q}$, and Λ_k can be calculated by (19).

IV. Simulation Results

In this section, the performance of the proposed algorithm is evaluated in terms of speech frequency weighted segment SNR (SSNR) [13], Perceptual Evaluation of Speech Quality (PESQ) in ITU-TP.862 [14] and the speech recognition accuracy of pocket sphinx system [15]. The frequency weighted SSNR is as follows

$$\text{fSNR}=\frac{10}{M}\sum_{m=0}^{M-1}\left(\sum_{j=1}^{K}B_j\log_{10}\left[\frac{F^2(m,j)}{(F(m,j)-\hat{F}(m,j))^2}\right]\Big/\sum_{j=1}^{K}B_j\right) \qquad (21)$$

where M is the total number of frames, B_j represents the weight of the j-th spectrum band and K denotes the number of bins in each band. $F(m,j)$ and $\hat{F}(m,j)$ respectively denote the filtered band's amplitude of the j-th spectrum band in the m-th frame of the clean speech signal and the enhanced signal. The main advantage in using the frequency weighted SSNR over the time-domain SSNR is the additional flexibility to place different weights for different frequency bands of the spectrum. The word accuracy in percentage, abbreviated as PA, is measured in [1] as

$$\text{PA}=(N-D-S-I)/N \qquad (22)$$

where *N* denotes the number of reference labels, and D, S, and *I* represent the number of deletions, substitutions and insertions, respectively. The PA value illustrates that the recognition result is correct only when results and reference labels are identical.

To test the proposed and existed algorithms in ASR system, 400 utterances were recorded at a sampling frequency of 16 kHz. They are mixed with four different types of noises taken from NOISE92 database [16], say, white noise, pink noise, babble noise, and factory noise, at four various input SNR levels between 0 to 15 dB. The utterances are Hann-windowed with the analysis frame size is 256. Let the degree of freedom $a=0.1$ because MMSE speech spectral amplitude estimation method can produce a best result for $a\in[0.05,0.2]$ [17]. The proposed algorithm is compared with the MMSE speech PSD estimator and the OM-LSA-TCS algorithm [3], all of which assume the speech spectral amplitude obeys Rayleigh distribution. A MMSE noise estimator [18] is employed to compute noise PSD for all algorithms. An *a priori SNR* estimator based on temporal cepstrum smoothing (TCS) [19] is used to calculate the *a priori SNR*. The probability of speech presence q is measured with the method in [3]. Fig.1 demonstrates the frequency weighted SSNR of the speech which is enhanced by three different algorithms for different input SNRs and noise types over 400 utterances. It can be seen that the proposed approach has a better frequency weighted SSNR than the other two algorithms. Compared with traditional speech PSD estimation whose spectral amplitude is assumed to be Rayleigh distributed, the proposed speech PSD estimation under Chi *prior* has an obvious improvement in terms of SNR (about 2.15 dB). In the case of babble noise, the proposed algorithm gains a 0.4dB higher SNR than the OM-LSA-TCS algorithm. As for the other three noises, the proposed algorithm outperforms the OM-LSA-TCS with a much better SNR of about 1.22dB.

In Fig.2, we evaluate the PESQ of the speech enhanced by the compared algorithms under different noise environment. Overall, the PESQ of three algorithms are very similar to the SNR results in Fig.1. The proposed algorithm respectively has a 0.2 and 0.58 higher PESQ than the OM-LSA-TCS and the MMSERaleigh-spu algorithms. Furthermore, under the babble noise, the proposed algorithm obtains a slightly higher result (about 0.07 dB) than OM-LSA-TCS.

Finally, the comparison of recognition accuracy for the three algorithms is made. Fig.3 illustrates the recognition accuracy of the original noisy speech is very low, and all other algorithms can improve the recognition accuracy, however, the proposed algorithm still gains a higher recognition accuracy. Compared with the MMSERaleigh-spu and the OM-LSA-TCS algorithms, the proposed algorithm respectively improves the recognition accuracy by 13% and 18% except it is slightly higher than OMLSA-TCS algorithm (about 1.19%) in babble noise case. In addition, we find that in low SNR condition, the MMSERaleigh-spu approach which achieves lower SNR and PESQ has a 5.3% higher recognition accuracy than the OM-LSA-TCS obtaining higher SNR and PESQ except in the babble noise, where they have the close recognition accuracy. This phenomenon implies that the MMSE estimation of speech PSD can get a better performance than the MMSE speech spectral amplitude estimator.

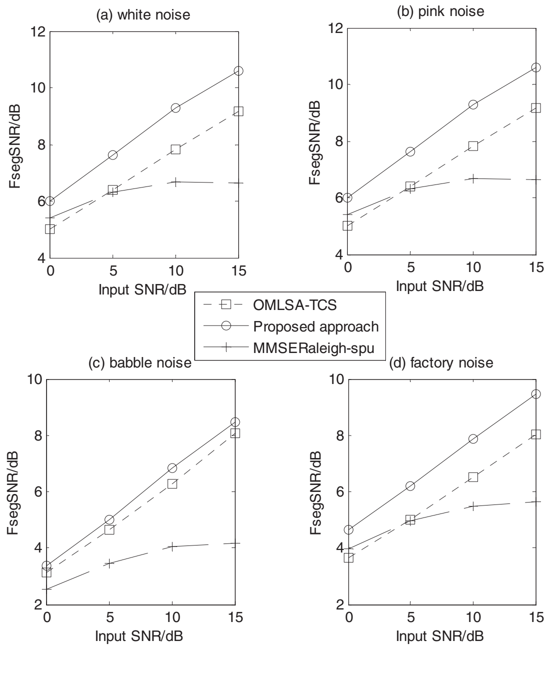

Fig.1. Comparison of frequency weighted segment SNR. (a), (b), (c) and (d) represent frequency weighted segment SNR results with white, pink, babble, factory noisy speech respectively.

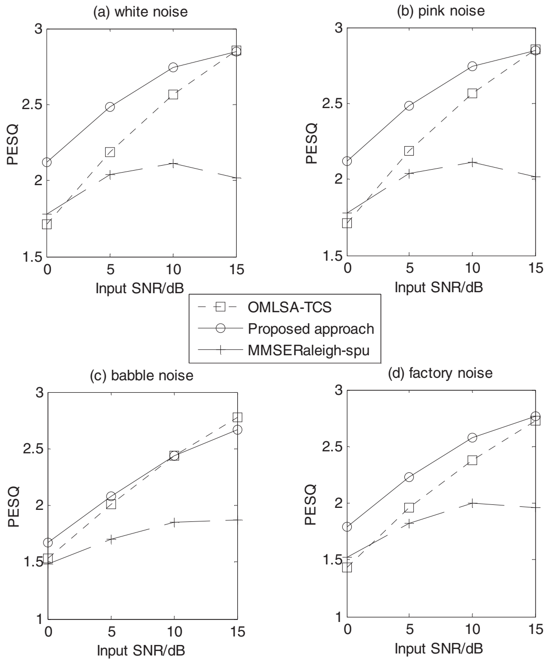

Fig.2. Evaluation of the PESQ for different input SNRs and noise types. (a), (b), (c) and (d) represent PESQ grade with white, pink, babble, factory noisy speech respectively.

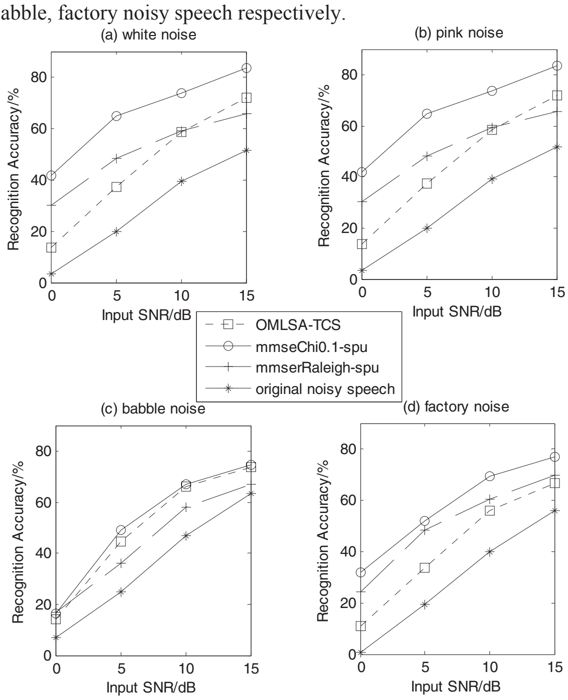

Fig.3. Recognition accuracy results from various enhanced speech. (a), (b), (c) and (d) respectively represent recognition accuracy with white, pink, babble, factory noisy speech.

V. CONCLUSIONS

This paper proposes a MMSE estimator of speech PSD with SPU for ASR. Both the speech PSD and the SPU estimator are based on the MMSE criterion and Chi spectral amplitude model. Experimental results indicate that the proposed algorithm has a better speech quality and higher recognition accuracy than the estimation method under Rayleigh distribution. Moreover, it also outperforms the estimation method, the log MMSE spectral amplitude OM-LSA-TCS, especially in low SNR scenarios.

ACKNOWLEDGEMENTS

This work is supported by the research projects of National Natural Science Foundation of China (No. 61501072) and of Chongqing Science and Technology Commission (cstc2015jcyjA40027).

APPENDIX

The following four equations (23)-(26) are used for the derivation of the estimation of clear speech PSD in Section III, which in Gradshteyn and Ryzhik's work [20] are 3.339, 6.631-4, 6.643-1, and 9.220-2.

$$I_0(|z|) = \tfrac{1}{2\pi}\int_0^{2\pi} \exp[\mathrm{Re}(z\,\mathrm{e}^{-j\theta_z})\,\mathrm{d}\,\theta_z], \qquad (23)$$

$$\int_0^\infty x^c e^{-ax^2} I_0(bx)\,\mathrm{d}\,x = \frac{\Gamma(0.5c+0.5)}{2a^{(1+c)/2}}\Phi(\frac{c+1}{2},1;\frac{b^2}{4a}), \qquad (24)$$

$$\int_0^\infty x^{\mu-0.5} e^{-\alpha x} J_{2\nu}(2\beta\sqrt{x})\,\mathrm{d}\,x = \frac{\Gamma(\mu+\nu+0.5)}{\Gamma(2\nu+1)\beta} e^{-\frac{\beta^2}{2\alpha}}\alpha^{-\mu} \qquad (25)$$

$$\times M\mu,\nu(\tfrac{\beta^2}{\alpha}),$$

$$M_{\lambda,\mu}(z) = z^{m+0.5} e^{-0.5z}\Phi(\mu-\lambda+0.5, 2\mu+1; z). \qquad (26)$$

where I_0 and $J_{2\nu}$ denote the zero order modified Bessel function and first kind Bessel function with order 2ν, respectively. $\Gamma(\cdot)$ is the gamma function. $M(\cdot)$ and $\Phi(\cdot)$ respectively represent the confluent hypergeometric function and Whittaker function.

VI. REFERENCES

[1] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*, Wiley, 2013.

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

[3] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113-116, 2002.

[4] 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 Trans. on Acoustics, Speech and Signal Processing* , vol.21, no.5, pp.1023-1034, 2013.

[5] J. Jensen and T. Zheng-Hua, "Minimum mean-square error estimation of Mel-Frequency cepstral features–a theoretically consistent approach," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.23, no.1, pp.186-197, 2015.

[6] K.M. Indrebo, R.J. Povinelli and M.T. Johnson, "Minimum mean-squared error estimation of mel-frequency cepstral coefficients using a novel distortion model," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.16, no.8, pp.1654-1661, 2008.

[7] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[8] T. Lotter and P. Vary, "Noise reduction by joint maximum a

posteriori spectral amplitude and phase estimation with super-Gaussian speech modelling,” *in Signal Processing Conference*, 2004, pp. 1457–1460.

[9] T. H. Dat, K. Takeda, and F. Itakura, “Generalized gamma modeling of speech and its online estimation for speech enhancement,” *IEEE ICASSP*, 2005, vol. 4, pp. 181–184.

[10] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 15, no. 6, pp. 1741 – 1752, 2007.

[11] A. Stark and K. Paliwal, "MMSE estimation of log-filterbank energies for robust speech recognition." *Speech Communication*, vol. 53, no.3, pp.403-416, 2011.

[12] B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma speech priors throughout," *IEEE ICASSP*, 2012, pp.4033-4036.

[13] J.M. Tribolet, P. Noll, B. McDermott and R.E. Crochiere , "A study of complexity and quality of speech waveform coders," in Acoustics, Speech, and Signal Processing, *IEEE ICASSP*, vol. 3, no., pp.586-590, 1978.

[14] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE ICASSP*, 2001, vol.2, pp.749-752.

[15] http://cmusphinx.sourceforge.net/wiki/download.

[16] A. Varga, and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 93, pp. 247-251.

[17] I. Andrianakis and R.W. Paul, "Mmse Speech Spectral Amplitude Estimators With Chi and Gamma Speech Priors," *IEEE International Conference on Acoustics, Speech, Signal Processing* , May 2006, vol. 3, pp. 14-19.

[18] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE ICASSP*, 2010, pp. 4266-4269.

[19] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in Proc. *IEEE ICASSP,* Apr. 2008, pp.4897-4900.

[20] I.S. Gradshteyn and I. M. Ryzhik "Table of Integrals, Series, and Products (Seventh Edition)." *Table of Integrals* vol. 103, no. 1, pp.1161–1171, 2007.