Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

INTER-FREQUENCY DEPENDENCY IN MMSE SPEECH ENHANCEMENT

Chunjian Li and Søren Vang Andersen

Aalborg University
Department of Communication Technology
Fredrik Bajers Vej 7 A3, 9220 Aalborg Ø
DENMARK
E-mail: cl@kom.auc.dk; sva@kom.auc.dk

ABSTRACT

In this paper an MMSE estimator of the complex short-time spectrum is considered for optimum noise reduction of speech. The correlation between frequency components is exploited to improve the estimation, especially of those components with low local SNR. Furthermore, by making use of both spectral envelope and time envelope, the estimator is able to suppress noise power in frequency domain and time domain simultaneously. The performance of the resulting estimator is found to be superior to the non-causal IIR Wiener filter. The enhanced signal suffers less spectral distortion, while achieving a lower mean squared error than the Wiener filter.

1. INTRODUCTION

In recent years, several MMSE approaches to speech enhancement appeared, including the non-causal IIR Wiener filter [1], the MMSE STSA estimator [2], and MMSE estimator using non-Gaussian priors [3]. Most of them can be characterized as short-time spectral amplitude estimators. A common characteristic of these methods is that they only process the spectral amplitude and use the noisy phase spectra to generate the enhanced signals (except for [3], in which the real parts and imaginary parts of the DFT coefficients are independently estimated). As an example, take the non-causal IIR Wiener filter with transfer function defined by

$$H_{WF}(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{vv}(\omega)} \tag{1}$$

where $P_{ss}(\omega)$ and $P_{vv}(\omega)$ denote the power spectral density of the speech signal and the uncorrelated additive noise, respectively. Hereafter we refer to (1) as the Wiener filter or WF. The transfer function of the WF is of zero phase and therefore it leaves the phase unprocessed. In addition, the WF does not exploit any inter-frequency dependency. This is a consequence of the stationarity assumption, and is another common point of the established MMSE approaches. One reason for not processing the phase spectrum is that phase is found to play a less important role in the human perception of speech [4]. An approximate threshold of phase perception was found in [4] corresponding to a local SNR of about 6 dB. If a frequency component in a frame has a local SNR higher than 6 dB, the phase distortion is not audible. The second common point comes as a consequence of assuming the speech frame to be infinitely long and stationary [5]. Although speech signals are known to be non-stationary and short-time processing is applied, this assumption is widely used in order to simplify the estimator.

In this paper we show that if these two restrictions are removed, better estimators are obtained.

This work was supported by the Danish National Center for IT research, Grant No.329.

2. PHASE SPECTRUM AND INTER-FREQUENCY DEPENDENCY

The motivation for involving phase information in the MMSE estimator is that, first of all, phase distortion is audible with low SNR speech. Processing low SNR speech with an estimator working only on the spectral amplitude brings reverberant effect and roughness to the enhanced speech. Recent works [6, 7] confirm that, especially for the voiced male speech, phase information is of clear perceptual importance. Moreover, the phase noise causes amplitude spectrum distortion through phase modulation when the signals are short-time processed using the overlap-add method. The rise of the spectrum in the valley between pitch harmonics causes audible artifacts and higher residual noise.

Secondly, phase coherence in the voiced speech is a significant source of correlation between frequency components. Two sources of correlation among frequency components can be identified. One is the finite-length window effect. It is known that the infinite Fourier matrix is the eigenvector matrix of an infinite Toeplitz matrix [8]. If we denote the covariance matrix of the speech samples, the in-

verse Fourier matrix, and the covariance matrix of the frequency components as $\mathbf{C_s}$, $\mathbf{F}$, and $\mathbf{C}_\theta$, respectively, we can write the covariance matrix as $\mathbf{C}_\theta = \mathbf{F}\mathbf{C_s}\mathbf{F^H}$. When $\mathbf{C_s}$ is a Toeplitz matrix, if the frame length of the Fourier analysis approaches infinity, $\mathbf{C}_\theta$ will become diagonal. However in general the speech signal is non-stationary, and very long windows are not applicable. The finite-length window effect causes the covariance matrix $\mathbf{C}_\theta$ to be generally non-diagonal. Therefore correlation exist among the frequency components. The second, and more interesting source of correlation is the phase coherence in voiced speech. Voiced speech can be modeled as an excitation pulse train filtered by an all-pole filter. The phase of the pulse train is approximately linear at pitch harmonic frequencies. After the filtering, the coherence in phase is maintained to some extend. If the phase coherence is lost, the voiced speech sounds reverberant [9]. The coherence in phase corresponds to energy localization in the time domain, which can be modeled by a time envelope.

Because of the importance of phase stated above, and because the optimum amplitude estimator and the optimum phase estimator do not coexist [2], we formulate the MMSE estimator as an estimate of the complex Fourier coefficients instead of independently derived spectral amplitude and phase estimators as in [2] or independent real parts and imaginary parts as in [3].

3. MMSE ESTIMATOR WITH TIME AND FREQUENCY ENVELOPES

The key feature of the new MMSE estimator is modeling the covariance matrix $\mathbf{C}_\theta$ as a full matrix instead of a diagonal matrix as in the WF. We will show the frequency domain MMSE estimator first and then transform it to time domain.

We use the following statistical model and problem formulation. The DFT coefficients of each speech segment are modeled as complex Gaussian random variables with zero mean and varying variance. Let $y(n,k)$, $s(n,k)$, $v(n,k)$ denote the n'th sample of noisy observation, speech, and additive white Gaussian noise of the k'th frame, respectively. Then

$$y(n,k) = s(n,k) + v(n,k). \tag{2}$$

Let $\theta(m,k)$ represent the m'th DFT coefficient of the k'th frame, defined by $\theta(m,k) = \sum_{n=0}^{N} s(n,k)exp(-j2\pi nm/N)$. For compactness we use vector representation and omit the index in the following discussion. Let $\mathbf{y}$, $\boldsymbol{\theta}$, $\mathbf{v}$, and $\mathbf{F}$ denote the vectors of y, θ, v and the inverse Fourier matrix respectively. Then (2) can be written as

$$\mathbf{y} = \mathbf{F}\boldsymbol{\theta} + \mathbf{v}. \tag{3}$$

The MMSE estimator can be shown to be the conditional mean [10]

$$\begin{aligned}\hat{\theta} &= E(\boldsymbol{\theta}|\mathbf{y}) \\ &= \mathbf{C}_\theta\mathbf{F^H}(\mathbf{F}\mathbf{C}_\theta\mathbf{F^H} + \mathbf{C_v})^{-1}\mathbf{y}\end{aligned} \tag{4}$$

where $(\cdot)^H$ denotes the Hermitian transpose and $\mathbf{C_v}$ denotes the covariance matrix of the noise $\mathbf{v}$. The covariance matrix $\mathbf{C}_\theta$ is generally unknown and must be replaced with an estimate. We propose here an approach to the estimation of $\mathbf{C}_\theta$ from the all-pole model of the speech. Let $q/A(z)$ denote the transfer function of the all pole model. Let $\mathbf{H}$ be the corresponding synthesis filter matrix derived from the all-pole model, and $\mathbf{r}$ be the residual vector, such that

$$\mathbf{s} = \mathbf{H}\mathbf{r}. \tag{5}$$

Since the residual is a white noise sequence with unit variance (for voiced speech it is a few impulses present periodically in the white noise), the covariance matrix $\mathbf{C_r}$ of $\mathbf{r}$ can be written as a diagonal matrix with the squared residual as the diagonal elements [1]. Once $\mathbf{C_r}$ is obtained, $\mathbf{C_s}$ and $\mathbf{C}_\theta$ can easily be found. We have

$$\mathbf{C_s} = \mathbf{H}\mathbf{C_r}\mathbf{H^H} \tag{6}$$

$$\mathbf{C}_\theta = \mathbf{F^H}\mathbf{C_s}\mathbf{F}. \tag{7}$$

Inserting (7) in (4) gives the MMSE short-time spectral estimator.

Fig.1 shows how the covariance matrix $\mathbf{C}_\theta$ estimated by this approach differs from the diagonal matrix underlying the standard WF. We can see that the off-diagonal elements are generally non-zero. At the brims of the matrix the cross-correlations are significant. This represents the windowing effect caused by the high spectral power at low frequencies. More interestingly, we see how inter-frequency dependency, especially between neighboring formants show up as significant off-diagonal elements in the covariance matrix. It is well known that a properly chosen window can reduce the correlation between frequency components but can not eliminate it. In Fig.1 a Hanning window is used, and we see that the remaining correlation is still significant and can be exploited to improve the estimator.

The frequency domain MMSE estimator given by (4) is mainly for the purpose of demonstrating the difference to the WF made by a full covariance matrix. In the estimation of the speech waveform, (4) is transformed back to time domain, giving the desired time domain MMSE estimator,

$$\hat{\mathbf{s}} = \mathbf{C_s}(\mathbf{C_s} + \mathbf{C_v})^{-1}\mathbf{y}. \tag{8}$$

Estimating the diagonal elements of $\mathbf{C_r}$ is equivalent to estimating the residual power distribution over the time axis. It can also be seen as estimating phase from the residual, because the power spectrum of the residual is known

[1] Here we ignore the long term correlation of the residual.

to be white. Estimating the squared residual from noisy observation is difficult. Our solution is to estimate the time envelope of the squared residual with simple shapes, i.e. a constant floor plus some pulses located periodically. These varying variances of residual represent time localization of energy. This is a major difference to the WF, which can be seen as using constant residual variance because of the stationary assumption. We estimate the residual envelope in a simple but effective way. The noisy speech signal is first lowpass filtered with cut-off frequency of 800 Hz. A 3-tap whitening filter is found by applying linear prediction on the filtered signal. The output of the low pass filter is then filtered by the whitening filter to get a reference residual. The position of the maximum in the reference residual is chosen as the first impulse position of the estimated residual envelope. According to an estimate of the pitch period the positions of remaining impulses are found. A pre-defined pulse shape is put on every impulse position. The pulse shape is chosen to be wider and smoother than a true residual impulse in order to gain robustness against error in estimating the impulse positions. The rest of the residual will be approximated with a constant whose amplitude is decided by keeping the average power of the estimated residual equal to unity. The estimation of the residual envelope is only needed for voiced frames. Fig.2 shows an example of the estimated residual envelope.

Because the above described MMSE estimator requires a spectral envelope and a temporal envelope as the prior knowledge, we hereafter refer to it as the Time-Frequency Envelope MMSE (TFE-MMSE) estimator.

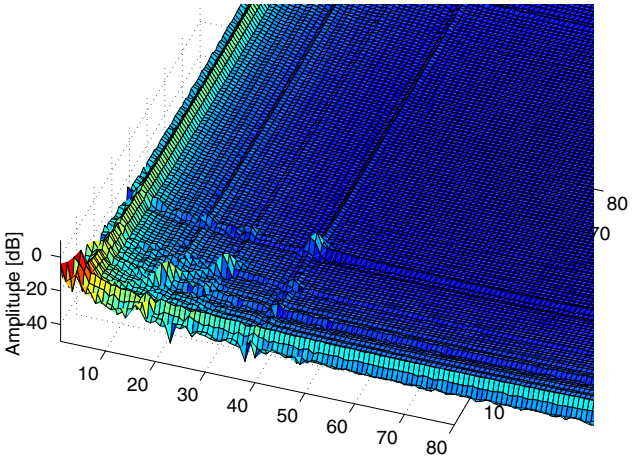

Fig. 1. Amplitude plot of the covariance matrix $\mathbf{C}_\theta$. Matrix size is 160 by 160 (only one corner of the matrix is shown).

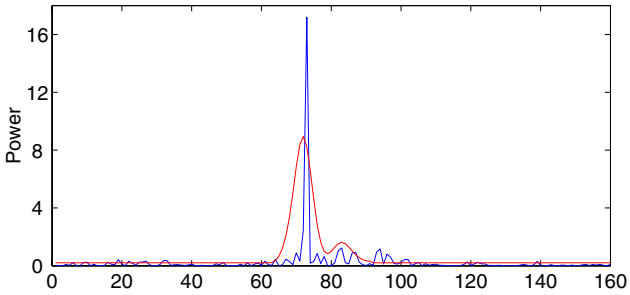

Fig. 2. The squared residual (blue) and the estimated envelope (red).

4. RESULTS

We first compare the performance of the TFE-MMSE estimator and the WF based on known spectral envelope of the signal. Since the purpose is to show that using the extra information about phase (or energy localization in time) it is possible to achieve lower mean squared error and lower spectral distortion at the same time, we first use known spectral envelopes for both estimators.

Both estimators run with 30 sentences from different speakers (15 male and 15 female) from the TIMIT database added with artificial white Gaussian noise at a signal-to-noise ratio of 0 dB. All sentences are 16kHz sampled, and segmented into frames of 160 samples. For the TFE-MMSE estimator, the time envelopes of the residual are estimated from noisy observations using the method described in section 3. For the output of both estimators, the SNR, Segmental SNR (segSNR) and Log-Spectral Distortion (LSD) to the original signal spectrum are calculated. The SNR is defined as the ratio of the total signal power to the total noise power in the sentence. The segSNR is defined as the average ratio of signal power to noise power per frame, omitting frames with a power more than 30 dB below average power. The LSD is defined as the distance between two log-scaled DFT spectra summed over all frequencies. The LSD is calculated only for voiced frames since for the unvoiced frames both estimators are identical.

From Table 1 we see consistent improvement of the TFE-MMSE estimator over WF in all three measurements. Fig.3 shows the signal spectrum of a voiced frame comparing with the spectrum of the output of the two estimators. Only the lower frequency half is plotted to show the details of the harmonic structure. It is seen that the TFE-MMSE estimator preserves the harmonic structure better than the WF.

To verify the performance in a practical scenario, estimated LPC coefficients are also used in the comparison. The LPC coefficients are estimated by a method similar to the decision directed method in [2]. The experimental setup is identical to the above one, except that input SNR

is now set to 10 dB. Table 2 shows the results. Significant improvements are observed with the segSNR measurement. The LSD of the TFE-MMSE estimator also improves significantly over the WF. Informal listening experiments show that the reduction of spectral distortion is significant.

	Male			Female		
	SNR	segSNR	LSD	SNR	segSNR	LSD
WF	10.73	5.21	290	10.57	5.59	347
TFE-MMSE	11.24	5.48	265	10.85	5.71	315
Improv.	0.51	0.27	25	0.28	0.12	32

Table 1. Performance of WF and the TFE-MMSE estimator with known AR coefficients. All SNR measures are in dB. Input SNR is 0 dB. Results are averaged over 30 sentences (by 15 male and 15 female speakers).

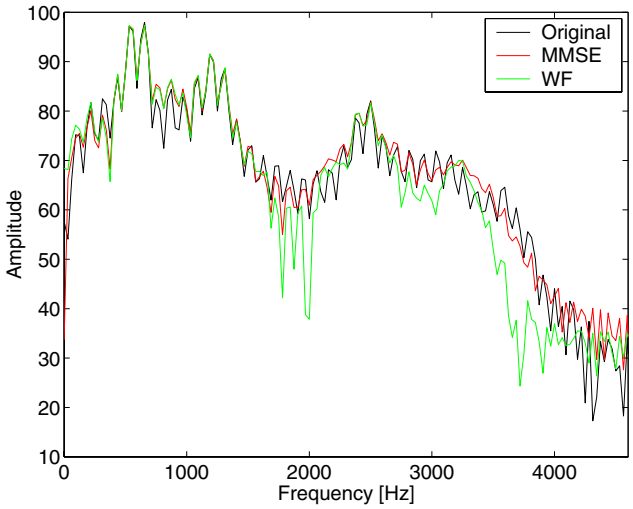

Fig. 3. A comparison of amplitude spectrum for the output of WF and the TFE-MMSE estimator to the original signal spectrum.

	Male			Female		
	SNR	segSNR	LSD	SNR	segSNR	LSD
WF	15.65	8.73	245	15.38	9.30	303
TFE-MMSE	16.71	9.42	183	16.48	9.83	231
Improv.	1.06	0.70	62	1.10	0.53	72

Table 2. Performance of WF and the TFE-MMSE estimator with estimated AR coefficients. Input SNR is 10 dB. Results are averaged over 30 sentences (by 15 male and 15 female speakers).

5. DISCUSSION

In the first part of this paper we stated the motivation of formulating an MMSE joint estimator of amplitude and phase spectrum, i.e., phase is of perceptual importance for low SNR sources, and estimating phase provides the additional information about the correlation of DFT coefficients which improves the amplitude spectrum estimation in return. We have avoided the widely used assumption of independent frequency components. This is justified by the fact that both finite-length window effect and time localization of energy (caused by phase coherence) in the voiced speech introduce correlation among the frequency components. Phase is known as hard to estimate, so we re-formulate the problem into estimating time envelope of the residual power. The MMSE joint spectral estimator (4) shows us that a full covariance matrix can exploit the inter-frequency dependency, achieving a better spectrum estimate. The algorithm is finally implemented as a time domain MMSE estimator (8).

The performance of the TFE-MMSE estimator and Wiener filter are compared based on known LPC coefficients as well as estimated ones. The TFE-MMSE estimator shows higher SNR and less spectral distortion than the WF. In the case of using estimated LPC coefficients, the improvement of segmental SNR and spectral distortion of the TFE-MMSE estimator over the WF is even more significant. This is because the spectral suppression and the temporal suppression benefit from each other making a better joint estimator.

6. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. vol.67, pp. 1586–1604, Dec. 1979.

[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. ASSP-32, pp. 1109–1121, Dec. 1984.

[3] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation With Gamma Distributed Speech Priors," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002. Proceedings.*, vol. 1, pp. 253–256, May 2002.

[4] P. Vary, "Noise Suppression By Spectral Magnitude Estimation - Mechanism and Theoretical Limits," *Signal Processing 8*, pp. 387–400, May 1985.

[5] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, New York : McGraw-Hill, 1958.

[6] H. Pobloth and W. B. Kleijn, "On Phase Perception in Speech," *ICASSP '99. Proceedings*, vol. 1, pp. 29–32, Mar. 1999.

[7] J. Skoglund, W. B. Kleijn, and P. Hedelin, "Audibility of Pitch-Synchronously Modulated Noise," *Speech Coding For Telecommunications Proceeding, IEEE*, vol. 7-10, pp. 51–52, Sept. 1997.

[8] R. M. Gray, *Toeplitz and Circulant Matrices : A review*, http ://ee.stanford.edu/ gray/toeplitz.pdf, 2002.

[9] T. F. Quatieri and R. J. McAulay, "Phase Coherence in Speech Reconstruction for Enhancement and Coding Applications," *International Conference on Acoustics, Speech, and Signal Processing, 1989.*, vol. 1, pp. 207–210, May 89.

[10] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*, Prentice Hall PTR, 1993.