SPEECH ENHANCEMENT USING MIXTURES OF GAUSSIANS FOR SPEECH AND NOISE

Ilyas Potamitis, Nikos Fakotakis, George Kokkinakis,

Wire Communications Lab., Electrical & Computer Engineering Dept.,
University of Patras, Rion-26 500, Patras, Greece
potamitis@wcl.ee.upatras.gr

Abstract. In this article we approximate the clean speech spectral magnitude as well as noise spectral magnitude with a mixture of Gaussians pdfs using the Expectation-Maximization algorithm (EM). Subsequently, we apply the Bayesian inference framework to the degraded spectral coefficients and by employing Minimum Mean Square Error Estimation (MMSE), we derive a closed form solution for the spectral magnitude estimation task adapted to the spectral characteristics and noise variance of each band. We evaluate our algorithm using true, coloured, slowly and quickly varying noise types (Factory and aircraft noise) and demonstrate its robustness at very low SNRs.

1 Introduction

In spite of key contributions on the subject of *Short-Time Spectral Attenuation* algorithms (STSA) as applied to speech enhancement [1], [2], [3], there is still need for further work primarily on the problem of balancing the trade-off between noise reduction and speech distortion. The STSA family of algorithms attempts to uncover the underlying spectral magnitude of speech by applying a gain function to the observed, noisy short-time spectra, where the gain function is related to the noise power spectrum.

In this work we propose a novel STSA algorithm that incorporates into a Bayesian formulation the long term pdf of each spectral band of an ensemble of clean recordings resulting in a better treatment of low energy spectral regions, while the spectral magnitude of noise is modelled by a mixture of Gaussians that allows for compensating the effect of time-varying noise types. A mixture of Gaussians is employed to account for the representation of the the magnitude of each spectral band of an ensemble of high quality speech (three minutes of phonetically balanced speech from speakers of both genders were found sufficient). The descriptive parameters of each mixture are derived from the observed spectral bands of the clean data by employing the EM algorithm. Assuming the availability of noisy data, we incorporate a Gaussian mixture model for the background noise and derive the descriptive statistics of the mixtures using the EM algorithm.

Objective (SNR and Itakura-Saito (IS) measures) as well as subjective evaluation of signals degraded with additive Factory noise, and DC-3 aircraft noise at low SNRs ranging from -10 to 10 dB confirm the benefit of our approach.

2 Description of the algorithm

Let s(m) denote the clean time-domain signal corrupted by noise n(m) where (m) is the discrete-time index. The observed signal x(m) is given by: x(m)=s(m)+n(m) and is subjected to Short Time Fourier Transform (STFT). Based on the generalized spectral subtraction framework [3], we can derive a linear-spectral representation of a clean speech signal corrupted by additive noise using a 2N point FFT as:

$$x^{\alpha}_{\kappa,l}=s^{\alpha}_{\kappa,l}+n^{\alpha}_{\kappa,l} \quad \kappa=0,\ldots,N. \tag{1}$$

$\{x_\kappa\}$ denotes the spectral magnitude of the degraded sub-band $\{\kappa\}$, $\{n_\kappa\}$ the noise spectral magnitude, {l} the frame index and $1\leq\alpha\leq2$. Prior knowledge about the time frequency distribution of $\{s_\kappa\}$ is provided by a mixture of Gaussians that model the undegraded spectral bands of the available clean speech corpora (Eq. 2). Practically, 2-3 minutes of clean speech, unrelated to the signals to be enhanced, were found sufficient to tune the free parameters of the algorithm.

For notational convenience we set $x=x^\alpha$, $s=s^\alpha$, $n=n^\alpha$ and we drop subscript $\{\kappa\}$, {l} implying that the subsequent analysis holds for every time-trajectory of spectral sub-band $\{\kappa\}$ independently, in the linear spectral domain. We have found that setting $\alpha=3/2$ optimises performance, though, the subsequent analysis holds for every $\{\alpha\}$.

$$f(s) = \sum_{m=1}^{M} p_m G(s;\mu_m,\sigma_m{}^2), \qquad \sum_{m=1}^{M} p_m = 1 \tag{2}$$

The pdf of the spectral magnitude of noise is modeled by a mixture of Gaussian as:

$$f(n) = \sum_{\kappa=1}^{K} p_\kappa G(n;\mu_\kappa,\sigma_\kappa{}^2), \qquad \sum_{\kappa=1}^{K} p_\kappa = 1 \tag{3}$$

where, {M} is the total number of mixture components, p_m, μ_m and σ_m are the prior probability, mean and standard deviation of the *m*th Gaussian speech mixture, while p_κ, μ_κ and σ_κ are the prior probability, mean and standard deviation of the *κ*th Gaussian noise mixture. The descriptive statistics of the Gaussian mixture i.e p_m, μ_m, σ_m, p_κ, μ_κ and σ_κ are computed by the EM algorithm. Means are initialized uniformly over the interval of each spectral band magnitude, while weights are set to equal values and variance is lower-bounded to avoid picking narrow spectral peaks. Subsequently we proceed in deriving the MMSE estimation of the underlying spectral coefficients {s} as $S_{MMSE}=E\{s|x\}=\int s f(s|x)ds$. The pdf of {s} given the observation {x} is derived by the Bayesian formula $f(s|x)=f(x|s)f(s)/f(x)$. Combining $f(s|x)$ and S_{MMSE} results in:

$$S_{MMSE} = \frac{\int s f(x|s)\, f(s)ds}{\int f(x|s)\, f(s)ds} \tag{4}$$

Substituting Eq. 2 and Eq. 3 into Eq. 4 and by carrying out some simple algebra, we derive the underlying spectral magnitude in terms of an integral which is expressed in closed form through parabolic cylinder functions. (See Appendix for details in the definition and evaluation of the integrals I_1, I_2). Based on the Gaussian assumption for the spectral magnitude pdf of noise the MMSE estimation of the underlying clean spectral magnitude is

$$S_{MMSE} = \frac{\sum_{\kappa=1}^{K}\sum_{m=1}^{M} \frac{p_\kappa}{\sigma_\kappa}\frac{p_m}{\sigma_m} I_2(b_{m,\kappa}, c_{m,\kappa}, d_{m,\kappa})}{\sum_{\kappa=1}^{K}\sum_{m=1}^{M} \frac{p_\kappa}{\sigma_\kappa}\frac{p_m}{\sigma_m} I_1(b_{m,\kappa}, c_{m,\kappa}, d_{m,\kappa})} . \quad (5)$$

Based on the fact that the information of the speech signal is encoded in the frequency domain and that human hearing is relatively insensitive to phase information, we focus on the short-time amplitude of the speech signal leaving the noisy phase unprocessed. After the enhancement procedure has been applied, noisy phase is added back and the time-domain signal is subsequently reconstructed using inverse FFT and the weighted overlap and add method.

3 Simulation and results

We performed speech enhancement experiments using real factory noise taken from the NOISEX-92 database as well as aircraft noise. Each noise type was added to 10 clean speech files of 5 sec. mean duration so that the corrupted waveform ranges from –10 to 10 SNRdB. The number of Gaussian mixtures is set to nine for noise and six for speech as the objective measures indicated marginal gain by augmenting the number of mixtures. The SNR and the IS measures of the enhancement obtained by our technique are shown in Fig. 1a and Fig. 1b respectively. The IS distortion measure is based on the spectral distance between AR coefficient sets of the clean and enhanced speech waveforms over synchronous frames of 15ms duration and is heavily influenced due to mismatch in formant locations. As indicated in Figs. 1a, 1b, our method consistently effected a strong enhancement over all SNRs while the low energy parts of the spectrum are preserved even at 0 dB SNR. We attribute this fact to the *a-priori* modelling of clean spectral bands and to the mixture modelling of noise that permits a variable weighting for the generalized magnitude of noise for each spectral band and each frame. Parallel listening tests are well correlated with the objective measures and indicate that periodic components are strongly suppressed.

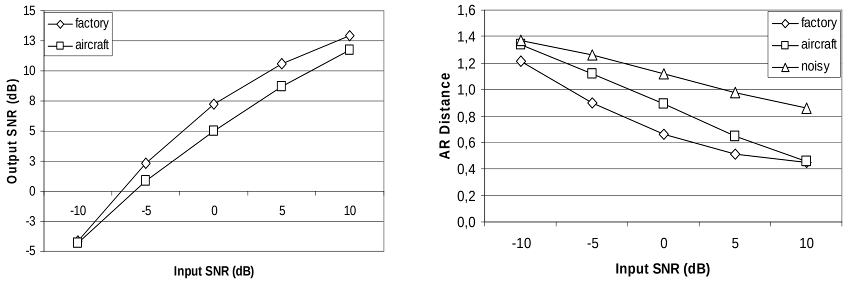

Fig. 1: a) SNR measurements, **b)** Itakura-Saito AR-distance measurements. Noise is Factory noise (NOISEX-92) and of an approaching DC-3 aircraft noise.

4 Conclusions

The application of SNR and Itakura-Saito (IS) measures confirmed the benefit of modeling clean spectral bands and the spectral bands of noise with a mixture of Gaussians. The key idea of independent modeling of the multimodal, heavy tail pdf of the magnitude of spectral bands with a mixture of Gaussians combined with MMSE formulation, can supply an efficient solution to a series of spectral estimation problems. We demonstrated the benefit of this enhancement technique at very low SNRs with true, slowly and quickly varying noise types. We suggest that the incorporation of the long term pdf of each band as *a-priori* information leads to estimators adapted to the spectral characteristics and noise variance of each band leading to better treatment of low energy time-frequency regions. Future work focuses on the incorporation of different techniques for the adaptive estimation of the variance of noise and the adaptive estimation of the descriptive statistics of the Gaussian mixture of noise as well as combining soft decision rules for estimating speech presence uncertainty.

Appendix

$$I_{\nu} = \int_{0}^{+\infty} s^{\nu-1} \exp(-b_m s^2 - c_m s - d_m) ds = \exp(-d_m)(2b_m)^{-\frac{\nu}{2}} \Gamma(\nu) \exp(\frac{c^2{}_m}{8b_m}) D_{-\nu}(\frac{c_m}{\sqrt{2b_m}})$$

$D_{\nu+1}$-zD_{ν}(z)+ν$D_{\nu\text{-}1}$(z)=0, (Eq. 3.462, [4])

$$D_{-1}(z) = \exp\left(\frac{z^2}{4}\right)\sqrt{\frac{\pi}{2}}\left\{1 - erf\left(\frac{z}{\sqrt{2}}\right)\right\} \; D_{-2}(z) = \exp\left(\frac{z^2}{4}\right)\sqrt{\frac{\pi}{2}}\left\{\sqrt{\frac{\pi}{2}}\exp\left(\frac{z^2}{4}\right) - z\left[1 - erf\left(\frac{z}{\sqrt{2}}\right)\right]\right\}$$

$$b_{m,\kappa} = \frac{1}{2}\left(\frac{1}{\sigma^2{}_m} + \frac{1}{\sigma^2{}_\kappa}\right), \; c_{m,\kappa} = -\left(\frac{\mu_m}{\sigma^2{}_m} + \frac{x - \mu_\kappa}{\sigma^2 \kappa}\right), \; d_{m,\kappa} = \frac{1}{2}\left(\frac{(x - \mu_\kappa)^2}{\sigma^2{}_\kappa} + \frac{\mu^2{}_m}{\sigma^2{}_m}\right)$$

References

[1] McAulay R., Malpass M., (1980), “Speech enhancement using a soft decision noise suppression filter,” *IEEE Trans. Speech and Audio Processing*, Vol. 28, no. 2, pp. 137-145.

[2] Ephraim Y., Malah D., (1984), “Speech Enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. ASSP*, Vol. 32, pp. 1109-1121.

[3] Gong Y., (1995), “Speech recognition in noisy environments: A survey,” *Speech Communication*, 16, pp.261-291.

[4] Gradshteyn I., Ryzhik M., Jeffrey A. editor., Fifth edition, (1994), “Table of Integrals, Series and Products,” *Academic Press*, pp. 1094-1095, Eq. 9.247, Eq. 9.254, Eq. 3.462.