On the Application of Markov Random Fields to Speech Enhancement

By Ioannis Andrianakis and Paul R. White

The Institute of Sound and Vibration Research,
University of Southampton, Southampton, SO17 1BJ, UK
e-mail: {ia,prw}@isvr.soton.ac.uk

Abstract

We report on the development of a novel Bayesian estimator for speech enhancement, which is capable of modelling the time and frequency dependencies of speech. Central to the development of the estimator is a conditional prior that is derived from the Markov Random Field theory. The proposed prior is a conditional Gaussian prior that defines the distribution of the amplitude of a speech STFT sample conditioned on the values of its time and frequency neighbours. This formulation allows the explicit inclusion in the estimation model of both time and frequency dependencies that exist among the amplitudes of speech STFT samples. The resulting estimator presents an enhanced ability in preserving the weaker speech spectral components compared to alternative estimators.

1. Introduction

The objective of this work is the development of a Bayesian speech enhancement estimator that will account for the time and frequency dependencies of speech. An inspection of a speech spectrogram reveals that successive spectral amplitude samples within a frequency bin are correlated. Additionally, frequency dependencies within the same time frame exist due to the common modulation, in time, of the frequency bins and also due to the spectral leakage caused by the windowing functions used in the STFT transformation.

In traditional amplitude estimation algorithms for speech enhancement (i.e. Ephraim & Malah (1984), Martin (2005)) the speech spectral amplitude samples are assumed to be mutually independent. The time dependencies of speech are incorporated in the estimation model with the decision directed method, which is used for the estimation of the a priori SNR. Little work, has been done to incorporate the frequency dependencies of speech.

In this work, we propose the modelling of speech spectral amplitudes with a prior that is capable of modelling both time and frequency dependencies of speech. The prior is derived from the theory of Markov Random Fields, and defines the distribution of a spectral amplitude sample conditioned on the values of some of its neighbours. The introduction of the neighbourhood allows the incorporation of the time and frequency dependencies in the estimation model. Based on the proposed prior, we then derive an efficient MAP estimator of the speech spectral amplitude.

The organisation of this paper is as follows: In section 2 we present the estimation framework and introduce the Markov Random Field prior. The MAP estimator is derived in section 3, where we also propose values for the prior's parameters. Finally, in section 4 we present results from an objective evaluation of the proposed algorithm and highlight its main benefits and limitations.

2. Statistical model

Let us assume that we observe the noisy speech signal x, which is the sum of the clean speech s and the noise signal n. The latter two signals are assumed to be zero mean and uncorrelated. For the STFT representations of x, s and n it will hold:

$$X(k,l) = S(k,l) + N(k,l) \tag{2.1}$$

where k and l are the frequency and time frame indices correspondingly. The typical objective of a speech enhancement algorithm is the estimation of $S(k,l)$ when only the STFT of the noisy speech is observed. $X(k,l)$ and $S(k,l)$ in their polar forms will be denoted as $X \equiv R\,\mathrm{e}^{j\psi}$ and $S \equiv A\,\mathrm{e}^{j\phi}$.

The algorithm we propose estimates the speech spectral amplitude A, which is then combined with the phase of noisy speech ψ to yield an estimate of S, that is, $\hat{S} = \hat{A}\,\mathrm{e}^{j\psi}$. The estimation of A is performed with the Maximum A Posteriori (MAP) estimator. In a traditional speech enhancement framework, the MAP estimator is formed as (see Andrianakis & White (2006)):

$$\hat{A}_i = \arg\max_{A_i} \left[\ln\left(p(R_i|A_i)p(A_i)\right)\right] \tag{2.2}$$

where the subscript i is a shorthand notation for the $(k,l)^{\text{th}}$ sample of an STFT quantity. The density $p(R_i|A_i)$ is often referred to as likelihood, while $p(A_i)$ is known as the prior. The inclusion of the time and frequency dependencies of speech into the estimation model can be performed by making the prior distribution $p(A_i)$ conditional on some samples of interest, $A_{\mathrm{n}(i)}$, which we call neighbours of A_i and will be defined in a subsequent paragraph. The conditional prior can be written as $p(A_i|A_{\mathrm{n}(i)})$ and the MAP estimator with the incorporation of the new prior will then become:

$$\hat{A}_i = \arg\max_{A_i} \left[\ln\left(p(R_i|A_i)p(A_i|A_{\mathrm{n}(i)})\right)\right] \tag{2.3}$$

The construction of the above conditional prior can be based on the theory of Markov Random Fields (MRF) (see Besag (1974)). MRFs can be considered as 2 dimensional extensions of Markov Chains and can be defined on N points A_i with $i \in P$ and $P = \{1, 2, ..., N\}$, which typically lie on a lattice. The specific MRF model we are using is the conditional Gaussian MRF prior (Besag (1974)), which is given by:

$$p(A_i|A_{\mathrm{n}(i)}) \propto \exp\left(-\tfrac{1}{2\sigma^2}\left(A_i - \textstyle\sum_{j\in\mathrm{n}(i)} b_{ij}A_j\right)^2\right) \tag{2.4}$$

where b_{ij} are the MRF parameters and σ^2 is the second moment of A_i.

The algorithm described here uses two neighbourhoods with the above priors: the first, which is used for the unvoiced or speech absent frames, is a 4-neighbour system given by:

$$A_{\mathrm{n}(i)} = \{A(k,l-1), A(k,l+1), A(k+1,l), A(k-1,l)\} \tag{2.5}$$

The second neighbourhood system we propose is used for the voiced frames and is a 'harmonic' neighbourhood:

$$A_{\mathrm{n}(i)} = \{A(k,l-1), A(k,l+1), A(k+k_f,l), A(k-k_f,l)\} \tag{2.6}$$

where k_f is the frequency bin number that corresponds to the pitch frequency of frame l.

3. Derivation of the Estimator

In this section we derive the MAP estimator that is based on the Gaussian MRF speech prior (eq. 2.3). To do so we first need to derive the expression for the likelihood $p(R_i|A_i)$. Assuming that the STFT coefficients of noise follow a complex Gaussian distribution, the likelihood is (Andrianakis & White (2006)):

$$p(R_i|A_i) = \frac{2R_i}{\sigma^2_{N,i}} \exp\left(-\frac{R_i^2+A_i^2}{\sigma^2_{N,i}}\right) I_0\left(\frac{2R_iA_i}{\sigma^2_{N,i}}\right) \tag{3.1}$$

where $\sigma^2_{N,i} = \mathrm{E}[|N_i|^2]$ and $I_0(x)$ is the zeroth order modified Bessel function of the first kind. Approximating the Bessel function using the formula $I_0(x) \approx \mathrm{e}^x/\sqrt{2\pi x}$, $p(R_i|A_i)$ can be written as:

$$p(R_i|A_i) \propto A_i^{-0.5} \exp\left(-\frac{(R_i-A_i)^2}{\sigma^2_{N,i}}\right) \tag{3.2}$$

Substituting eqs. 3.2 and 2.4 in 2.3 and solving the maximisation problem by taking the first derivative of the resulting expression w.r.t A_i, the estimator turns out to be:

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{3.3}$$

where $\zeta_1 = \dfrac{2R_i\sigma^2 + \sigma^2_{N,i}\sum_{j\in \mathrm{n}(i)} b_{ij}A_j}{2(2\sigma^2+\sigma^2_{N,i})}$ and $\zeta_2 = \dfrac{1}{4}\dfrac{2\sigma^2\sigma^2_{N,i}}{2\sigma^2+\sigma^2_{N,i}}$

For certain values of its input parameters the square root in the above estimator can have a negative value. This is due to a singularity at zero, introduced by the approximation of the Bessel function $I_0(x)$. Following the same rationale as in Andrianakis & White (2006), the output of the above estimator is used when the argument of the square root is non negative, otherwise, the noisy sample R_i is suppressed by a fixed amount (50 dB).

The definition of ζ_1 in eq. 3.3 reveals that the values of $A_{\mathrm{n}(i)}$ are required for the estimation of A_i. However, these values are not available as we do not directly observe A. In the processing of a speech utterance we assume that the estimation proceeds from smaller to larger time frame and frequency indices. During the estimation of A_i, an estimate for a set $A_{j\in Y}$ of its neighbours has already been calculated, while for a second set of neighbours $A_{j\in U}$ an estimate is not yet available. The sets Y and U are $Y = \{(k-1,l),(k,l-1)\}$, $U = \{(k+1,l),(k,l+1)\}$ for the unvoiced or speech absent frames and $Y = \{(k-k_f,l),(k,l-1)\}$, $U = \{(k+k_f,l),(k,l+1)\}$ for the voiced frames. The estimates we use for $A_{j\in U}$ are $\hat{A}_j = \max\left(\left(X_j^2 - \sigma^2_{N,j}\right)^{0.5}, \epsilon\ \sigma^2_{N,j}\right)$ with $\epsilon = 0.0032$.

The weights we used for the neighbours were $b_{ij} = 0.1$ if $j \in Y$ and $b_{ij} = 0.06$ if $j \in U$ and were determined by simulations. The value of σ^2 was calculated from the a priori SNR ξ based on the relation $\xi_i = \sigma_i^2/\sigma^2_{N,i}$. The a priori SNR was calculated with the decision directed approach proposed in Ephraim & Malah (1984) and smoothing parameter $\alpha = 0.98$.

4. Results

The proposed algorithm was evaluated using the average segmental SNR (SegSNR) and the PESQ (ITU recommendation P.862) objective measures. The clean speech, which consisted of 48 sentences from the TIMIT database, was corrupted by white Gaussian noise and recorded car noise in 3 different input SegSNR levels. The results of the proposed algorithm (MAP-MRF) are compared against the results of the MAP amplitude estimator with Chi speech priors and shape parameter $a = 1$ (see Andrianakis & White

	SegSNR [dB]			PESQ		
Input	0	10	20	2.11	2.80	3.46
MAP-Chi	7.56	13.30	20.66	2.72	3.28	3.88
MAP-MRF	7.69	13.97	21.50	2.87	3.46	3.96

TABLE 1. Objective measure results for white noise

	SegSNR [dB]			PESQ		
Input	0	10	20	2.89	3.49	4.07
MAP-Chi	10.82	17.03	24.22	3.40	3.88	4.22
MAP-MRF	10.89	17.14	24.40	3.47	3.93	4.24

TABLE 2. Objective measure results for car noise

(2006) for details). The latter algorithm (MAP-Chi) is a special case of the proposed MAP-MRF algorithm for $b_{ij} = 0$.

The MAP-MRF algorithm requires a pitch estimate for each STFT frame. The pitch estimates were extracted with the 2.4 kbps MELP Proposed Federal Standard Speech Coder, which is based on autocorrelation. To simulate a realistic scenario the pitch estimates were extracted from the noisy speech, which was preprocessed with the MAP-Chi algorithm. The latter algorithm is computationally efficient and hence suitable as a preprocessing step.

The objective scores results for the two examined algorithms are shown in tables 1, 2. The proposed MAP-MRF algorithm achieves higher scores in both objective measures and for all input SegSNR levels. Informal listening tests and examination of spectrograms reveal that the coupling imposed by the MRF prior achieves the recovery of some weak speech spectral components which are suppressed by the MAP-Chi algorithm. In particular, the frequency coupling recovers some speech harmonics, which have a low SNR, while the time coupling results in a better recovery of speech at its onset. The main drawback of the proposed method is that the spurious spectral peaks have a higher amplitude compared to those of the MAP-Chi algorithm and as a result the residual noise has more musical character. Our current research efforts are trying to address this problem and result in a scheme that will combine the uniform residual noise with the enhanced ability of the proposed MRF prior to restore the weaker speech spectral components.

REFERENCES

ANDRIANAKIS, I. & WHITE, P. R. 2006 MMSE speech spectral amplitude estimators with Chi and Gamma speech priors. In *Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06* **3**, 1068–1071.

BESAG J. 1974 Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**, no.2, 192–236.

EPHRAIM, Y. & MALAH, D. 1984 Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-32**, no.6, 1109–1121.

MARTIN R. 2005 Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech and Audio Processing* **13**, no.5, 845–856.