A Speech Enhancement Algorithm Based on a Chi MRF Model of the Speech STFT Amplitudes

Yiannis Andrianakis and Paul R. White, *Member, IEEE*

***Abstract*—A speech enhancement algorithm that takes advantage of the time and frequency dependencies of speech signals is presented in this paper. The above dependencies are incorporated in the statistical model using concepts from the theory of Markov Random Fields. In particular, the speech short-time Fourier transform (STFT) amplitude samples are modeled with a novel Chi Markov Random Field prior, which is then used for the development of an estimator based on the Iterated Conditional Modes method. The novel prior is also coupled with a 'harmonic' neighborhood, which apart from the immediately adjacent samples on the time frequency plane, also considers samples which are one pitch frequency apart, so as to take advantage of the rich structure of the voiced speech time frames. Additionally, central to the development of the algorithm is the adaptive estimation of the weights that determine the interaction between neighboring samples, which allows the restoration of weak speech spectral components, while maintaining a low level of uniform residual noise. Results that illustrate the improvements achieved with the proposed algorithm, and a comparison with other established speech enhancement schemes are also given.**

***Index Terms*—Chi, Gaussian, Markov random fields, speech enhancement, short-time Fourier transform (STFT) estimation.**

I. Introduction

THE short-time Fourier transform (STFT) of speech is a representation that has a rich structure with dependencies both along the time and frequency axes. Cohen [1] has shown that consecutive samples within a frequency bin are highly correlated, while the decision directed approach [2] for the estimation of the *a priori* signal-to-noise ratio (SNR) owes its success largely to the exploitation of the dependencies between successive spectral amplitude samples of speech. Correlations also exist between consecutive samples along the frequency axis of the STFT, which stem not only from the spectral leakage caused by the tapered windows used in the calculation of the STFT, but also from the common modulation of the amplitude of samples that belong to adjacent harmonics of the voiced speech frames, as reported in [3]. All this information that is encapsulated in the STFT representation of speech can prove very helpful in its restoration when degraded by noise.

In this paper, we present an algorithm that enhances speech corrupted by additive and uncorrelated noise, by taking into account the time and frequency dependencies of speech signals. The incorporation of these dependencies into the algorithm is achieved by modeling the amplitude of the speech STFT with a Markov Random Field (MRF) prior. MRFs are spatial stochastic processes, which can be considered as two-dimensional extensions of Markov Chains. Therefore, as the value of a random variable (r.v.) in a Markov Chain depends on some r.v.'s that precede it, the value of a r.v. in an MRF depends on a number of r.v.'s which are considered as its neighbors, in the two dimensional space in which the MRF is defined.

MRFs have found extensive application in image processing problems, due to their ability to model the spatial dependencies of images, particularly over smooth areas. Some characteristic examples can be found in [4]–[8]. In speech processing on the other hand, MRFs have not been widely used to date. We indicatively mention the work of Gravier *et al.* [9] , where MRFs were employed in a speech recognition problem, and the work of Andia [10], where MRFs were used for the restoration of STFT data that were missing due to severe contamination from tonal noises. To the best of our knowledge, here is the first time that MRFs are used in the enhancement of speech that is corrupted with broadband noise.

In the present work, we introduce a generalization of the established Gaussian MRF, which extends our previous work that incorporated it [11] and allows the derivation of a well defined speech spectral amplitude estimator, capable of enhancing speech while avoiding the artefacts known as musical noise. The novel model is termed Chi MRF, because it constitutes a generalization of the Gaussian MRF, in the same sense that the Chi density function [12] is a generalisation of the Gaussian density. We also introduce a "harmonic" neighborhood, according to which, each sample in an unvoiced or speech absent time frame interacts with the closest four neighbors on the time-frequency plane, while for the voiced speech frames, the samples which are one pitch frequency apart are also considered.

A key feature of the proposed algorithm is the estimation scheme of the MRF prior's weights, which determine the interaction of each sample with its neighbors. The weights are estimated adaptively, via a function of the speech spectral variance at the sites of the neighbors. According to this scheme, neighbors with a variance higher than that of the noise are more likely to contain speech and contribute more to the final estimate. On the other hand, neighbors with a small variance should contain mostly noise and thus have a relatively smaller influence. The

Manuscript received July 16, 2008; revised April 08, 2009. First published May 05, 2009; current version published August 14, 2009.This work was supported by the EPSRC under Grant GR/S30238/01. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yariv Ephraim.

Y. Andrianakis is with the National Oceanography Center, Southampton, SO14 3ZH, U.K. (e-mail: i.andrianakis@soton.ac.uk).

P. R. White is with the Institute of Sound and Vibration Research, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail: prw@isvr.soton.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2022199

1558-7916/$26.00 © 2009 IEEE

result is the ability of restoring speech spectral components that are immersed in noise, while keeping the level of the residual noise low.

The outline of this paper is as follows. In Section II, we review the fundamental elements of the MRF theory, on which the proposed Adaptive Chi MRF (ACMRF) algorithm is based. In Section III, the statistical model is discussed, which includes the introduction of the Chi MRF priors and the definition of the "harmonic" neighborhood. The speech spectral amplitude estimator is derived in Section IV, where the formulae for the adaptive estimates of the neighbors' weights are also given. The proposed algorithm is evaluated in Section V, where a comparison with other established speech enhancement algorithms is also shown and finally, Section VI concludes this paper.

II. Theoretical Background

In this section, we present the basic theory and some fundamental concepts of Markov Random Fields. Our presentation, which is primarily based on [13] and to some extend on [5], [7], and [8], is focused on those aspects of the MRF theory that will be necessary for the development of the proposed speech enhancement algorithm in the subsequent sections. A more extensive treatment of the theory of MRFs can be found in the above references.

A. Markov Random Fields and the Hammersley–Clifford Theorem

Suppose that we have a vector of random variables $X = [X_1, \ldots, X_q]$ and let $x = [x_1, \ldots, x_q]$ denote a realization of X. We define the space $\mathcal{S}_i$ of the random variable X_i as

$$\mathcal{S}_i = \{x_i : p(x_i) > 0\}, \quad \text{with} \quad i \in Q = \{1, \ldots, q\}$$

where $p(x_i)$ is the probability density function of X_i. The joint probability density function of the X_i random variables is denoted as $p(x) = p(x_1, \ldots, x_q)$. The space $\mathcal{S}$ of the vector of random variables X is given by the Cartesian product of the individual $\mathcal{S}_i$'s

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_q.$$

A central concept in the development of Markov Random Fields is that of a *neighbor*. Given two random variables X_i and X_j with $i \neq j$, we say that X_j is a neighbor of X_i if and only if the conditional distribution $p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_q)$ is a function of x_j. The neighbors of the random variable X_i are denoted by $X_{\mathrm{n}(i)}$. We also require that if the realizations $X_1 = x_1, \ldots, X_q = x_q$ can occur individually, they can also occur simultaneously. More formally, if $p(x_i) > 0, \forall i \in Q$ then $p(x_1, \ldots, x_q) > 0$. The last condition is called the *positivity* condition and is usually satisfied in practice.

Definition: A Markov Random Field is a collection of interacting random variables with joint probability density function $p(x)$ for which:

1) the positivity condition holds;
2) for each X_i there is a defined set of r.v.'s $X_{\mathrm{n}(i)}$, which are called neighbors and the following statement is true

$$p(x_i | x_{\{Q-i\}}) = p(x_i | x_{\mathrm{n}(i)}) \quad \forall i \in Q.$$

.	.	.	.	.
.	.	X	.	.
.	X	O	X	.
.	.	X	.	.
.	.	.	.	.

Fig. 1. First-order neighborhood. The cells marked with "x" are the neighbors of "o". The distribution of "o" is independent of all the other cells if the values of the "x"s are known.

where $\{Q - i\}$ is a shorthand notation for the set of indices $j \in Q$ with $j \neq i$.

An intuitively appealing method of constructing an MRF is via the definition of conditional density functions. This method allows to define explicitly the interactions between a random variable and its neighbors, which is not as straightforward to achieve with the direct construction of a joint density function. The conditional density approach however, is hindered by the disadvantage that not all conditional densities yield a valid joint distribution for the process. Instead, there is a number of unobvious consistency conditions that the conditional densities need to satisfy, in order to yield a valid joint MRF density function. The most general form a conditional density can assume can be obtained from the class of joint densities that constitute valid MRF schemes, as it was shown in [13]. The class of valid MRF densities is defined by the Hammersley–Clifford theorem.

Theorem (Hammersley–Clifford): Let $p(x) > 0$ with $x \in \mathcal{S}$, denote a probability density function satisfying the positivity condition. Then, $p(x)$ is a Markov Random Field if and only if

$$p(x) \propto \prod_{\mathcal{C}} \Psi_{\mathcal{C}}(x_{\mathcal{C}}). \tag{1}$$

The functions $\Psi_{\mathcal{C}}(x_{\mathcal{C}})$ are chosen arbitrarily, subject to $0 < \Psi_{\mathcal{C}}(x_{\mathcal{C}}) < \infty$ for all $x \in \mathcal{S}$. The sets of indices $\mathcal{C} \subseteq Q$ define sets of random variables $x_{\mathcal{C}}$, which in the MRF literature are termed *cliques*. A clique is any set that consists of random variables which are mutually neighbors. A set that consists of a single random variable (singleton) is also considered a clique.

Although a number of different neighborhood schemes exist [13], we will only be concerned with first order schemes, as the one depicted in Fig. 1. According to this scheme, the random variables are arranged on a rectangular lattice and each one of them depends on the values of its four nearest neighbors. As the lattice is not infinite, the r.v.'s at its edges will only have three neighbors, while the r.v.'s at the corners will only have two. The cliques in this spatial scheme consist only of singletons and pairs of neighbors. Therefore, only pairwise interactions are allowed between the random variables.

B. Gaussian Markov Random Fields

An example of a common MRF, which is a special case of the Chi MRFs that we introduce in this paper, and aids the clarification of the concepts discussed in the previous section, is the Gaussian MRF. Its conditional density function is

$$p(x_i | x_{\mathrm{n}(i)}) \propto \exp\left[-\frac{1}{\theta_i}\left(x_i - \sum_{j \in \mathrm{n}(i)} b_{ij} x_j\right)^2\right] \tag{2}$$

where θ_i is related to the variance of x_i and controls the scaling of the prior, while b_{ij} is a weight that determines the influence of x_j on x_i. The joint density function can be derived via the factorization

$$\frac{p(x)}{p(z)} = \prod_{i \in Q} \frac{p(x_i | x_1, \ldots, x_{i-1}, z_{i+1}, \ldots, z_q)}{p(z_i | x_1, \ldots, x_{i-1}, z_{i+1}, \ldots, z_q)} \tag{3}$$

where by x and z we denote two realizations of the vector of random variables X. Substituting the expression for the conditional density from (2) into the above factorization, and assuming the symmetry condition $b_{ij}/\theta_i = b_{ji}/\theta_j$, the derived expression for the joint density is

$$p(x) \propto \exp\left[-\left(\sum_{i \in Q} \frac{x_i^2}{\theta_i} - \sum_{\{i,j\} \in C} \frac{2b_{ij}}{\theta_i} x_i x_j\right)\right] \tag{4}$$

where C denotes the unordered set of pairs of indices, such that $\{i, j\} \in C$ if and only if x_i and x_j are neighbors. Note also that $x_{\mathcal{C}} = \{x_Q, x_C\}$, or in other words the cliques in a first-order neighborhood consist of the r.v.'s that form the MRF (singletons) and the pairs of r.v.'s which are mutually neighbors. A comparison with (1) reveals that (4) has the form required by the Hammersley–Clifford theorem.

C. Estimation With MRF Priors

Suppose that we observe a set of random variables $Y = [Y_1, \ldots, Y_q]$, which are modeled as a random function of the random variables X that constitute an MRF. An example of such a random function could be the addition of a Gaussian noise vector to X. We additionally suppose that the random variables Y are mutually independent when the values of X are known, while Y_i is independent of all X, except for X_i. For the joint density function of Y we therefore have

$$p(y) = \prod_{i \in Q} p(y_i | x_i). \tag{5}$$

A typical estimation problem under the above scenario is to find an optimal, in some sense, estimate of X when only Y is observed, given that the joint density of X belongs to the class of Markov Random Fields.

An estimator that has been widely used is the maximum *a posteriori* (MAP), which according to Bayes' theorem, can be written as [14]

$$\hat{x} = \arg\max_x p(y|x)p(x). \tag{6}$$

The above optimization problem can be difficult to solve due to the large number of the random variables involved in many real problems. In an image processing scenario for example, even a small picture (256×256) contains 2^{16} pixels. A relatively efficient, although still computationally demanding optimization method, was proposed by Geman and Geman [4] involving simulated annealing and the Gibbs sampler. Apart from the heavy computational load, an additional disadvantage of this type of global optimization is that it can induce correlations between random variables that are arbitrary far from each other [5], while it is generally desirable to have models whose dependencies are only local.

An alternative local, as opposed to global, optimization method was proposed by Besag [5], which was termed Iterated Conditional Modes (ICM). Under this estimation scheme, the proposed estimate $\hat{x}_i$ is the one with the maximum probability given the observation y_i and the neighbors $x_{\mathrm{n}(i)}$. That is

$$\hat{x}_i = \arg\max_{x_i} p(y_i | x_i) p(x_i | x_{\mathrm{n}(i)}). \tag{7}$$

The ICM method circumvents the problems posed by the computational load and the large scale dependencies of the global optimization methods. However, the ICM does not always converge to the global estimate of (6), which is the theoretically sound solution according to the MRF model specification. Furthermore, the ICM method does not require the strict adherence to genuine MRFs as predicted by the Hammersley–Clifford theorem [5]. Nevertheless, its computational efficiency and the avoidance of large scale dependencies make it a very attractive method, and for these reasons it is employed in the present study.

III. Statistical Model

A. Problem Formulation

The problem we consider in this work is the enhancement of speech that is corrupted by additive and uncorrelated noise. The enhancement of noisy speech is formulated as an estimation problem, according to which, an optimal estimate of the clean speech STFT amplitudes is sought, when only the STFT of the noisy speech is observed and a statistical model for the clean speech and noise is assumed. The linearity of the Fourier transform implies the following relationship for the ith sample of the STFT representations of the noisy speech, the speech and the noise signals, respectively

$$R_i \mathrm{e}^{\mathbf{i}\psi_i} = A_i \mathrm{e}^{\mathbf{i}\phi_i} + N_i \mathrm{e}^{\mathbf{i}\omega_i}. \tag{8}$$

In the above equation, R_i, A_i, and N_i are the STFT amplitudes of the noisy speech, the speech and the noise, ψ_i, ϕ_i, and ω_i are the respective phases and $\mathbf{i} \equiv \sqrt{-1}$.

For the noise STFT coefficients we use the standard modeling assumption (e.g., [2]), which is the zero mean complex Gaussian distribution with independent and identically distributed real and imaginary parts. We denote the second moment of the noise spectral amplitude $\mathrm{E}[N_i^2]$ by λ_{N_i}.

The model we assume for the speech phase is that it follows a uniform distribution, which is independent from the amplitude [15], [16] (i.e., $p(\phi_i) = 1/2\pi$, $p(A_i, \phi_i) = p(A_i)p(\phi_i)$). This model seems to agree very well with the speech data presented in [16] and has the additional advantage that the optimal estimate of the speech phase is the noisy phase itself [2], [15]. This implies that for the estimation of the clean speech STFT it suffices to estimate the amplitude only, which is then combined with the noisy phase to produce an estimate of the clean speech waveform. The statistical model that we are assuming for the speech spectral amplitude samples is detailed in the next section.

B. Speech Spectral Amplitude Prior

The STFT amplitude of speech signals has a rich structure across both time and frequency. Cohen [1] investigated the correlation of successive (in time) speech spectral amplitude samples using scatter plots and by calculating their correlation coefficient. For STFT frame overlap of 50% the correlation coefficient for STFT samples adjacent in time was 0.7 while for a 75% overlap it increased to 0.85. It is worth mentioning that the respective values for white noise signals were reported to be significantly lower, which implies that the correlation between consecutive STFT samples is a property of speech signals and is not solely due to the STFT frame overlap. Zavarehei *et al.* [3] claimed that the amplitudes of adjacent speech harmonics within the same time frame are highly correlated, reporting correlation coefficients in the region of 0.75–0.85. Additionally, samples which are adjacent in frequency within a STFT frame will also be correlated to a certain extent, due to the spectral leakage caused from the windowing operation involved in the calculation of the STFT.

A major contribution of this paper is the introduction of a speech spectral amplitude prior that is capable of taking into account the above dependencies that are present in the speech STFT amplitudes. The prior we propose, which we term Chi MRF prior, is given by

$$p(A_i|A_{\mathrm{n}(i)}) \propto A_i^{a-1} \exp\left[-\frac{1}{\theta_i}\left(A_i - \sum_{j\in \mathrm{n}(i)} b_{ij}A_j\right)^2\right]. \tag{9}$$

The above density is an extension of the Gaussian MRF conditional density, in the same manner that the Chi density [12] is a generalisation of the Gaussian density. Equation (9) yields the Gaussian MRF conditional density for $a = 1$. Since the parameter a in the case of the Chi density is called shape parameter, we maintain the same terminology here. The parameter θ_i controls the scaling of the density, while the parameters b_{ij} determine the influence between the neighbors i and j. The joint density can be found by considering the symmetry condition $b_{ij}/\theta_i = b_{ji}/\theta_j$ and by substituting the conditional density from (9) into the factorization in (3), which after some algebraic manipulation yields [17]

$$p(A) \propto \prod_{i\in Q}\left(A_i^{a-1}\right) \exp\left[-\sum_{i\in Q}\frac{A_i^2}{\theta_i} + \sum_{\{i,j\}\in C}\frac{2b_{ij}}{\theta_i}A_iA_j\right]. \tag{10}$$

Chi MRF priors are introduced in this work in order to sidestep the shortcomings of a Gaussian MRF based speech enhancement algorithm that was proposed in [11]. The latter resulted in musical residual noise, while the estimator was not well defined for all the values of its input parameters. The generalization provided by the Chi MRF priors mitigates both problems as it will be shown in Section V.

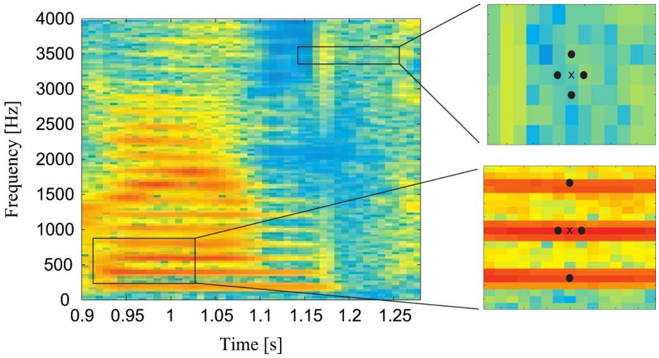


Fig. 2. Illustration of the proposed harmonic neighborhood. Upper right figure shows the neighbors of a sample that belongs to an unvoiced frame and lower right figure shows the neighbors used for the samples of the voiced frames.

Definition of the Neighborhood

The selection of the neighbors of the sample A_i determines its interaction with the rest of the spectral amplitude samples. To capture both the time and frequency dependencies of speech signals we propose a “harmonic” neighborhood, similar to the one found in [10]. In our scheme, the neighbors of each spectral amplitude sample in time, are the two adjacent samples within the same frequency bin. For the unvoiced or the speech absent frames, the frequency neighbors are the samples that are adjacent in frequency in the same time frame. For the voiced speech frames however, the frequency neighbors are k_{f0} frequency bins apart, where k_{f0} is the frequency bin number that corresponds to the pitch frequency of the current frame (assuming that the DC frequency bin has the number 0). The definition of the harmonic neighborhood is given in (11) and is illustrated in Fig. 2. See equation (11) at the bottom of the page.

As a shorthand notation for the neighbors of $A_i \equiv A(k,l)$ we introduce the notation $A_{\mathrm{n}(i)} = \{A_{\mathrm{S}}, A_{\mathrm{N}}, A_{\mathrm{W}}, A_{\mathrm{E}}\}$, which denote the south, north, west and east neighbors, respectively. If the frame l is unvoiced we denote $A_{\mathrm{S}} \equiv A(k-1,l)$ and $A_{\mathrm{N}} \equiv A(k+1,l)$, while if frame l is voiced $A_{\mathrm{S}} \equiv A(k-k_{f0},l)$ and $A_{\mathrm{N}} \equiv A(k+k_{f0},l)$. In both cases, $A_{\mathrm{W}} \equiv A(k,l-1)$ and $A_{\mathrm{E}} \equiv A(k,l+1)$.

The samples which lie on the edges of the STFT, and therefore have less than four neighbors, are treated by an appropriate modification of their weights, as discussed at the end of Section IV-B. Additionally, for the samples that belong to a voiced frame and correspond to frequencies smaller than the frame’s pitch frequency, the neighborhood of the unvoiced frames is used. That is because these samples typically have very low energy and this neighborhood avoids contamination

$$A_{\mathrm{n}(k,l)} = \begin{cases} \{A(k-1,l), A(k+1,l), A(k,l-1), A(k,l+1)\} & \text{if } l \text{ unvoiced} \\ \{A(k-k_{f0},l), A(k+k_{f0},l), A(k,l-1), A(k,l+1)\} & \text{if } l \text{ voiced} \end{cases} \tag{11}$$

from samples above the pitch frequency, which typically have higher energy levels.

The above type of neighborhood requires a pitch estimate for each of the voiced frames. The estimates are obtained with the pitch estimator of the 2400-bps Federal Standard Speech Coder [18]. This pitch estimation algorithm is based on autocorrelation and the application of error correcting procedures for common errors such as pitch doubling.

The proposed speech enhancement algorithm is robust against small errors in the pitch estimates. The reason is that only the frequency bin number that corresponds to the pitch frequency is required and not the actual pitch frequency. In our experiments we used analysis windows of 256 samples for calculating the STFT coefficients, while the sampling frequency was 8 KHz. This implies that each frequency bin has a bandwidth of 31.25 Hz; hence, pitch errors smaller than 15 Hz can be tolerated.

The voiced/unvoiced classification of the time frames was carried out with a voice activity detector (VAD), which was based on the average log-spectral difference between the noisy speech and the noise estimate for each time frame. We observed that applying the voiced frame neighborhood to an unvoiced or noise only time frame did not have a detrimental effect, because coupling frequency bins that were not adjacent still aided the recovery of broadband speech components and the uniform suppression of noise. On the other hand, using the unvoiced neighborhood for voiced frames did not allow the adequate recovery of the weaker speech harmonics.

In the light of the above observation, one could avoid the use of a VAD and treat all frames as voiced [i.e., discarding the first leg of (11)], using the pitch estimate of the last voiced frame for the unvoiced/speech absent frames. We prefer however to maintain this distinction, because a basic VAD is either computationally cheap and simple to implement or it comes with the pitch estimation algorithm at no extra computational cost, and in this way we avoid the oxymoron of requiring a pitch estimate for the unvoiced or speech absent frames.

IV. ACMRF Algorithm

A. Derivation of the Estimator

Applying the ICM method to the estimation of the speech spectral amplitudes yields

$$\hat{A}_i = \arg\max_{A_i} p(R_i|A_i)p(A_i|A_{\mathrm{n}(i)}). \tag{12}$$

Based on the Gaussian noise model, the expression for the likelihood in (12) is given by [2], [19]

$$p(R_i|A_i) = \frac{2R_i}{\lambda_{N_i}} \exp\left[-\frac{R_i^2 + A_i^2}{\lambda_{N_i}}\right] I_0\left(\frac{2R_iA_i}{\lambda_{N_i}}\right) \tag{13}$$

where $I_0(.)$ is the modified Bessel function of the first kind. The prior term in (12) is the Chi MRF prior given in (9). An analytical expression for the estimator in (12) can be found by maximizing the logarithm of $p(R_i|A_i)p(A_i|A_{\mathrm{n}(i)})$ term, w.r.t. A_i. This procedure is outlined in the following.

The modified Bessel function is first approximated by the formula [19]

$$I_0(x) = \frac{\mathrm{e}^x}{\sqrt{2\pi x}} \tag{14}$$

as this allows the derivation of the estimator in a closed form. Discarding the terms of (12) that are constant w.r.t. A_i, the expression that has to be maximized is

$$\hat{A}_i = \arg\max_{A_i}\left[\ln A_i^{a-3/2} - \frac{(A_i - R_i)^2}{\lambda_{N_i}} - \frac{\left(A_i - \sum_{j\in\mathrm{n}(i)} b_{ij}A_j\right)^2}{\theta_i}\right]. \tag{15}$$

Maximizing (15) leads to the following expression for the estimator:

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{16}$$

where

$$\zeta_1 = \frac{R_i\theta_i + \lambda_{N_i} \sum_{j\in\mathrm{n}(i)} b_{ij}A_j}{2(\theta_i + \lambda_{N_i})} \tag{17}$$

and

$$\zeta_2 = (1.5 - a)\frac{\lambda_{N_i}\theta_i}{2(\theta_i + \lambda_{N_i})}. \tag{18}$$

The following sections discuss the selection of the MRF prior parameters θ_i, b_{ij} and provide a method for the practical implementation of the algorithm.

B. MRF Parameter Selection

In image processing problems, fixed values for the parameters of the MRFs are often used (e.g., [5], [7]). For speech enhancement, the fixed parameters control the tradeoff between noise suppression and fidelity of the recovered speech [11], [17]. The reason being that the neighbors $A_{\mathrm{n}(i)}$ exert a constant influence on A_i, independent of the level of speech they contain. Ideally, one would like a neighbor to have a greater contribution when it contains significant speech information and smaller contribution when it contains mostly noise.

To implement such a policy we propose a set of adaptive MRF parameters, which are functions of the spectral variances of speech and noise. The proposed estimates for θ_i and b_{ij} are

$$\theta_i = \frac{w_{ii}\lambda_{A_i}\lambda_{N_i}}{\sum_{m\in\mathrm{n}(i)} w_{im}\lambda_{A_m} + \frac{\lambda_{N_i}a}{2}} \tag{19}$$

and

$$b_{ij} = \frac{w_{ij}\sqrt{R_i^2\lambda_{A_j}}}{\sum_{m\in\mathrm{n}(i)} w_{im}\lambda_{A_m} + \frac{\lambda_{N_i}a}{2}}. \tag{20}$$

In the above equations, $\lambda_{A_i} \equiv \mathrm{E}[A_i^2]$ and $\lambda_{N_i} \equiv \mathrm{E}[N_i^2]$ are the variances of the ith spectral components of speech and noise, respectively. The constants w_{ij} are weights that provide further control to the interaction between neighbors. An interpretation of the above equations and their function within the MRF prior is given in Appendix I. Briefly, note that θ_i (or b_{ij}) increases when the variance of A_i (or A_j) is greater than the variances of noise and the other neighbors and vice versa.

C. Implementation

The estimation proceeds from smaller to larger frequency indices k and subsequently from smaller to larger time frame indices l. According to this schedule, during the estimation of A_i there are estimates available for A_{S} and A_{W}, but not for A_{N} and A_{E}. For the two latter quantities we need to calculate temporary estimates. The temporary estimate for A_{N} is found using a spectral subtraction type estimate

$$\hat{A}_{\mathrm{N}} = \left(R_{\mathrm{N}}^2 - \lambda_{N_{\mathrm{N}}}\right)^{0.5} \tag{21}$$

and a similar formula is used for the estimation of A_{E}. For the variances of the neighbors we propose the estimates

$$\hat{\lambda}_{A_j} = \hat{A}_j^2, \quad j \in \mathrm{n}(i) \tag{22}$$

while for the variance of the sample A_i we use

$$\hat{\lambda}_{A_i} = R_i^2 - \lambda_{N_i}. \tag{23}$$

The above strategy for the estimation of the parameters that are involved in (19), (20) allows us to write the ACMRF estimator in a very compact form and provides some further insight on its behavior. We begin by noting that according to (22), $\sqrt{R_i^2\hat{\lambda}_{A_j}}\hat{A}_j = R_i\hat{\lambda}_{A_j}$. Using this last result, substitution of the expressions for θ_i and b_{ij} (19), (20) in the equation for ζ_1(17) yields, after some algebraic manipulation

$$\zeta_1 = \frac{R_i \sum\limits_{m\in\{i,\mathrm{n}(i)\}} w_{im}\hat{\lambda}_{A_m}}{2\left(\sum\limits_{m\in\{i,\mathrm{n}(i)\}} w_{im}\hat{\lambda}_{A_m} + \frac{\lambda_{N_i}a}{2}\right)}. \tag{24}$$

If we finally define an *a priori* SNR estimate ξ_i as

$$\xi_i \equiv \sum_{m\in\{i,\mathrm{n}(i)\}} w_{im}\frac{\hat{\lambda}_{A_m}}{\lambda_{N_i}} \tag{25}$$

then ζ_1 can be further simplified to

$$\zeta_1 = \frac{\xi_i R_i}{2\left(\xi_i + \frac{a}{2}\right)}. \tag{26}$$

Following the same procedure, ζ_2 can be reduced to

$$\zeta_2 = (1.5 - a)\frac{w_{ii}\hat{\lambda}_{A_i}}{\xi_i + \frac{a}{2}}. \tag{27}$$

The ACMRF estimator can therefore be summarized as

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{28}$$

where ζ_1 and ζ_2 are given by (26), (27).

TABLE I
PSEUDOCODE FOR THE ACMRF ALGORITHM

```
For all time frames $l$
  For all frequency bins $k$
    Find $A_{\mathrm{N}}$, $A_{\mathrm{E}}$ with (21)
    Find $\lambda_{A_{\mathrm{n}(i)}}$ with (22)
    Find $\lambda_{A_i}$ with (23)
    Find $\xi_i$ with (25)
    Find $\hat{A}_i$ with (26), (27), (28)
```

The above form of the estimator is similar to the MAP estimator that uses the Chi speech prior, presented in [20] and is a generalization of the MAP spectral amplitude estimator proposed by Wolfe and Godsill [21]. A major difference of the MRF-based estimator however, is that for the estimation of the *a priori* SNR ξ_i the variances of A_i and a number of neighbors $A_{\mathrm{n}(i)}$ are taken into account, while in the traditional Decision Directed approach [2] for example, only the variances of $A(k,l)$ and $A(k,l-1)$ are considered.

The w_{ij} weights are empirically selected as $w_{i\mathrm{W}} = 0.49$, $w_{i\mathrm{S}} = 0.48$, $w_{i\mathrm{E}} = 0.01$, $w_{i\mathrm{N}} = 0.01$, and $w_{ii} = 0.01$. The $w_{i\mathrm{W}}$ and $w_{i\mathrm{S}}$ weights are significantly larger because during the estimation of A_i, the A_{W} and A_{S} have already been estimated with the ACMRF algorithm and therefore the estimates are more accurate. For the samples that lie on the DC and Nyquist frequency bins we use $w_{i\mathrm{W}} = 0.98$, $w_{i\mathrm{S}} = 0.00$, $w_{i\mathrm{E}} = 0.01$, $w_{i\mathrm{N}} = 0.00$ and $w_{ii} = 0.01$, while for voiced frame samples that are less than a pitch frequency apart from the Nyquist frequency, w_{N} is set to zero.

Some final implementation details include limiting the variances (22), (23) to non-negative numbers and forcing ξ_i to be greater than -25 dB for perceptual reasons. Also for $a < 1.5$, when $\zeta_1^2 - \zeta_2 < 0$ (28) cannot be applied and the noisy samples are suppressed by a fixed amount (50 dB). The ACMRF algorithm is summarized in Table I.

V. Evaluation

The proposed algorithm was evaluated with a series of objective and subjective tests. For the objective evaluation we used 48 sentences from the TIMIT database, sampled at 8 KHz and uttered by three male and three female speakers. The speech was corrupted with four different types of noise, computer generated white Gaussian, train and car noise recorded by our colleagues, and babble noise from the Noizeus database [22]. The objective measures we used in the evaluation were the Segmental SNR (SegSNR) [23], the Log Spectral Distortion (LSD) [1], and the PESQ. The last measure is the ITU-T P.862 recommendation [24], [25], and is an objective measure that is designed to predict the results of Mean Opinion Score tests (MOS). The PESQ scores lie on a scale from 1 (= bad) to 4.5 (= no distortion) and are reported to correlate well with subjective MOS results [26].

The subjective evaluation consisted of a listening test (A-B preference test) in which the ACMRF algorithm was compared with the Log STSA (LS) [27]. In this test, ten sentences, uttered by five male and five female speakers, were corrupted

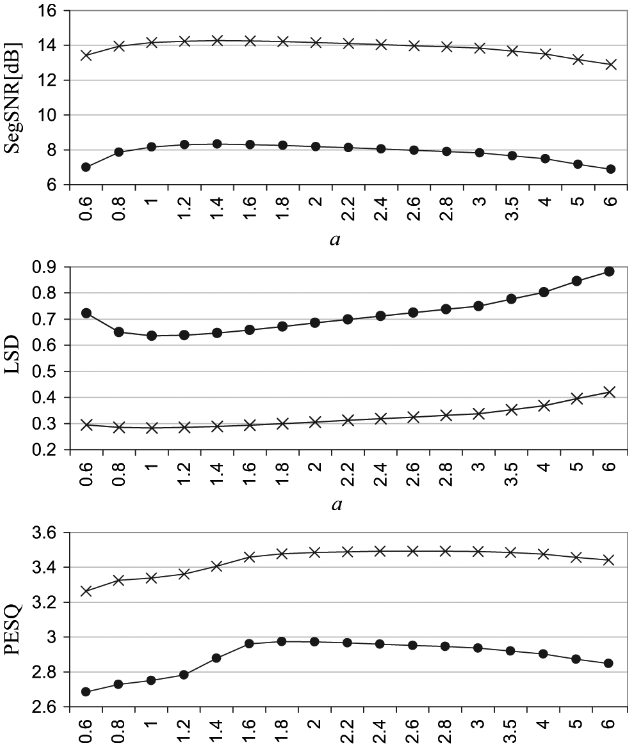


Fig. 3. Objective measures scores as a function of the shape parameter a. The corrupting noise is white Gaussian at 0 (circles) and 10 (crosses) dB input SegSNR.

with the noises used for the objective evaluation at two different SNR levels. Ten normal hearing listeners, age 20–30 years, were asked to identify the algorithm that provided the best quality of enhanced speech. The sentences were presented to each listener in a random order and the order of presenting the two enhanced versions of the same sentence was also randomized.

For both evaluations the transformation to the STFT domain was performed with Hamming windows of 256 samples, using an overlap of 75%. In order to isolate the effect of the performance of a practical noise estimation algorithm (e.g., [28]), the noise power was estimated from the noise samples directly with a first-order recursive system as in [1]. The pitch estimates needed for the definition of the neighborhood were obtained with the pitch estimator described in [18] directly from the noisy speech signals. The VAD had a high sensitivity, because as explained in Section III-C, treating a voiced frame as unvoiced (false negative) hindered its enhancement, while the opposite (false positive) had a rather negligible effect.

We begin our evaluation by examining the effect of the shape parameter a on the quality of speech.

A. *Effect of the Shape Parameter a*

Fig. 3 shows the results in the three objective measures as a function of the value of a, for white noise and two input SegSNR levels. The respective curves for other noise types and input SegSNR levels have a similar shape to those shown here. Further examples can be found in [17].

The objective measures indicate that the optimal range of the shape parameter is $1.8 < a < 2.8$, while their values do not vary significantly when a lies in the above interval. Informal listening tests and observation of spectrograms indicate that these values of a result in a good preservation of speech spectral components in combination with uniform residual noise. For $a < 1.5$, spurious spectral peaks begin to appear, which are perceived as musical noise and/or speech distortion. Finally, for a larger than 3 the preservation of the speech spectral components gradually deteriorates, which is reflected in the drop of the objective measures scores. Based on the above analysis we use for the remainder of this evaluation $a = 2$.

B. *Comparison With Alternative Algorithms*

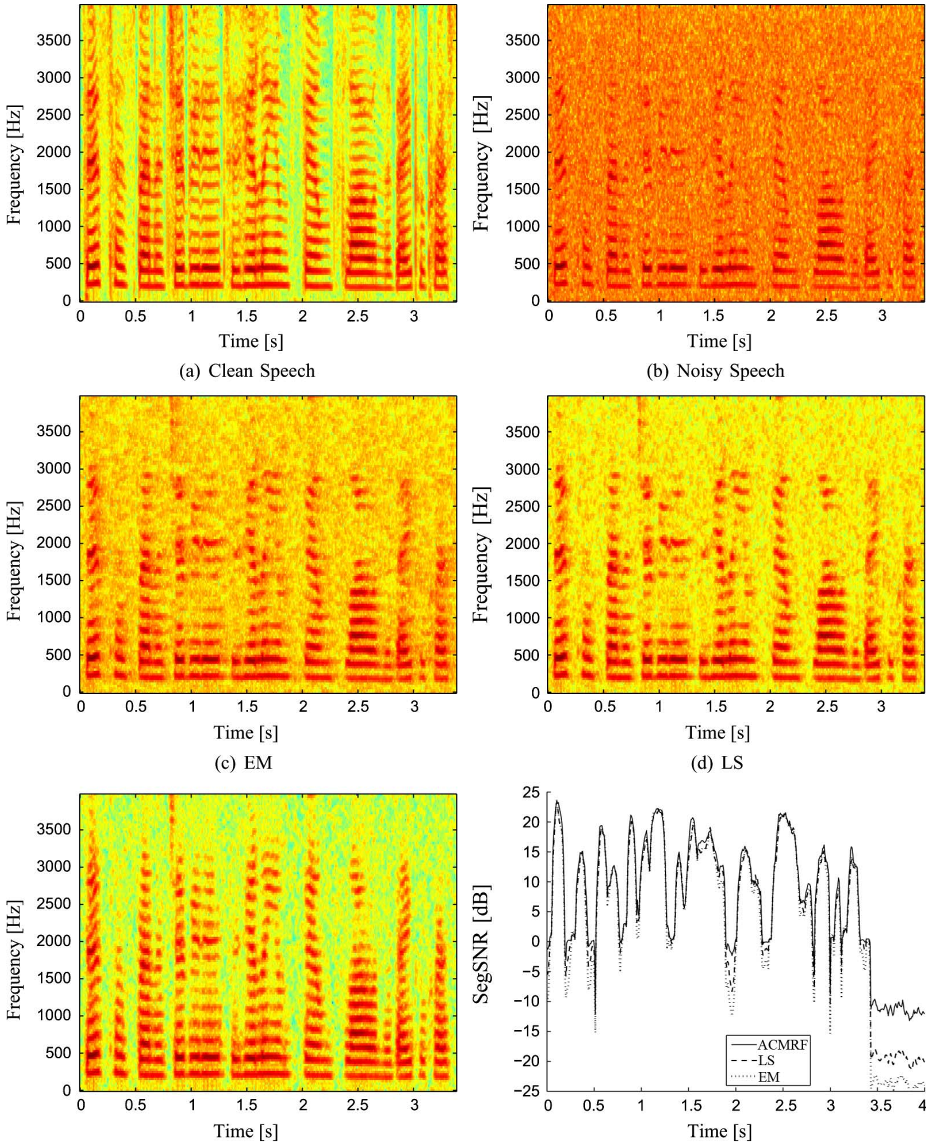
This section compares the proposed algorithm with two well established alternatives, the Ephraim–Malah MMSE STSA (EM) [2] and the Log STSA (LS) [27].

Fig. 4 shows spectrograms of the sentence "*This has been attributed to the helium film flow in the vapour pressure thermometer*" corrupted with white Gaussian noise at 0 dB input SegSNR and enhanced with the three algorithms. Fig. 4(f) shows the frame SNRs for the same sentence that is shown in the spectrograms, augmented by half a second of silence, in order to highlight the noise suppression ability of the ACMRF algorithm.

A comparison of Figs. 4(c)–4(e) shows that the ACMRF algorithm restores a large number of speech spectral components, which are missed by the two alternative algorithms and are unidentifiable even by a visual inspection of the noisy speech spectrogram in Fig. 4(b). The recovery of these spectral components is attributed to a large extent to the frequency coupling that is imposed by the MRF prior. Additionally, in noise dominated regions, the smoothing that is achieved by considering the values of the four nearest neighbors, results in a uniform noise of much lower level. Inspection of the last time frames in Fig. 4(f) reveals that the ACMRF algorithm results in 8 and 12 dB more suppression of the residual noise compared to the LS and EM algorithms, respectively, in the silent segments of the utterance.

Table III shows the SegSNR scores of the three algorithms for the different noises and input SegSNR levels. The ACMRF algorithm results consistently in higher scores compared to the two alternatives. For example the improvements over the EM algorithm are as high as 3 dB in low input SegSNR conditions. Tables IV and V show the respective LSD and PESQ results, which in all cases favor the ACMRF algorithm.

Table II shows the results of the subjective listening test. The results indicate that the listeners preferred the ACMRF algorithm from 76% to 91% of the time, depending on the background noise and input SegSNR level. It was constantly reported that the ACMRF algorithm resulted in lower levels of residual background noise. For the low SNR condition, the higher noise suppression capability of the ACMRF algorithm made some artefacts of the noise suppression process more perceivable. We attribute the lower scores for the low SNR condition to this characteristic. An exception to this observation was the car noise. Because most of its energy is concentrated in the lower frequency bands, the processing artefacts were masked by the speech energy and the listeners overwhelmingly chose the ACMRF algorithm for the low SNR. For the higher



(a) Clean Speech

(b) Noisy Speech

(c) EM

(d) LS

(e) ACMRF

(f) Frame SNR

Fig. 4. Spectrograms of clean speech, speech corrupted with Gaussian white noise at 0-dB input SegSNR and enhanced with the EM, LS, and ACMRF algorithms. (f) shows the frame SNR of the above utterance enhanced with the three algorithms, augmented with half a second of silence.

TABLE II
PERCENT PREFERENCE OF THE ACMRF ALGORITHM OVER THE LS FOR DIFFERENT NOISE TYPES AND INPUT SEGSNR LEVELS

Noise	% ACMRF pref.	Noise	% ACMRF pref.
White, 0 dB	76 %	Car, 0 dB	91 %
White, 10 dB	86 %	Car, 10 dB	77 %
Babble, 0 dB	82 %	Train, 0 dB	80 %
Babble, 10 dB	87 %	Train, 10 dB	90 %

SNR condition, the listeners reported that for some of the sentences the two algorithms produced very similar results, so their preference for the ACMRF was not as strong. In conclusion, the proposed algorithm was chosen at least 3 times out of 4, across all noise types and levels.

Regarding the computational efficiency, unlike the EM and LS algorithms, the ACMRF requires only the evaluation of elementary operations. It also requires a VAD and a pitch estimator, which are additional to most speech enhancement schemes. Excluding the cost of the latter two modules, the ACMRF has comparable computational requirements to the efficient algorithms proposed in [21].

VI. CONCLUSION

In this paper, we presented a speech enhancement algorithm that models the time and frequency dependencies of the speech STFT amplitudes. The modeling of the above dependencies was pursued using concepts from the theory of Markov Random Fields. A novel model, the Chi MRF, was introduced and it was shown that it constitutes a generalization of the established Gaussian MRF. The conditional Chi MRF density was then employed as a speech prior for the development of a MAP speech spectral amplitude estimator.

The proposed prior was combined with a “harmonic” neighborhood, in which the four nearest neighbors of each sample

TABLE III
SegSNR Results Obtained With the EM, LS, and ACMRF Algorithms for Different Noise Types and Input SegSNR Levels

	White			Car			Train			Babble		
	EM	LS	ACMRF	EM	LS	ACMRF	EM	LS	ACMRF	EM	LS	ACMRF
-5	3.26	4.23	5.52	5.59	6.80	8.24	3.08	4.11	5.44	1.86	3.24	4.86
0	6.41	7.05	8.21	9.27	10.21	11.37	6.39	7.14	8.25	5.26	6.31	7.80
5	9.65	10.00	11.07	12.88	13.54	14.55	9.87	10.38	11.34	8.80	9.54	10.87
10	13.08	13.18	14.19	16.51	16.89	17.80	13.53	13.83	14.67	12.54	12.98	14.20

TABLE IV
LSD Results Obtained With the EM, LS, and ACMRF Algorithms for Different Noise Types and Input SegSNR Levels

	White			Car			Train			Babble		
	EM	LS	ACMRF	EM	LS	ACMRF	EM	LS	ACMRF	EM	LS	ACMRF
-5	1.01	1.01	0.93	0.73	0.67	0.59	0.95	0.91	0.82	1.06	0.98	0.84
0	0.75	0.78	0.68	0.49	0.46	0.40	0.67	0.67	0.60	0.76	0.72	0.58
5	0.53	0.56	0.47	0.32	0.32	0.27	0.46	0.47	0.41	0.51	0.50	0.40
10	0.35	0.37	0.31	0.21	0.22	0.18	0.30	0.31	0.27	0.33	0.33	0.26

TABLE V
PESQ Results Obtained With the EM, LS, and ACMRF Algorithms for Different Noise Types and Input SegSNR Levels

	White			Car			Train			Babble		
	EM	LS	ACMRF	EM	LS	ACMRF	EM	LS	ACMRF	EM	LS	ACMRF
-5	2.45	2.55	2.69	3.09	3.20	3.29	2.68	2.75	2.83	2.49	2.58	2.62
0	2.78	2.87	2.98	3.34	3.45	3.54	2.95	3.03	3.13	2.79	2.89	2.98
5	3.05	3.14	3.24	3.57	3.67	3.76	3.21	3.28	3.38	3.06	3.16	3.27
10	3.30	3.39	3.50	3.79	3.86	3.95	3.43	3.51	3.61	3.32	3.41	3.52

were considered in the unvoiced speech or speech absent frames, while the samples that were one pitch frequency apart were used for the frames that contained voiced speech. The pitch estimation was performed with an "off-the-shelf" pitch estimator, which estimated the pitch directly from the noisy samples.

An important aspect of the proposed algorithm was the adaptive selection of the weights that determined the influence of the neighbors on the estimated sample. These were selected in such a way that samples with large variance, which were more likely to contain speech, exerted more influence on the estimated sample, compared to samples with smaller variance, which were more likely to correspond to noise. This strategy allowed the recovery of weak speech spectral components, while keeping the residual noise level low.

The comparison of the proposed algorithm with other widely used speech enhancement schemes, highlighted its ability to recover speech spectral components that are immersed in noise. This attribute of the proposed algorithm is largely due to the time-frequency coupling that the MRF priors imposed. Additionally, in noise dominated areas, the consideration of a larger number of neighboring samples compared to more traditional approaches, resulted in lower levels of residual noise, which was also free of musical noise artefacts. The above two characteristics of the proposed algorithm were also illustrated in a number of objective and subjective speech quality tests. Finally, the derivation of the proposed estimator in a simple, closed form, allows the easy implementation of the ACMRF algorithm, and keeps its computational complexity low.

Appendix
Interpretation of the θ_i and b_{ij} Parameters

The Chi MRF prior (9) can be written as

$$p(A_i|A_{\mathrm{n}(i)}) \propto A_i^{a-1}\exp\left[-\frac{1}{\theta_i}\left(A_i-\sum_{j\in\mathrm{n}(i)} b'_{ij}\right)^2\right] \qquad (29)$$

with θ_i as defined in (19) and

$$b'_{ij} = \frac{w_{ij}\lambda_{A_j}R_i}{\sum\limits_{m\in\mathrm{n}(i)} w_{im}\lambda_{A_m} + \frac{\lambda_{N_i}a}{2}}. \qquad (30)$$

When the variances of the neighbors $\lambda_{A_{\mathrm{n}(i)}}$ tend to zero, because $A_{\mathrm{n}(i)}$ contain mostly noise, then $\sum_{j\in\mathrm{n}(i)} b'_{ij}$ also tends to zero and the prior tends to be proportional to the Chi density, which is given by $p(A_i) = 2A_i^{a-1}/(\theta_i^{a/2}\Gamma(a/2))\exp(-A_i^2/\theta_i)$. The parameter θ_i, which is the scale parameter of the Chi density, tends to $2w_{ii}\lambda_{A_i}/a$, which is in accordance with the expression for the second moment of the Chi density, which is $\mathrm{E}[A_i^2] \equiv \theta_i a/2$. In other words, when the neighbors $A_{\mathrm{n}(i)}$ contain mostly noise, the prior degenerates to a univariate Chi density with variance $w_{ii}\lambda_{A_i}$; hence, the similarities of the ACMRF algorithm with the MAP-Chi estimator presented in [20].

On the other hand, when the variance of the neighbors is much higher than λ_{A_i} and λ_{N_i}, then θ_i tends to zero, and $\sum_{j\in\mathrm{n}(i)} b'_{ij}$ tends to R_i. This implies that the prior degenerates to a point mass, which is centered at R_i. The result is that when the neighbors of A_i contain mainly speech, then the prior strongly favors the observation R_i as an estimate for A_i.

In preliminary work, we experimented with the parameter

$$b_{ij}^{\mathrm{alt}} = \frac{w_{ij}\lambda_{A_j}A_j}{\sum\limits_{m\in \mathrm{n}(i)} w_{im}\lambda_{A_m} + \frac{\lambda_{N_i}a}{2}}. \tag{31}$$

Under this scenario, when λ_{A_j} is much greater than the rest of the variances, then the prior tends to a point mass centred at A_j and effectively we have $\hat{A}_i \approx A_j$. In an image processing scenario, where MRFs have been extensively used, assigning to a pixel the value of its neighbor might be desirable, assuming that both pixels represent the same color. For the restoration of the speech STFT amplitudes however, this approach was found to generate significant distortions and was abandoned.

Acknowledgment

The authors would like to thank S. Vaseghi, E. Zavarehei, Q. Yan, B. Milner, and J. Darch for their helpful comments and discussion. They would also like to thank the listeners who participated in the listening tests, as well as to the reviewers for their constructive comments.

References

[1] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.

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

[3] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1194–1203, May 2007.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, pp. 721–741, Nov. 1984.

[5] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Statist. Soc., Ser. B*, no. 48, pp. 259–302, 1986.

[6] T. P. O'Rourke and R. L. Stevenson, "Improved image decompression for reduced transform coding artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 490–499, Dec. 1995.

[7] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 296–310, Jul. 1993.

[8] P. Pèrez, "Markov random fields and images," *CWI Quarterly*, vol. 11, no. 4, pp. 413–437, 1998.

[9] G. Gravier, M. Sigelle, and G. Chollet, "A Markov random field based multi-band model," in *Proc. 25th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-00*, 2000, vol. 3, pp. 1619–1622.

[10] B. I. Andia, "Restoration of speech signals contaminated by stationary tones using an image perspective," in *Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, 2006, vol. 3, pp. 61–64.

[11] I. Andrianakis and P. R. White, "On the application of Markov random fields to speech enhancement," in *Proc. IMA Int. Conf. Math. Signal Process.*, 2006, pp. 198–201.

[12] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. New York: Wiley, 1994, vol. 1.

[13] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Statist. Soc., Ser. B*, no. 36, pp. 192–236, 1974.

[14] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*. New York: Wiley, 1968.

[15] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. 9th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-84*, 1984, vol. 9, pp. 53–56.

[16] R. C. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized-gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[17] I. Andrianakis, "Bayesian estimators for speech enhancement," Ph.D. dissertation, ISVR, Univ. of Southampton, , U.K., 2007.

[18] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: The new Federal standard at 2400 bps," in *Proc. 22nd IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-97*, 1997, vol. 2, pp. 1591–1594.

[19] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[20] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," in *Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, 2006, vol. 3, pp. 1068–1071.

[21] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process.*, vol. 10, pp. 1043–1051, 2003.

[22] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, 2006, vol. 1, pp. 153–156.

[23] J. R. J. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[24] *Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T Rec. P. 862 Std, Feb. 2001.

[25] 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," in *Proc. 26th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-01*, 2001, vol. 2, pp. 749–752.

[26] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," *Proc. Interspeech 2006*, 2006.

[27] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Proc. IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[28] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

Yiannis Andrianakis graduated from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and received the M.Sc. degree in sound and vibration studies and the Ph.D. degree in the development of Bayesian estimators for speech enhancement, both from the Institute of Sound and Vibration Research, University of Southampton, U.K., in 2004 and 2007, respectively.

He is currently with the National Oceanography Center, Southampton, U.K., where he is working on a project titled "Managing Uncertainty in Complex Models."

Paul R. White (M'01) received the Ph.D. degree from the Institute of Sound and Vibration Research, University of Southampton, Southampton, U.K., in 1992.

He is currently a Professor of statistical signal processing at the University of Southampton. His research interests are centered on digital signal processing (DSP). In particular, he is interested in the application of DSP to problems in underwater acoustics, speech and image processing, condition monitoring, and biomedicine. This work involves the development of novel DSP methods for the analysis of nonlinear or time-varying systems, statistical modeling, and adaptive algorithms.