A New Metric for VQ-based Speech Enhancement and Separation

Christensen, Mads Græsbøll; Mowlaee, Pejman

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2011.5947420

Publication date:
2011

Document Version
Accepted author manuscript, peer reviewed version

A NEW METRIC FOR VQ-BASED SPEECH ENHANCEMENT AND SEPARATION

Mads Græsbøll Christensen

Dept. of Arch., Design and Media Technology
Aalborg University, Denmark
mgc@create.aau.dk

Pejman Mowlaee

Institut für Kommunikationsakustik
Ruher-Universität Bochum, Germany
pejman.mowlaee@rub.de

ABSTRACT

Speech enhancement and separation algorithms frequently employ two-stage processing schemes, where the signal is first mapped to an intermediate low-dimensional parametric description. Then, these parameters are mapped to vectors in codebooks trained on individual noise-free sources using a vector quantizer. To obtain accurate parameters, one must employ an estimator that takes the signal characteristics into account. An open question is, however, how to derive metrics for use in the vector quantization process. In this paper, we present and derive a new metric aimed at exactly this, and we exemplify and demonstrate its use in sinusoidal modeling. The metric takes into account that parameters may have different uncertainties and dependencies associated with them and thus leads to more accurate estimates, as is demonstrated in experiments. Moreover, we incorporate the metric in a recently proposed speech separation algorithm and compare its performance to state-of-the-art methods.

Index Terms— Speech processing, vector quantization, speech enhancement

1. INTRODUCTION

Speech separation and enhancement algorithms play important roles in many speech processing applications as subsequent processing stages and models can be greatly simplified if they do not have to take into account the presence of noise and multiple sources. Examples of applications that benefit from this are speech recognition and speech coding, both of which can exclude models of background noise and interfering speakers when only the signal of interest is present. There are of course also other applications where such algorithms may be useful, including hearing aids, in which they alleviate listener fatigue and hold the promise of increased speech intelligibility for the hearing impaired. A common approach to speech separation and enhancement is vector quantization (VQ), where codebooks are trained offline for each noise-free source. These codebooks are then used for estimating the individual speech sources from a mixture, or the speech signal from a noisy observation, as is the case in speech enhancement. Instead of the time-domain signal, often low-dimensional parameter vectors are used as an intermediate representation of the sources as this leads to better and faster training of the codebooks and faster separation and enhancement algorithms. Some examples of VQ-based enhancement and separation methods are [1–4] and [5–9], respectively. In finding the parameters of the intermediate representation, standard estimation algorithms such as the maximum likelihood estimator based on well-known metrics can be used. However, the question then arises which metric to use in the vector quantization process.

In this paper, we seek to answer this question as we develop a new metric for this purpose based on statistical arguments and exemplify its use in a specific form of speech processing, more specifically sinusoidal modeling. The effect of the metric on the vector quantization process is twofold: Firstly, it takes into account that the estimates of different parameters may have different uncertainties associated with them, and, secondly, it also takes into account that there may exist dependencies between the parameter estimates.

The rest of the paper is organized as follows. In Section 2, we define the considered problem and present and derive the proposed metric based on statistical arguments. In the next section, Section 3, we exemplify the use of the derived metric on a sinusoidal model and present simulation results in Section 4. Finally, we conclude on our work in Section 5.

2. THEORETICAL DEVELOPMENT

We will now proceed to derive the proposed metric, but first we will define the problem under consideration, and we will do this based on the following signal model:

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{s}_k + \mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ is the observed signal, $\mathbf{e}$ the observation noise, and $\mathbf{s}_k$ the kth signal of interest. Each signal of interest $\mathbf{s}_k$ is characterized by (possibly nonlinear) parameters $\boldsymbol{\theta}_k$. Note that, for simplicity, $\boldsymbol{\theta}_k$ denotes both the true parameter vector and the unknown parameter vector, depending on the context. When we refer to a specific estimate, this will be denoted as $\hat{\boldsymbol{\theta}}_k$. The full parameter set is denoted $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$ and similarly for estimates. The problem of interest is then to find estimates $\{\hat{\boldsymbol{\theta}}_k\}$ of $\{\boldsymbol{\theta}_k\}$ from $\mathbf{x}$ where the parameters are in a codebook, i.e., $\boldsymbol{\theta} \in \mathcal{C}$, and this codebook is a subset of the full space, i.e., $\mathcal{C} \subset \mathbb{R}^M$.

Some VQ-based speech separation and enhancement algorithms work in a way, where, instead of finding directly the codebook entries that best match the observation in some sense, they first go through an intermediate step wherein a parametrization of the signal is obtain. In math, this can be described as

$$f : \mathbb{R}^N \rightarrow \mathbb{R}^M, \tag{2}$$

where $f(\cdot)$ is then the estimator. Using this estimator, intermediate parameters $\tilde{\boldsymbol{\theta}}$ are found as $\tilde{\boldsymbol{\theta}} = f(\mathbf{x})$. This is often beneficial as the dimension of the parameter vector will be lower (and often much lower) than the observation vector, i.e., $M < N$, whereby not only the training procedure but also the separation or enhancement algorithm are simplified. These intermediate parameters are then

mapped to codebook entries via a vector quantizer, here a function $g(\cdot)$, defined as

$$g : \mathbb{R}^M \to \mathcal{C}. \tag{3}$$

The final estimates are then obtained as $\hat{\boldsymbol{\theta}} = g(\tilde{\boldsymbol{\theta}})$. The question to be answered is then how the functions $f(\cdot)$ and $g(\cdot)$ relate and how they should be chosen. An estimator of the intermediate parameters $\tilde{\boldsymbol{\theta}}$ should be chosen such that the found parameters are most likely to explain the observation, i.e., it should take the characteristics of the noise $\mathbf{e}$ into account. An obvious choice here that does this is the maximum likelihood estimator, which is well-known to exhibit a number of desirable properties, including asymptotic optimality.

Assuming that a maximum likelihood estimator $f(\cdot)$ is used and that the data satisfies some regularity conditions, the so-obtained estimates $\tilde{\boldsymbol{\theta}}$ are asymptotically distributed as (see, e.g., [10])

$$\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})) \tag{4}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix, $\sim$ means distributed according to and $\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$ denotes the normal probability density function (pdf) with mean $\boldsymbol{\theta}$ and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\theta})$. Specifically, the Fisher information matrix is defined as

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}\left\{ \frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}, \tag{5}$$

where $p(\mathbf{x}; \tilde{\boldsymbol{\theta}})$ is the likelihood function of the observed signal parametrized by the parameters $\boldsymbol{\theta}$. The above then also means that the estimation error $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is distributed as $\mathcal{N}(0, \mathbf{I}^{-1}(\boldsymbol{\theta}))$, i.e., the estimates are asymptotically unbiased and attain the Cramér-Rao lower bound. It then follows that the likelihood function for the intermediate parameters is given by

$$p(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{M}{2}} \det\left(\mathbf{I}^{-1}(\boldsymbol{\theta})\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T \mathbf{I}(\boldsymbol{\theta})\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)}, \tag{6}$$

where the pdf can be seen to be parametrized by the unknown parameters $\boldsymbol{\theta}$. Choosing now as our vector quantization function $g(\cdot)$ the maximum likelihood estimator, we obtain

$$\hat{\boldsymbol{\theta}} = g(\tilde{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\theta} \in \mathcal{C}} \ln p(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}) \tag{7}$$

$$= \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{I}(\boldsymbol{\theta}) \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{8}$$

This criterion, which is a weighted squared error metric, essentially takes into account that different parameters in $\boldsymbol{\theta}$ may have different uncertainties associated with them in the vector quantization process. As can be seen, the resulting estimator is a weighted least-squares estimator.

One last difficulty remains, however. The metric in the estimator in (8) requires knowledge of the true parameters to compute $\mathbf{I}(\boldsymbol{\theta})$. Instead of using $\mathbf{I}(\boldsymbol{\theta})$, we can use an approximation based on the intermediate parameters $\tilde{\boldsymbol{\theta}}$ (see [11, 12]), i.e.,

$$\mathbf{I}(\boldsymbol{\theta}) \approx -\mathrm{E}\left\{ \frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}\Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \triangleq \mathbf{W}, \tag{9}$$

and this leads to the following estimates:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{W} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} J, \tag{10}$$

which is the fundamental result that we will use here. The metric essentially takes into account that the individual intermediate parameters will have different uncertainties associated with them in the vector quantization process. It should be noted that the obtained weighting matrix may also be valid for suboptimal estimators that produce estimates that are not distributed according to (4) as long as the covariance matrix is related to the inverse Fisher information matrix as $\kappa \mathbf{I}^{-1}(\boldsymbol{\theta})$, where κ is a positive constant. We note that the metric in (10) is also the metric one obtains using the so-called EXIP principle [13].

We note that for the fairly general case of Gaussian signals with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q})$ where $\mathbf{Q}$ is the noise covariance matrix, Slepian-Bang's formula can be used for determining a more specific expression for the Fisher information matrix. More specifically, it is given by

$$\left[\mathbf{I}(\boldsymbol{\theta})\right]_{nm} = \frac{\partial \boldsymbol{\mu}^T(\boldsymbol{\theta})}{\partial \theta_n} \mathbf{Q}^{-1} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_m}, \tag{11}$$

which requires only that the partial derivatives of the mean with respect to all unknown parameters $\frac{\partial \boldsymbol{\mu}^T(\boldsymbol{\theta})}{\partial \theta_n}$ for all n be determined, something that is often fairly simple to do.

3. AN EXAMPLE

We will now exemplify the use of the proposed metric with a specific parametrization of the observed signal $\mathbf{x}$. More specifically, we will use a sinusoidal model that is characterized by frequencies $\{\omega_l\}$, amplitudes $\{A_l\}$, and phases $\{\phi_l\}$. In this case, the signal model is given by

$$\mathbf{x} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{12}$$

where $\mathbf{Z} \in \mathbb{C}^{N \times L}$ is a Vandermonde matrix constructed from L complex sinusoidal vectors as $\mathbf{Z} = [\, \mathbf{z}(\omega_1) \cdots \mathbf{z}(\omega_L) \,]$ with $\mathbf{z}(\omega) = [\, 1 \;\; e^{j\omega} \;\; \cdots \;\; e^{j\omega(N-1)} \,]^T$, and $\mathbf{a} \in \mathbb{C}^L$ a vector containing the complex amplitudes as $\mathbf{a} = [\, a_1 \;\; \cdots \;\; a_L \,]^T$ where $a_l = A_l e^{j\phi_l}$. Note that, as before, we assume that we are here dealing with real signals, which means that the complex sinusoid come in complex-conjugate pairs. The parameter vector for each sinusoid is defined as $\boldsymbol{\theta}_l = [\, A_l \; \phi_l \; \omega_l \,]$. Assuming that the noise $\mathbf{e}$ is white Gaussian, the Fisher information matrix is fortunately well-known (see, e.g., [10]). For sufficiently large N and a distinct set of frequencies, it exhibits the following block-diagonal structure:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{W}_L \end{bmatrix}, \tag{13}$$

which means that the associated metric is additive over the sub-matrices, which yields the estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \sum_{l=1}^{L} \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right)^T \mathbf{W}_l \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right) \tag{14}$$

$$\triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \sum_{l=1}^{L} J_l. \tag{15}$$

The sub-matrices are given by

$$\mathbf{W}_l = \frac{1}{4\sigma^2} \begin{bmatrix} 2N & 0 & 0 \\ 0 & 2N\tilde{A}_l^2 & N^2\tilde{A}_l^2 \\ 0 & N^2\tilde{A}_l^2 & \frac{2}{3}N^3\tilde{A}_l^2 \end{bmatrix}. \tag{16}$$

Note that the noise variance is multiplied onto all elements and can therefore be ignored. In sub-band processing, like [7], this may not be the case, however, as the noise level may vary from one sub-band to another, meaning that we would have a σ_l^2 for each sub-band.

Using (16) along with the definition of the parameter vector $\boldsymbol{\theta}_l$, J_l can be expressed as

$$J_l = \frac{1}{4\sigma^2}\left(\begin{bmatrix}\tilde{A}_l\\ \tilde{\phi}_l\\ \tilde{\omega}_l\end{bmatrix} - \begin{bmatrix}A_l\\ \phi_l\\ \omega_l\end{bmatrix}\right)^T \times \begin{bmatrix}2N & 0 & 0\\ 0 & 2N\tilde{A}_l^2 & N^2\tilde{A}_l^2\\ 0 & N^2\tilde{A}_l^2 & \frac{2}{3}N^3\tilde{A}_l^2\end{bmatrix}\left(\begin{bmatrix}\tilde{A}_l\\ \tilde{\phi}_l\\ \tilde{\omega}_l\end{bmatrix} - \begin{bmatrix}A_l\\ \phi_l\\ \omega_l\end{bmatrix}\right). \quad (17)$$

When codebooks for the parameter subsets $\boldsymbol{\theta}_l$ are used so that $\hat{\boldsymbol{\theta}}_l \in \mathcal{C}_l$, (14) simplifies further as

$$\hat{\boldsymbol{\theta}}_l = \arg\min_{\boldsymbol{\theta}_l \in \mathcal{C}_l} \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right)^T \mathbf{W}_l \left(\tilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l\right). \quad (18)$$

It should be noted that in speech enhancement and separation, the phase is often omitted. We have, however, retained it here for completeness.

The question remains which estimators to use for the parameters of the sinusoidal model. Our derivations were based on the assumption that the estimated parameters will be distributed according to (4). An asymptotically optimal estimator of the frequencies is the periodogram while for the complex amplitudes, the least-squares estimator is efficient for white Gaussian noise and asymptotically so for colored noise [14].

4. RESULTS

Next, we will present some simulations results. The aim of the first experiments reported here is to demonstrate that the proposed metric leads to superior estimates as compared to naive approaches ignoring the different uncertainties associated with the intermediate parameters, corresponding to using $\mathbf{W} = \mathbf{I}$, as is commonly done in the literature. This results in a least-squares (LS) estimator in the vector quantization process. Additionally, we will compare to the optimal performance obtained as follows: for each codebook entry, a signal is reconstructed and the 2-norm of the error between this signal and the observed signal is measured. The codebook entry that leads to the lowest error is then chosen as the estimate. This approach, which we refer to as analysis-by-synthesis (AbS) is optimal in the sense that it chooses the codebook entry that best explains and reconstructs the observed signal. It is, however, also computationally expensive as it measures distances of N-dimensional signals rather than the M-dimensional parameter vectors. The experiments are carried out by generating a signal $\mathbf{x}$ from a set of parameters from the codebook after which noise is added. The intermediate parameters are then found using a 8192 point FFT and these are then quantized using the different metrics. We here use the sinusoidal model, the metric, and the estimators discussed in Section 3. Moreover, we use a random codebook of size 4096 which has been populated by realization of uniformly distributed phases and frequencies between 0 and 2π and Rayleigh distributed amplitudes. The performance is measured as the percentage of correctly estimated codebook entries. If, as has been hypothesized, the proposed metric is good, it should lead to better estimates than the simple least-squares estimates and to estimates close to those obtained with the AbS method. The results are depicted in Figures 1 and 2 as functions of the signal-to-noise ratio (SNR) with $N = 50$ and the number of samples N with $SNR = 0$ dB, respectively. The SNR is here defined as $10\log_{10} A_1^2/\sigma^2$. For each data point, 1000 Monte Carlo trials were run. The figures show that the proposed metric outperforms least-squares in the regions of interest while for extremely low and high SNRs, all the methods approach 0 and 100 %, respectively, and similarly for low and high N. It can also be seen that the proposed metric leads to results that are very close to those obtained using the AbS approach, and this is done at a significantly reduced computation time.

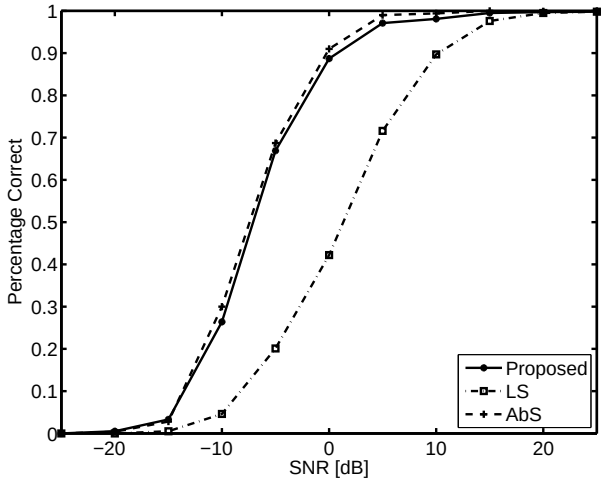

Fig. 1. Percentage of correctly estimated codebook entries as a function of the SNR for $N = 50$.

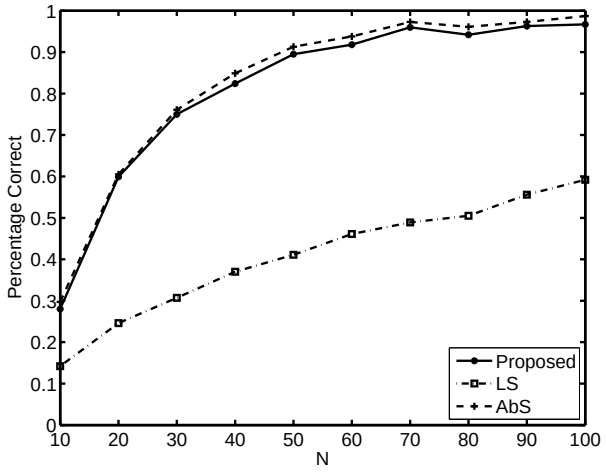

Fig. 2. Percentage of correctly estimated codebook entries as a function of the number of samples N for an SNR of 0 dB.

In the second part of our experiments, we will now, as a proof of concept, assess the performance of the proposed metric incorporated in [7] using the speech separation database in [15] (we refer to these papers for further details on the system and the speech database, respectively) and compare it to that obtained using well-known methods. The signals were down-sampled to 8 kHz and we used windows of length 32 ms along with a frame-shift of 8 ms along with codebooks we used 11 bits for amplitude and 3 bits for frequency part with a sinusoidal model order of 50. For training purposes, 500 utterances were used and these were subsequently excluded from the evaluation. The test data is mixtures of two signals, a target signal and a masker, at different signal-to-signal ratios (SSRs). The results shown here are for speakers 9 and 23 from the database, and we mea-

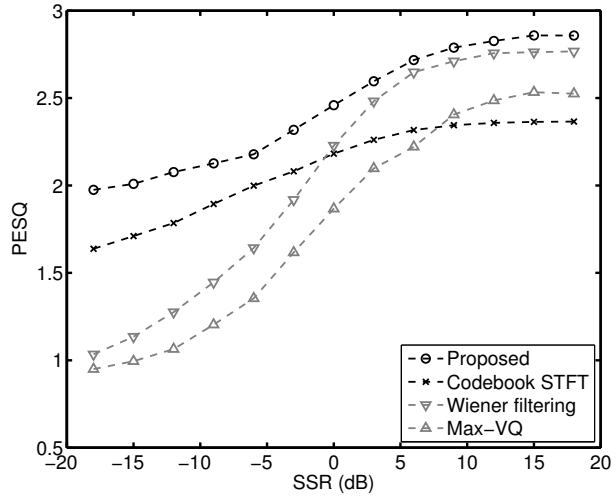

Fig. 3. PESQ scores obtained for the reconstructed signal as a function of the SSR (in dB).

sure the quality of the reconstructed signals using PESQ [16]. In Figure 3, the scores obtained for the different methods are reported as averages over the test set. The methods are: the STFT-based method of [8], Max-VQ [5], and Wiener filtering in the STFT domain [17]. Codebooks of size 2048 were used for these methods with the same training data as before. It can be observed that the proposed metric in combination with the system of [7] performs well, attaining good quality for the separated target speech signal for a wide range of SSRs. In fact, the method using the proposed metric achieves the highest score for the range of SSR values considered here.

5. CONCLUSION

In this paper, a new metric for VQ-based speech enhancement and separation has been proposed and its use exemplified on a specific parametrization, namely sinusoidal modeling. The metric was derived based on statistical arguments and expressions for the asymptotic distribution of maximum likelihood estimators. It essentially takes the uncertainties of different parameters into account in the quantization process. This was then demonstrated to lead to superior estimates in Monte Carlo simulations with a vector quantizer as compared to the commonly used squared error measure. In fact, the proposed metric proved to perform close to the optimal performance, showing that only a small loss is incurred by operating on intermediate parameters. Moreover, the proposed metric was incorporated in a speech separation algorithm that was demonstrated to generally perform favorably compared to state-of-the-art methods. The proposed methodology can of course be adapted to other parametrizations and could lead to improved results there too.

6. REFERENCES

[1] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15(4), pp. 1194–1203, May 2007.

[2] S. Srinivasan, J. Samuelsson, and W.B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14(1), pp. 163–176, Jan. 2006.

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15(2), pp. 441–452, Feb. 2007.

[4] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 4(5), pp. 383–389, 1996.

[5] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, 2003, pp. 1009–1012.

[6] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24(1), pp. 30–44, Jan. 2010.

[7] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. on Audio, Speech and Language Processing*, 2010, accepted.

[8] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, pp. 957–960, May 2006.

[9] R. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24(1), pp. 16–29, 2010.

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

[11] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE SP Mag.*, vol. 21(4), pp. 36–47, July 2004.

[12] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Processing*, vol. 46(10), pp. 2726–2735, Oct. 1998.

[13] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle," *Elsevier Signal Processing*, vol. 17, pp. 383–387, 1989.

[14] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results and an application," *IEEE Trans. Signal Processing*, vol. 48(2), pp. 338–352, Feb. 2000.

[15] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24(1), pp. 1–15, 2010.

[16] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Recommendation P.862, International Telecommunication Union, Feb. 2001.

[17] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15(6), pp. 1766–1776, Aug. 2007.