Conditional Vector Quantization for Speech Coding

Yannis Agiomyrgiannakis and Yannis Stylianou

Abstract—**In many speech-coding-related problems, there is *available* information and *lost* information that must be recovered. When there is significant correlation between the available and the lost information source, coding with side information (CSI) can be used to benefit from the mutual information between the two sources. In this paper, we consider CSI as a special VQ problem which will be referred to as conditional vector quantization (CVQ). A fast two-step divide-and-conquer solution is proposed. CVQ is then used in two applications: the recovery of highband (4–8 kHz) spectral envelopes for speech spectrum expansion and the recovery of lost narrowband spectral envelopes for voice over IP. Comparisons with alternative approaches like estimation and simple VQ-based schemes show that CVQ provides significant distortion reductions at very low bit rates. Subjective evaluations indicate that CVQ provides noticeable perceptual improvements over the alternative approaches.**

I. Introduction

THERE is a constant need for speech codecs with decreased bit rate, increased quality, robustness to bit errors and data losses. The speech signal has considerable redundancy that has been used in many ways for speech coding.

Several speech coding problems, like Speech Spectrum Expansion (the reconstruction of 4–8 kHz speech spectrum) and the recovery from packet losses in voice over IP (VoIP), face the following situation: there is *available* information and *lost* information, and the lost information has to be -somehow- *recovered* from the available information. This is an *estimation* problem when there is no possibility to transmit additional data, and a *coding* problem when data transmission is permitted. In a simple coding scenario where the available information is coded independently of the lost information (however, useful to the decoder), there is no benefit from the mutual information between the two sources: the *lost* information and the *available* information. Therefore, it is desirable to encode the former having the latter as *side* information.

In terms of (Conditional) rate-distortion theory, this is referred to as a coding with side information (CSI) problem [1], [2], and is schematically shown in Fig. 1, where Y is the information that will be coded, and $\hat{X}$ is the side information (with distortion) available at the encoder and the decoder. Estimation can be seen as a particular case of CSI, where the transmitted bit stream is empty. In this paper, we show that CSI can have many applications in speech coding like wideband speech coding, bandwidth expansion, and packet-loss concealment.

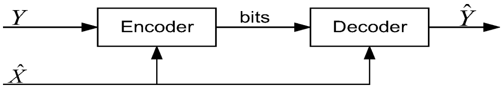

Fig. 1. Coding with side information.

There has been much effort in the enhancement of the narrowband (0.3–3.4 kHz) Public Switch Telephone Network (PSTN) speech signal by bandwidth expansion; the highband is estimated from the narrowband using several methods like vector quantization (VQ) mapping [3], Gaussian mixture model (GMM)-based estimators [4], [5], and hidden Markov models (HMMs) [6]. These attempts report an improvement over narrowband speech, although the resulting speech signal suffers from artifacts. The quality of the reconstructed speech is bounded by the relatively low mutual information between the two bands [7], [8] and the poor performance of estimation [9]. On the other hand, the acceptable performance of these methods indirectly states that the required bit rate for high-quality highband reconstruction should be low. Coding the highband without taking advantage of the highband knowledge carried at the narrowband, results in a higher bit rate. Therefore, it is beneficial to encode the highband having the narrowband as side information available to the encoder and the decoder.

It is widely accepted that for many speech sounds, the lower frequencies are perceptually more important than the higher frequencies. Therefore, in wideband speech coding, it may be desirable to separately encode the spectral envelope of the higher frequencies from the spectral envelope of the lower frequencies. Moreover, different fidelity requirements may be used in each band. For example, memoryless coding of the wideband spectral envelopes (0–8 kHz) using 14 line spectrum frequencies (LSFs) requires ≈ 41 bits/frame, while coding narrowband spectral envelopes (0–3.4 kHz) using 10 LSFs requires ≈ 24 bits/frame [10]. Because a high distortion is, in general, acceptable at the higher frequencies the use of a nonweighted single fidelity criterion to the whole wideband spectral envelope is perceptually not optimal. Furthermore, different bands may need to be encoded using different analysis/synthesis rates. Splitting the wideband spectral envelope in two bands and coding them with different fidelity criteria can be quite advantageous, but it results to an information loss equal to the mutual information between the two spectra. Coding with side information may use most of the mutual information, by reestablishing the broken dependencies between the two information sources [1].

New packet-based applications like VoIP generate new demand for codecs. Packets, typically containing 10–30 ms of encoded speech, may be lost or unacceptably delayed. A lookahead buffer called "jitter buffer" containing a few packets of speech is used to counteract small delays of packet arrivals. One lost packet results to the loss of 1–2 speech frames and

Manuscript received May 19, 2005; revised March 10, 2006. This work was supported by the General Secretary of Research Technology, Hellas and ICS FORTH, under an ARISTEIA grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

The authors are with the Department of Computer Science, Institute of Computer Science, Foundation of Research and Technology Hellas, University of Crete, Hellas, Heralou, Greece (e-mail: jagiom@ics.forth.gr; styliano@ics.forth.gr).

Digital Object Identifier 10.1109/TASL.2006.881702

1558-7916/$25.00 © 2006 IEEE

depending on the speech codec used, the reconstruction error can be propagated to several following frames [11]. An obvious way to cope with this is to use forward error correction (FEC) [11]; the information of the current frame is repeated in the next frame, but the added redundancy does not take into account the information carried at the neighboring frames. Some researchers try to estimate the lost spectral envelope from the previous frame(s) [12], [13]. Coding with Side Information can be used to introduce a small size corrective bit-stream that provides an enhanced estimation/coding of the lost spectral envelope(s), up to a pre-defined fidelity requirement. In other words, the idea is to repair the loss, not to repeat the loss.

Coding with Side Information is not something completely new in speech coding. In fact, various forms of predictive coding can be seen as CSI; the current frame is coded having the previous frame as side information under certain distortion requirements. In this perspective, CSI can be seen as a generalization of Predictive Coding, with complex nonlinear input-output space relationships, where adverse but relevant information sources (like LSFs, energy, voicing, pitch) can be used as side information.

In this paper, we suggest a VQ-based solution to the CSI problem. In Section II, the CSI problem is discussed using conditional rate-distortion theory arguments, in comparison with estimation and simple VQ. The role of mutual information is discussed and a distortion-rate bound for CSI is given. The discussion is supported by a toy example. In Section III we formulate/simplify the CSI problem as a generalization of VQ, which will be referred to as the conditional vector quantization (CVQ) problem, and suggest a fast divide-and-conquer two-step solution. CVQ assumes a piecewise one-to-many mapping between input space X (the side information) and output space Y (the coded information). Section IV describes three estimation methods. The following sections discuss two applications of CSI. In Section V, we use CVQ to encode the highband 4–8 kHz LSFs using the narrowband 0–4 kHz LSFs as side information. We show that, provided an appropriate excitation, only 134 bits/s are enough for a high-quality highband reconstruction. In Section VI, CVQ is used to generate a repairing bit stream for the VoIP problem and encode the current spectral envelope, using the previous and the next spectral envelopes as side information. Using LSFs for the parameterization of the spectral envelopes, we show that a very low bit stream of 400 bits/s can significantly reduce the reconstruction distortion for single and double packet losses.

II. Coding With Side Information

Let us consider two correlated sources X, Y, and their joined source $Z = [XY]^T$. Source X is already transmitted from the encoder to the decoder, while source Y must be, somehow, reconstructed at the decoder. Three options are available then:

- estimate Y given X. In most cases mutual information $\mathcal{I}(x;y)$ between the two sources cannot be fully utilized;
- encode Y with a CSI system having X as side information. Mutual information $\mathcal{I}(x;y)$ can be effectively utilized;
- encode Y. In this case, mutual information is lost.

The best option for reconstructing Y will depend on the amount of mutual information, the available bit rate and the fidelity requirement. In this section we discuss about the benefits and the limits of CSI (as shown in Fig. 1), using rate-distortion theory arguments. The distortion-rate Shannon lower bound (SLB) for CSI will be provided, and a nontight distortion bound for estimation will be given as a special case.

A. Conditional Rate Distortion

Let $\mathcal{R}_x(\Delta_x)$, $\mathcal{R}_y(\Delta_y)$ and $\mathcal{R}_{xy}(\Delta_x,\Delta_y)$ be the rate-distortion functions for X, Y and Z, respectively, where Δ_x, Δ_y is the fidelity constraint for each of the corresponding variables. Let $D_x(x,\hat{x})$, $D_y(y,\hat{y})$ be some distortion measures over X-space and Y-space, respectively. Rate-distortion theory [14] states that

$$\mathcal{R}_y(\Delta_y) = \inf_{p(\hat{y}|y): E_{y,\hat{y}}\{D_y(y,\hat{y})\} \le \Delta_y} \mathcal{I}(y;\hat{y}) \tag{1}$$

where $\mathcal{I}(y;\hat{y})$ is the mutual information between the source and the encoded source. For the CSI problem, we are mainly interested in rate $\mathcal{R}_{y|x}(\Delta_y)$ which is the rate of the system depicted in Fig. 1. The formula for the conditional rate-distortion function [1] is analogous to (1)

$$\mathcal{R}_{y|x}(\Delta_y) = \inf_{p(\hat{y}|y,x): E_{x,y,\hat{y}}\{D_y(y,\hat{y})\} \le \Delta_y} \mathcal{I}(y;\hat{y}|x) \tag{2}$$

Note that $\mathcal{R}_{y|x}(\Delta_y)$ is the rate of the CSI system when side information X is provided with zero distortion. The conditional rate-distortion function satisfies the following inequalities [1]:

$$\mathcal{R}_{xy}(\Delta_x,\Delta_y) \ge \mathcal{R}_{y|x}(\Delta_y) + \mathcal{R}_x(\Delta_x) \tag{3}$$

$$\mathcal{R}_{y|x}(\Delta_y) \ge \mathcal{R}_y(\Delta_y) - \mathcal{I}(x;y) \tag{4}$$

$$\mathcal{R}_{xy}(\Delta_x,\Delta_y) \ge \mathcal{R}_y(\Delta_y) + \mathcal{R}_x(\Delta_x) - \mathcal{I}(x;y) \tag{5}$$

where $\mathcal{I}(x;y)$ is the mutual information between the two sources. Under moderate assumptions, inequalities (3)–(5) become equalities [1]. The assumptions are that there are no restricted transitions between X and Y (for any x and y, $P(y|x)$ is nonzero), and that distortions Δ_x and Δ_y are sufficiently small. When these assumptions do not hold, the above inequalities provide the performance bounds. On the other hand, when the assumptions hold there is no rate penalty for encoding source Y with a CSI system instead of jointly encoding X and Y. Therefore, coding X with fidelity Δ_x, and Y with fidelity Δ_y at a specific rate can be made either way: with typical source coding of the joined source Z or with CSI. Additionally, CSI has the advantage of being applicable in cases where the two sources X and Y are defacto separated. Furthermore, (4) states the role of mutual information: $\mathcal{I}(x;y)$ is the rate loss for encoding Y without knowing X.

Note that in [1] inequalities (3)–(5) are proven for X and Y taking values from finite alphabets. However, it is quite straightforward to extend the proof of the corresponding theorem to continuous sources.

B. Mutual Information

Mutual information provides the rate gain when a CSI system is used for coding Y, instead of a typical source coding system.

Furthermore, mutual information is provided in closed form [14]:

$$\mathcal{I}(x;y) = E_{x,y}\left\{\log\frac{p(x,y)}{p(x)p(y)}\right\} \tag{6}$$

When densities $p(x,y)$, $p(x)$, $p(y)$ are available through a continuous parametric model like a GMM, the integral in (6) can be approximated by stochastic integration [7], [8], according to the law of big numbers

$$\mathcal{I}(x;y) \approx \frac{1}{N}\sum_{n=1}^{N}\log\frac{p(x_n,y_n)}{p(x_n)p(y_n)} \tag{7}$$

where x_n and y_n are drawn from the joint pdf $p(x,y)$.

Several properties of mutual information provide further insight to the CSI problem. For example, theoretically we cannot increase the rate gain of a CSI system by using other transformations (1-1 mapping functions $g(\cdot)$, $f(\cdot)$) of either X or Y, because a transformation can only decrease mutual information, as stated by the data processing inequality [14]

$$\mathcal{I}(X,Y) \geq \mathcal{I}\left(g(X), f(Y)\right). \tag{8}$$

C. Distortion-Rate for CSI

A distortion-rate bound for CSI and squared error distortion measure can easily be derived via SLB for vector processes

$$\mathcal{D}_y(R_y) \geq \frac{d}{2\pi e}\exp\left(\frac{2}{d}\left(h(y) - R_y\right)\right) \tag{9}$$

where $h(y)$ is the differential entropy of source Y, and d the dimensionality of Y-space. Using inequalities (4) and (9) we can derive a SLB for the distortion rate function of vector processes for CSI

$$\mathcal{D}_y(R_{y|x}) \geq \frac{d}{2\pi e}\exp\left(\frac{2}{d}\left(h(y) - R_{y|x} - \mathcal{I}(x;y)\right)\right). \tag{10}$$

Note that inequality (4) is also valid for vector processes ([15, exer. 4.4]) and continuous sources.

In the CSI framework, estimation can be seen as the attempt to recover Y at the decoder without transferring any bits ($\mathcal{R}_{y|x} = 0$). By setting $R_{y|x} = 0$ we obtain a boundary to the performance of an estimator of Y given X

$$\mathcal{D}_y \geq \frac{d}{2\pi e}\exp\left(\frac{2}{d}\left(h(y) - \mathcal{I}(x;y)\right)\right). \tag{11}$$

This is the same estimation bound with the one provided in [7]. However, note that the bound is not tight [7]. Based on the discussion developed in Section II-A, this is expected since the estimation distortion is rather high and mutual information is gained only when distortions Δ_x and Δ_y are sufficiently small.

The evaluation of CSI via the SLB is not practical for many sources (including the speech spectral envelopes) for two reasons: it is not always feasible to determine the tightness of the SLB and it is not always possible to make an accurate estimation of the differential entropy $h(y)$. Note that the estimation of differential entropy is not a trivial task when data lay on a manifold, since then $h(y)$ must be computed over the manifold. Furthermore, there is evidence that the spectral envelopes of speech lay on manifolds [16]. In such cases, the evaluation of CSI can be made via an estimation of the mutual information, e.g. as presented in Section II-B.

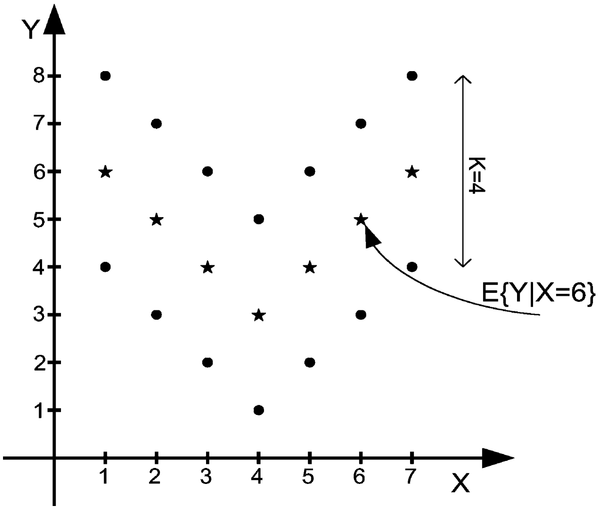

Fig. 2. Toy example.

D. A Toy Example

A toy example, similar to the one provided in [7], will be given to illustrate the notions described in previous subsections. Let $X \in \{1\ldots7\}$ and $Y \in \{1\ldots8\}$ be random variables taking values from finite alphabets. Let X, Y follow the joined distribution depicted in Fig. 2. The joint distribution codepoints (dots) have equal probability $p = 1/14$. Three bits are needed to describe Y. If we perform an estimation $\hat{y} = E_y\{Y|X\}$ of Y from X, we get the stars between the codepoints. Estimation $\hat{y}$ depends on the distance k between the two codepoints corresponding to the value of X. Note that for any $k \geq 3$, mutual information is constant ($\mathcal{I}(x;y) = 1.95$) bits and entropy is fixed to $\mathcal{H}(y) = 2.95$ bits. Therefore, the distortion-rate function $\mathcal{D}_y(R_{y|x})$ is independent of k. Obviously, estimation distortion can be arbitrary large for the given statistics. An important remark can be made: if 1 bit is provided, the reconstruction distortion falls to zero. For a given X, two codepoints may be chosen. The extra bit helps in choosing among these codepoints. In terms of our previous discussion, distortion Δ_y in the case of estimation (rate $R_{y|x} = 0$) is too large to take advantage of the mutual information. If 1 bit is provided, Δ_y becomes small enough ($= 0$) to gain $\mathcal{I}(x;y)$.

III. Conditional Vector Quantization

Intuitively, each value of X-space generates a different conditional pdf $p(y|x)$ for Y-space. We will try to capture the coarse structure of this mapping, using a VQ framework, which is re-

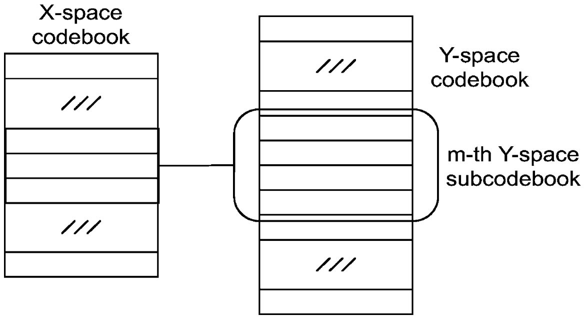

Fig. 3. CVQ.

ferred to as CVQ. The main idea is that each region in X-space is mapped to a different codebook of Y-space.

The problem of CVQ will be approached through a probabilistic point of view. Let $\vec{x} \in \mathcal{R}^P$ and $\vec{y} \in \mathcal{R}^D$ be random vectors of X-space and Y-space, respectively. The CVQ problem consists of constructing two linked codebooks $C_x \equiv \{\hat{\vec{x}}_m : m = 1 \dots M\}$ and $C_y \equiv \{\hat{\vec{y}}_{m,k} : m = 1 \dots M, k = 1 \dots K\}$, for X-space and Y-space respectively. Each codevector in C_x is linked to K codevectors in C_y, which form the mth subcodebook of C_y. The encoder finds the nearest C_x codevector and transmits the index of the nearest C_y codevector of the linked C_y subcodebook. The decoder locates the nearest C_x codevector and takes the estimation from the linked C_y subcodebook according to the transmitted index. Fig. 3 illustrates the two codebooks C_x and C_y, for $K = 4$. CVQ can be seen as a form of classified vector quantization [17], where the classification rule is taken from a VQ of X-space.

The CVQ reconstruction of $\vec{y}$ is a function of $\vec{y}$, $\vec{x}$, C_x, and C_y

$$\hat{\vec{y}}_{m,k} = Q_{y|x}\left(\vec{y}, Q_x(\vec{x}, C_x), C_y\right) \tag{12}$$

where $Q_x(.)$ is the quantization rule for X-space and $Q_{y|x}(.)$ the quantization rule for Y-space depending on X-space. The encoding rule can be expressed as

$$k = \arg\min_{k'} \left\{ d(\vec{y}, \hat{\vec{y}}_{m,k'}) \right\}, \text{ where } m = \arg\min_{m'} \left\{ d(\vec{x}, \hat{\vec{x}}_{m'}) \right\} \tag{13}$$

where $d(.,.)$ is some distortion measure. If we assume that $\hat{\vec{x}}_m$ and $\hat{\vec{y}}_{m,k}$ are random vectors spanning the discrete spaces C_x, C_y, respectively, then the average distortion of the CVQ encoding/decoding process becomes

$$D = \sum_{m=1}^{M} \sum_{k=1}^{K} \int\int p(\vec{x}, \vec{y}, \hat{\vec{x}}_m, \hat{\vec{y}}_{m,k}) d(\vec{y}, \hat{\vec{y}}_{m,k}) d\vec{x} d\vec{y}. \tag{14}$$

The joint probability $p(\vec{x}, \vec{y}, \hat{\vec{x}}_m, \hat{\vec{y}}_{m,k})$ can be analyzed to $p(\vec{x}, \vec{y}, \hat{\vec{x}}_m, \hat{\vec{y}}_{m,k}) = p(\vec{x}, \vec{y}) p(\hat{\vec{x}}_m|\vec{x}, \vec{y}) p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{y}, \vec{x})$ using the Bayes rule. The latter expression can be simplified with two CVQ-related assumptions. The first assumption is that the decoder cannot have knowledge of $\vec{y}$, and therefore $\hat{\vec{x}}_m$ is conditionally independent of $\vec{y} : p(\hat{\vec{x}}_m|\vec{x}, \vec{y}) \equiv p(\hat{\vec{x}}_m|\vec{x})$. The second assumption is that $\hat{\vec{y}}_{m,k}$ is conditionally independent of $\vec{x} : p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{x}, \vec{y}) \equiv p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{y})$ stating the piecewise mapping nature of the CVQ model; that no higher than first-order local statistics are taken into account when mapping a X-space region to K Y-space regions. Using these two assumptions, we conclude that

$$D = \int\int p(\vec{x}, \vec{y}) \sum_{m=1}^{M} p(\hat{\vec{x}}_m|\vec{x}) \times \sum_{k=1}^{K} p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{y}) d(\vec{y}, \hat{\vec{y}}_{m,k}) d\vec{x} d\vec{y}.$$

If the number of samples $[\vec{x}_n, \vec{y}_n]$, $n = 1, 2, \dots, N$ is large enough, then the law of big numbers states that D can be approximated by

$$D \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} p(\hat{\vec{x}}_m|\vec{x}_n) \sum_{k=1}^{K} p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{y}_n) d(\vec{y}_n, \hat{\vec{y}}_{m,k}). \tag{15}$$

The conditional probability $p(\hat{\vec{x}}_m|\vec{x}_n)$ is the association probability relating the input vector $\vec{x}_n$ with codevector $\hat{\vec{x}}_m$, while the association probability $p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{y}_n)$ relates the output vector $\vec{y}_n$ with the codevector $\hat{\vec{y}}_{m,k}$ of the mth subcodebook of C_y. The conditional dependence of $\hat{\vec{y}}_{m,k}$ with $\hat{\vec{x}}_m$ states that $\hat{\vec{y}}_{m,k}$ belongs to the mth subcodebook of C_y. Although the CVQ problem considers hard association probabilities taking values in {0,1}, the distortion formula (15) does not explicitly impose regular partitions. Therefore, minimization of D can also be made with nonregular partitions, i.e. Gaussians, in X-space and/or Y-space.

The minimization of D is a hard problem, but the complexity can be reduced if it is broken into several easier subproblems: first, compute a VQ of X-space and then minimize D. Since the partitioning of X-space determines the association probabilities $p(\hat{\vec{x}}_m|\vec{x}_n)$ and the codevectors $\hat{\vec{x}}_m$, the minimization problem breaks into a series of M typical weighted VQ minimization subproblems D_m

$$D \approx \sum_{m=1}^{M} \left[\frac{1}{N} \sum_{n=1}^{N} p(\hat{\vec{x}}_m|\vec{x}_n) \sum_{k=1}^{K} p(\hat{\vec{y}}_{m,k}|\hat{\vec{x}}_m, \vec{y}_n) d(\vec{y}_n, \hat{\vec{y}}_{m,k}) \right] = \sum_{m=1}^{M} D_m.$$

Furthermore, with hard association probabilities each of the M minimization subproblems, D_m operates in a subset of Y-space vectors providing, therefore, a significant computational advantage.

The resulting algorithm for hard association probabilities is:

- compute a VQ of X-space (M codevectors)
- for every $\hat{\vec{x}}_m \in C_x$:
- find the Y-space vectors corresponding to the X-space vectors that are nearest to $\hat{\vec{x}}_m$.
- perform a VQ on these Y-space vectors (K codevectors) to compute the mth Y-space subcodebook

At the case where $K = 1$, the CVQ problem is similar to the generalized VQ (GVQ) [18] problem, and the proposed solution is reduced to the nonlinear interpolative VQ (NLIVQ) [19] solution of GVQ. CVQ has also been used in [3]. Note, however, that in [3] the Y-space codebooks are taken from a Y-space partitioning that is trained independently of the X-space codebooks.

This solution is not consistent with (15) where it is clearly shown that the Y-space codewords depend directly on the X-space partition and not via a precomputed partitioning of Y-space.

IV. Estimation

In some applications like speech spectrum expansion (SSE) and VoIP packet loss concealment, the lost information Y is usually estimated from the available information X. The performance of the estimation is not always adequate in terms of subjective quality. CSI can overcome this limitation by providing an "enhanced" estimation at the cost of a few extra bits. A comparison between CSI and estimation is then necessary to indicate the practical performance gain when this strategy is adopted.

For this purpose, we focus on three memoryless mapping estimators; Linear Prediction, a simple VQ mapping called NLIVQ [19] and GMM-based estimation which will be referred to as GMM Conversion Function [5], [20]. The linear estimator provides a well-known baseline because it corresponds to the optimal linear relationship between the two spaces. The NLIVQ estimator provides useful insight as a special CVQ case (CVQ with $K = 1$). The GMM Conversion Function is a robust state-of-the-art estimator able to handle complex input–output space relationships.

A. *Linear Estimation*

In linear estimation, the estimated $\hat{\vec{y}}_t$ is a linear combination of the available information: $\hat{\vec{y}}_t = A\vec{x}_t$. Linear Estimation is also referred to as *linear prediction* [17], when the past is used to estimate the future.

B. *NLIVQ*

The NLIVQ method [19] uses two equal-sized codebooks, one for X-space codevectors and one for Y-space codevectors. The X-space vector is classified to the nearest X-space codevector which is mapped to one Y-space codevector. The X-space codebook is constructed by a variant of the well known binary split LBG VQ algorithm. The Y-space codebook is constructed from the means of Y-space vectors corresponding to X-space vectors that are nearest to the linked X-space codevector. NLIVQ is essentially the same to the CVQ method proposed in Section III when $K = 1$.

C. *GMM Conversion Function*

The GMMCF estimator uses an experts-and-gates regression function to "convert" the narrowband vectors to the wideband vectors. Both input and output spaces are modelled through GMM. The GMM conversion function is defined by

$$\hat{\vec{y}} = \sum_{m=1}^{M} p(\omega_m|\vec{x}) \left[\vec{y}_m + \Sigma_{yx}^m \left(\Sigma_{xx}^m\right)^{-1} (\vec{x} - \vec{x}_m)\right] \tag{16}$$

where $\vec{x}$ is the input vector associated with X-space, $\hat{\vec{y}}$ the estimation of $\vec{y}$, $\vec{x}_m$ and $\vec{y}_m$ denote the centroids of the mth Gaussian of X-space and Y-space, respectively, and Σ_{xx}^m is the covariance matrix of the mth X-space Gaussian, Σ_{yx}^m is the cross-covariance matrix that relates the mth Gaussians of X-space and Y-space, and ω_m denotes the mth class of X-space. Finally, $p(\omega_m|\vec{x})$ is the gating probability given by

$$p(\omega_m|\vec{x}) = \frac{p(\omega_m)\left|\Sigma_{xx}^m\right|^{-0.5} e^{-0.5(\vec{x}-\vec{x}_m)^T(\Sigma_{xx}^m)^{-1}(\vec{x}-\vec{x}_m)}}{\sum_{n=1}^{M} p(\omega_n)\left|\Sigma_{xx}^n\right|^{-0.5} e^{-0.5(\vec{x}-\vec{x}_n)^T(\Sigma_{xx}^n)^{-1}(\vec{x}-\vec{x}_n)}}. \tag{17}$$

The learning process for the GMM-based estimation function comprises of two stages. In the first stage a GMM of the X-space is estimated via the standard EM algorithm, while in the second stage the Y-space means $\vec{y}_m$ and the matrices Σ_{yx}^m are computed using a least-squares criterion [20]. For the experiments, we used diagonal covariance matrices Σ_{xx}^m and full cross-covariance matrices Σ_{yx}^m.

V. Application: CVQ of Highband Spectral Envelopes for Speech Spectrum Expansion

The problem of SSE has gained attention as a cost effective way to enhance narrowband speech into wideband. The main assumption is that narrowband (NB) speech contains enough information for the reconstruction of the missing highband (HB) frequencies. Another assumption is that the listener does not need an exact reconstruction of the lost frequencies but a perceptually valid one. Consequently, many researchers try to estimate the lost information from the transmitted information [3]–[6], [9]. Narrowband features like spectral envelopes under several parameterizations, pitch, voicing, zero-crossings, etc., have been extracted from the narrowband speech signal and used for the estimation of a highband features. The highband is then reconstructed from these features, usually an LSF spectral envelope and a gain parameter. The highband excitation is often an altered form of the narrowband excitation [6] or modulated white noise [21]. Reconstructed speech suffers from artifacts like whistling sounds and crispy sounds whose nature is associated with the employed excitation. These artifacts disappear if the highband LSFs are encoded with a few bits. However, the distortion at which this happens is significantly lower that the distortion resulting from the estimation. Therefore, it seems that a high-quality reconstruction of the highband cannot be based solely on estimation.

This observation is also supported by mutual information measurements using formula (7) in [7] which show that under several parameterizations, highband spectral envelopes and narrowband spectral envelopes share approximately 2.3 bits of mutual information. Furthermore, experimental setups in [3] with several estimators and parameterizations provide similar results.

A. *Objective Results*

We conducted several experiments to evaluate the quality of the reconstruction of highband spectral envelopes using the previously presented estimators, CVQ and simple VQ. All experiments were conducted using the TIMIT database. LSF parameterization was used for representing the spectral envelopes in the low and in the high-band using 14 and ten size vectors, respectively. Each experiment involves the use of approximately

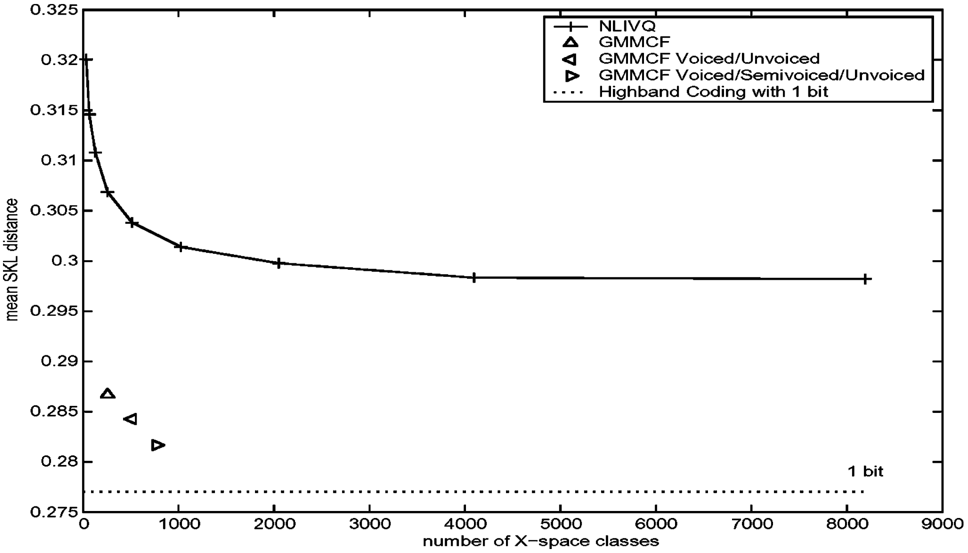

Fig. 4. Performance (SKL mean distance) of a NLIVQ estimator and three GMMCF-based estimators, in comparison with the SKL distortion of a simple highband VQ with 1 bit.

730 000 LSF vectors for training and about 270 000 LSF vectors for testing, while frames considered as silence were excluded from the training or the testing corpus. A pre-emphasis filter with $\mu = 0.95$ was applied on the narrowband signal. The length of the analysis window was set to 30 ms. Voicing decisions -when needed- were made according to the energy ratio between the narrowband and the highband. As an objective metric, we used the symmetric Kullback Leibler (SKL) distance given by

$$d_{\mathrm{SKL}}(P,Q) = \frac{1}{2\pi}\int_{0}^{2\pi}(P(\theta)-Q(\theta))\log\frac{P(\theta)}{Q(\theta)}d\theta \qquad (18)$$

where $P(\theta)$ and $Q(\theta)$ are the two power-normalized spectral envelopes. The SKL distance can also be seen as a weighted formant distance [22] and it seems to reflect the perceptual differences between AR spectra [23]. The SKL distance was chosen as a better alternative to spectral distortion.

Fig. 4 depicts the mean SKL distance of the presented estimators. The horizontal axis refers to the number of X-space classes used by the estimator. For example, the NLIVQ estimator has been tested for 16, 32, ..., 2048 4096 classes, while the GMMCF estimator has been tested for 128 classes. Accordingly, a multiple estimator system with two GMMCF estimators (one for voiced frames and one for unvoiced frames) had $2 * 128 = 256$ classes, and a voiced/semivoiced/unvoiced system had 384 classes. Results from the NLIVQ estimator are linked with a line to indicate the convergence of the estimator. The horizontal dotted line shows the mean SKL distance achieved when the highband is encoded with just 1 bit. From this figure, it is worthwhile to note that even the best estimator cannot provide 1 bit regarding the highband spectral envelope.

The performance of CVQ for 1, 2, 3, and 4 bits/frame and 128 classes for the X-space is shown in Fig. 5, where we have also included the performance of simple Y-space VQ with 1...5 bits, and the performance of the previously mentioned estimators. Clearly, CVQ outperforms VQ. Notice that CVQ benefits more from the mutual information, as the number of bits, $\log_2(K)$, is increasing.[1] For CVQ with 1 bit/frame, the distortion is slightly below the distortion of VQ with the same rate. It is a slight improvement compared to the performance of the best estimator (nearly 1 bit/frame), but it is much better than the performance of the NLIVQ estimator. Note that the best estimator has extra voicing information and uses second order local statistics (covariances) to perform the mapping between X-space and Y-space. Therefore, CVQ can be directly compared with NLIVQ which is a special case of CVQ ($K = 1$). As coding rate $\mathcal{R}_{y|x} = \log_2 K$ increases, CVQ gains approximately 1 bit from the available mutual information, in terms of the SKL-based distortion. In relative terms, CVQ offers a 20% improvement over simple VQ.

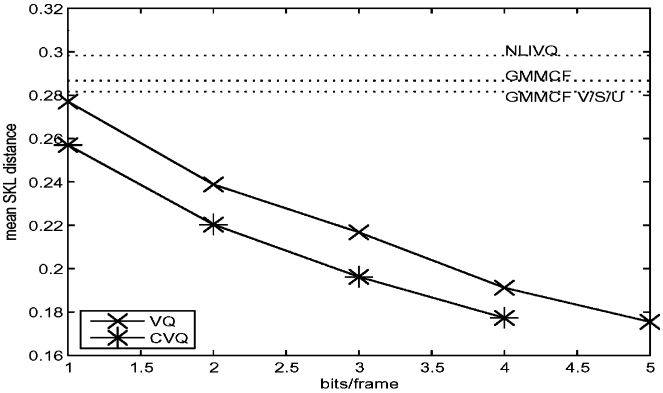

Fig. 5. Performance of CVQ with 128 X-space classes, in comparison with the SKL distortion of a simple highband VQ with 1, 2, 3, 4, and 5 bits. The performance of the estimators is indicated with horizontal lines.

B. Subjective Results

We conducted a subjective evaluation of a Speech Spectrum Expansion system with $\mathcal{R}_{y|x} = 4\ \mathrm{bits/frame}$ and an analysis/synthesis rate of 33.3 frames/s, and found that ≈ 134 bits/s for the highband spectral envelope were enough to provide a high-quality highband reconstruction when modulated white noise

[1] K is the size of each linked subcodebook.

is used as excitation signal for the highband and the highband energy is considered to be known. For the modulation of the white noise excitation signal, the time envelope of the 3–4 kHz band signal was used [21]. Since synthesis of noise using OLA (Overlap and Add) introduces audible fluctuations [24], we used a time variable lattice filter obtained by a sample by sample interpolation of their (reflection) coefficients. The highband signal is then scaled according to the highband energy. Finally, narrowband speech and the resulting highband speech are combined to synthesize the wideband speech signal.

Original excitation of the highband exhibits a specific time-domain structure in terms of energy localization. The time-domain modulation of the white noise tries to simulate this property of the original excitation signal. However, this modulation is not always successful. When highband spectral envelopes are well estimated, errors in the excitation signal are not perceived; then a high-quality wideband signal is obtained. To the contrary, when highband spectral envelopes are not well estimated, errors in the highband excitation signal tend to be amplified resulting in a reconstructed wideband signal of poor quality.

A further insight to the SSE problem requires the study of the complex auditory masking phenomena that take place in the reconstructed wideband signal. Most probably, the highband distortion is masked by a combination of *time-masking* and *frequency-masking* phenomena. Time-masking is partially exploited here by the time-domain modulation of the noise excitation. Frequency masking is directly related to the highband gain. For example, a lower highband gain might cause several highband frequency components to fall below the masking threshold imposed by the much stronger (in terms of energy) lower frequency formants. Therefore, the highband gain should be studied independently of the highband spectral envelope in order to isolate artifacts related to spectral shape from artifacts related to the relative energy of the highband. This section focuses only on CVQ of highband spectral envelopes.

Some artifacts that mainly occur in unvoiced parts of the speech, are caused by rapid amplitude variations of the time-envelope. These variations produce a "crispy" character to some consonants. To overcome these problems, we follow a strategy similar to [21] and filter the time-envelope with a low-pass variable filter controlled by a simple voicing criterion, based on the energy ratio between the two bands. Smoothing is performed mainly in unvoiced parts of speech, leaving the time-envelope of voiced speech almost untouched.

We have subjectively evaluated the described speech spectrum expansion system for the three following cases:

- original highband LSFs;
- estimated highband LSFs by NLIVQ with 128 classes;
- CVQ coded highband LSFs with 134 bits/s.

The degradation category rating (DCR) test was used to measure the quality degradation of the reconstructed wideband speech when the latter is compared with the original wideband speech [25].

A first test was conducted to determine an upper bound of reconstructed speech quality for the implementation of the described highband SSE system. A second test provides an example of quality achieved by an NLIVQ estimator. All presented estimators showed unnoticeable differences in terms of perceived quality and NLIVQ was chosen for being the simplest among all. In a third test, CVQ was used with 128 X-space classes and 4 bits/frame. A frame rate of 33.3 frames/s was found to be sufficient. Therefore, the total bandwidth requirements are 134 bits/s.

TABLE I
DCR Test Rating (and 95% Confidence Intervals) Using the Original Wideband Signal as Reference

Method	DCR score (95% CI)
NLIVQ estimator with 128 classes	3.59 (0.23)
CVQ with 4 bits/frame	4.41 (0.20)
ORIGINAL highband envelope	4.67 (0.15)

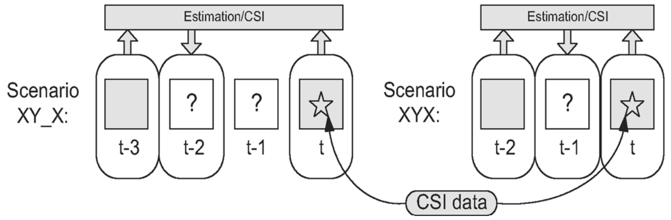

Fig. 6. Two CSI scenarios for recovery from single and double packet losses, assuming a two-packet jitter buffer. The boxes indicate lost/received packets. A lost packet is CSI encoded using neighboring packets. In each scenario, the CSI data -when needed- is stored in the packets with the star.

For the first two tests 29 listeners participated and they were asked to vote for 41 utterances from several speakers. From these utterances, a random subset was presented to each listener; 14 utterances for the NLIVQ estimator, 14 utterances using the original LSFs, a null-set of five stimuli and four repeated stimuli per test. Listeners that were severely biased and inconsistent were not taken into account. The CVQ utterances were evaluated with 19 listeners, using 16 utterances from the test set, four repeated stimuli, and five null-set stimuli, under the very same conditions.

The results from the DCR tests are shown in Table I. The DCR score of the first test proves that the SSE system used here provides high-quality reconstruction of the 4–8 kHz speech spectrum. The low DCR score of the NLIVQ estimator was mainly attributed to some crispy noise artifacts. The proposed CVQ coding at 4 bits/frame and 33.3 frames/s provides a very good DCR score, which is quite close to the score obtained using the original LSFs. Results can be found in "http://www.ics.forth.gr/~jagiom/SpeechSpectrumExpansion.html".

VI. Application: CVQ of Lost Spectral Envelopes for Voice Over IP

Speech signal contains considerable temporal correlations. These correlations can be used to tackle the packet loss problem in VoIP. For example, the LSF parameters of adjacent frames are highly correlated and this has been successfully used in modern codecs for packet loss concealment (PLC) [26]. Waveform substitution PLC algorithms try to reconstruct the lost speech giving emphasis to the continuity of the speech waveform [27]. However, waveform substitution techniques do not ensure the continuity of the sinusoidal tracks nor phase coherency. These desirable properties can be provided by sinusoidal PLC schemes [28] which outperform waveform PLC schemes [27]. Sinusoidal PLC schemes require the knowledge of the spectral envelope(s)

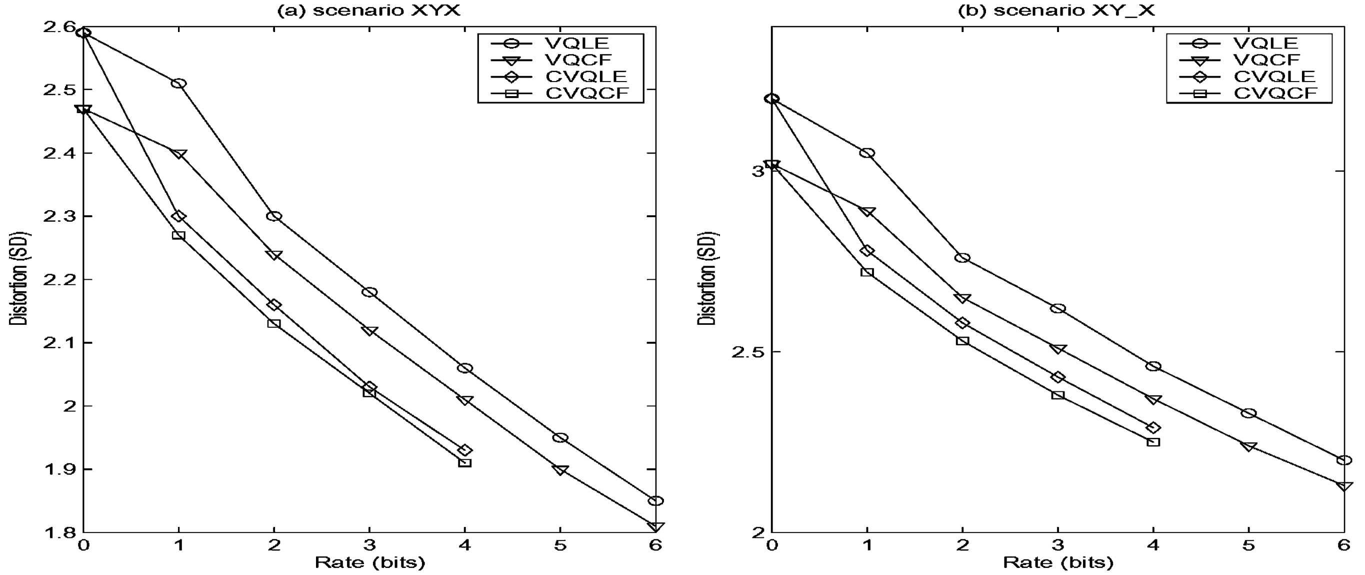

Fig. 7. Distortion-rate measurements for the two scenarios XYX, XY_X.

of the lost speech frame(s). The lost spectral envelopes can be recovered with a repetition scheme or with more sophisticated estimators [12], [13].

The performance of the estimators is bounded by the mutual information and the structure of the underlying probability space. To overcome these problems, FEC techniques have been proposed [11]. These algorithms require full repetition of the information for each packet consuming, however, bandwidth (by doubling the bit rate of a code.) CSI can be used to provide an adequate reconstruction of the lost spectral envelopes with minimal extra bandwidth. More specifically, past and future spectral envelopes (contained in the jitter buffer) can be used as side information for encoding the lost spectral envelope(s). In [25, p. 158], a deterministic frame-fill technique has been used to increase the temporal resolution of coarsely sampled (every 30 ms) spectral envelopes. CVQ is the stochastic counterpart of this frame-fill technique and it is capable of handling the complicated correlations between the received and the lost spectral envelopes.

A typical jitter buffer usually contains 1–2 packets (20–40 ms) of speech. With a jitter buffer of two packets, CVQ can be used to effectively handle single and double packets losses. We will focus on the narrowband spectral envelopes, typically encoded with ten LSFs per frame, assuming that each packet contains one spectral envelope. Note, however that CVQ can be also be used for other parameters like pitch and gain.

Let $\vec{v}_t$, $t = \{1, 2, \ldots\}$ be the sequence of transmitted LSF vectors, and $\vec{v}_t$ be the last received LSF vector. Single packet losses can be recovered with a CSI scheme that encodes $\vec{v}_{t-1}$ having $\vec{v}_{t-2}$ and $\vec{v}_t$ as side information. This case will be referred to as XYX *scenario*. Double packet losses can be recovered in two steps: first reconstruct $\vec{v}_{t-2}$ with a CSI scheme that uses $\vec{v}_t$ and $\vec{v}_{t-3}$ as side information, and then use the recovered $\hat{\vec{v}}_{t-2}$ and $\vec{v}_t$ to reconstruct $\vec{v}_{t-1}$. The first step will be referred to as XY_X *scenario*, while the second step is identical to the XYX scenario. This two-step procedure effectively *reuses* the single-frame corrective bit-stream. In fact, objective measurements show that $\vec{v}_{t-1}$ is recovered with *less* distortion than $\vec{v}_{t-2}$, at the rate of 4 bits of side information per lost spectral envelope. The two scenarios are depicted in Fig. 6.

A direct employment of CVQ on both scenarios provides poor results. However, as M increases, CVQ performance also increases, showing that for reasonable memory requirements, the size of the linked codebooks is not enough to model the correlation between the two spaces (X-space and Y-space). CVQ memory requirements can be reduced if a portion of the available mutual information is removed by estimation. Therefore, we performed CVQ on the *estimation* residual $\vec{\epsilon}_t = \vec{y}_t - \hat{\vec{y}}_t$, where $\vec{y}_t$ is the true value of the lost spectral envelope, and $\hat{\vec{y}}_t$ is an estimation of this value given the side information.

Estimation residual $\vec{\epsilon}_t$ has considerable correlation with side information $\vec{x}_t$. For example, in scenario XYX, mutual information measurements according to the procedure described in Section II-B have shown that $\vec{y}_t$ and $\vec{x}_t$ share 7 bits, while the GMMCF estimation residual $\vec{\epsilon}_t$, and $\vec{x}_t$ share 2.61 bits. In other words, nearly 62% of the initial mutual information is removed by the estimation step. To further benefit from the remaining mutual information, CVQ can now be used with reduced memory requirements. Analogous measurements for XY_X scenario showed similar results. All mutual information measurements were made using diagonal covariance GMMs with 1024 Gaussians and 10^6 samples for the stochastic integration.

For the experiments in this section we used the default training set and testing set as these are defined in the TIMIT database. The AR filter was computed from the narrowband (0–4 kHz) signal with the autocorrelation method using preemphasis ($\mu = 0.95$). The spectral distortion measure defined as

$$\mathcal{D}(X_t, \tilde{X}_t) = \frac{1}{\pi} \int_0^{\pi} \left(20 \log_{10} \frac{\left|X_t(e^{j\omega})\right|}{\left|\tilde{X}_t(e^{j\omega})\right|} \right)^2 d\omega \qquad (19)$$

was used in all the experiments, where $|X_t(e^{j\omega})|$, $|\tilde{X}_t(e^{j\omega})|$ is the original spectrum and the reconstructed spectrum, respectively. In this section we chose the spectral distortion measure instead of the SKL distance metric used in the previous section because the correlation of this measure with the subjective quality is well known for narrowband spectral envelopes.

The distortion-rate measurements for both scenarios are shown in Fig. 7. We examine four different cases of CSI. The first two cases, referred to as VQLE and CVQLE encode the residual from the Linear Estimation using VQ and CVQ, respectively. The other two cases, referred to as VQCF and CVQCF, encode the residual from the GMMCF estimation. For each case, the performance of the corresponding estimator is presented at the rate of 0 bits/frame. This allows a direct comparison of CSI techniques and estimation methods in terms of distortion. In all scenarios, CVQ had $M = 256$ X-space classes and GMMCF had 128 X-space Gaussians.

Compared to estimation, just 4 bits per lost vector encoded via CVQCF provide a benefit of 0.56 dB (-22.7%) and 0.77 dB (-25.5%) for scenarios XYX and XY_X, respectively. Furthermore, the (mean) reconstruction distortion in scenario XYX falls below the 2-dB threshold that is considered to be the threshold for outliers [25]. In both scenarios CVQCF approximately gains 1.3 bits and CVQLE gains at least 1 bit, compared to VQLE. Therefore, a Linear Estimator should be preferred over a GMM-based estimator since it is less computationally expensive. The scenarios examined in this section are not directly comparable to the "predictive" scenarios used in the literature [12], [13]. Such comparisons are available in [29].

We conducted an informal listening test to evaluate the effect of the reported distortion reduction. The original excitation was used in all the reconstructed frames. The test was restricted to single and double losses of consequent LSF vectors. Compared to simple linear interpolation, the suggested CVQLE-based scheme using 4 bits/frame for XYX scenario and 4 bits/frame for XY_X scenario provides reconstructed speech with much fewer and/or significantly milder envelope related artifacts.

The results from the reported subjective tests show that artifacts related to spectral envelope distortions can be efficiently removed based on the proposed approach. More details regarding the subjective evaluation can be found in [29]. For speech codecs that rely explicitly on the use of an excitation signal (e.g., CELP-based coders), additional tests should be conducted including the coding of the excitation signal. Obviously, in this case a deterioration of the obtained quality is expected. On the other hand, the spectral envelope information is very important for the quality of the reconstructed signal for speech coders based on the sinusoidal representation [25], where the excitation signal is obtained through a phase model that is based on the spectral envelope information.

VII. Conclusion

We address the problem of CSI from a VQ-based perspective, formulating it as the CVQ problem, and provide a two-step solution. Summarizing literature results, we examine CSI using conditional rate-distortion arguments and link it to the mutual information. CVQ is then used in two applications, showing that minimal bit streams provide significant distortion reduction over estimation and compare favorably with VQ and VQ of an estimation residual. This distortion reduction effectively removes artifacts in the presented applications.

CVQ performance is, however, inevitably limited by memory requirements; therefore applicable only for very low bit rates, as an alternative to estimation when data transmission is possible. Furthermore, the proposed CVQ solution is suboptimal in many ways, i.e. input space partitioning is not made according to the minimization of the output space coding distortion. A better solution can be provided via gradient methods, but at the expense of a much higher computational cost.

References

[1] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 480–489, Jul. 1973.

[2] T. Linder, R. Zamir, and K. Zeger, "On source coding with side information dependent distortion measures," *IEEE Trans. Inform. Theory*, vol. 46, no. 11, pp. 2697–2704, Nov. 2000.

[3] J. Epps, "Wideband extension of narrowband speech for enhancement and coding," Ph.D. dissertation, Univ. New South Wales, Sydney, NSW, Australia, 2000.

[4] Q. Yasheng and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, QC, Canada, 2004.

[5] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000.

[6] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden markov model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China, 2003, vol. 1.

[7] P. Jax, "Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds," Ph.D. dissertation, Inst. Communication Systems and Data Processing (IND), Rheinisch-Westfdlische Technische Hochschule (RWTH), Aachen, Germany, 2002.

[8] M. Nilsson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, 2002.

[9] Y. Agiomyrgiannakis and Y. Stylianou, "Combined estimation/coding of highband spectral envelopes for speech spectrum expansion," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, QC, Canada, 2004.

[10] S. Stephen and K. P. Kuldip, "Multi-frame GMM-based block quantization of line spectral frequencies for wideband speech coding," in *Proc. ICASSP*, Philadelphia, PA, 2005.

[11] L. Roch, G. Philippe, and S. Redwan, "A study of design compromises for speech coders in packet networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, QC, Canada, 2004.

[12] R. Martin, C. Hoelper, and I. Wittke, "Estimation of missing LSF parameters using gaussian mixture models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001.

[13] J. Lindblom, J. Samuelsson, and P. Hedelin, "Model based spectrum prediction," in *IEEE Workshop on Speech Coding*, Delavan, WI, 2000.

[14] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[15] R. M. Gray, *Source Coding Theory*. Norwell, MA: Kluwer, 1990.

[16] R. Togneri, M. D. Alder, and Y. Attikiouzel, "Dimension and structure of the speech space," *IEE Proc.—I Communications, Speech and Vision*, vol. 139, no. 2, pp. 123–127, 1992.

[17] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[18] A. Rose, D. Rao, K. Miller, and A. Gersho, "A generalized VQ method for combined compression and estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996, pp. 2032–2035.

[19] A. Gersho, "Optimal nonlinear interpolative vector quantization," *IEEE Trans. Commun.*, p. 1285, 1990.

[20] Y. Stylianou, O. Cappe, and M. Eric, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, 1998.

[21] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model," in *Proc. IEEE Int. Conf. Acoust.*, Istanbul, Turkey, 2000, pp. 1153–1156.

[22] V. Raymond and K. Esther, "On the computation of the Kullback-Leibler measure for spectral distances," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 1, pp. 100–103, Jan. 2003.

[23] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP*, 2001.

[24] H. Pierre and D.-C. Myriam, "Adapting the overlap-add method to the synthesis of noise," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFx-02)*, Hamburg, Germany, 2002.

[25] B. W. Kleijn and K. K. Paliwal*, Speech Coding and Synthesis*. New Providence, NJ: Elsevier, 1995.

[26] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech Audio Process.*, 2004.

[27] U-T Recommendation G.711, A High Quality Low-Complexity Algorithm for Packet Loss Concealment With G.711 1999.

[28] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling," in *Proc. IEEE Workshop on Speech Coding*, Orlando, FL, 2002, vol. 1, pp. 173–176.

[29] Y. Agiomyrgiannakis and Y. Stylianou, "Coding with side information techniques for LSF reconstruction in voice over IP," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, 2005.

Yannis Agiomyrgiannakis received the B.Sc. degree in computer science and the M.Sc. degree in networks and telecommunications in 1999 and 2002, respectively, from the University of Crete, Heraklion, Crete, where he is currently pursuing the Ph.D. degree.

He has worked on low-footprint DSP implementations of speech coding and speech processing algorithms. His research interests include digital signal processing, speech processing, speech coding/enhancement, source/channel coding, and voice-over-IP.

Yannis Stylianou received the electrical engineering diploma from the National Technical University of Athens (NTUA), Athens, Greece, in 1991 and the M.Sc. and Ph.D. degrees in signal processing from the Ecole National Superieure des Telecommunications (ENST), Paris, France, in 1992 and 1996, respectively.

From 1996 to 2001, he was with AT&T Labs Research, Murray Hill/Florham Park, NJ, as a Senior Technical Staff Member. In 2001, he joined Bell-Labs Lucent Technologies, Murray Hill. Since 2002, he has been with the Computer Science Department, University of Crete, Heraklion, Crete, where he is currently an Associate Professor with the Department of Computer Science. He holds eight patents and participates in the SIMILAR Network of Excellence (6th FP) coordinating the task on the fusion of speech and handwriting modalities.

Dr. Stylianou was Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 1999 to 2002. He is currently Associate Editor of the EURASIP *Journal on Speech, Audio and Music Processing*.