# CODING WITH SIDE INFORMATION TECHNIQUES FOR LSF RECONSTRUCTION IN VOICE OVER IP.

*Yannis Agiomyrgiannakis*

Institute of Computer Science,
Foundation of Research & Technology Hellas
P.O.Box 1385 Heraklio, GR-711-10 GREECE
jagiom@ics.forth.gr

*Yannis Stylianou*

Department of Computer Science,
University of Crete, P.O.Box 2208,
Heraklion, Crete, GR-714 09 GREECE
yannis@csd.uoc.gr

**ABSTRACT**

In applications like VoIP, speech codecs have to deal with excessive packet losses, caused by network errors and/or delays. In this paper a new method for the reconstruction of lost speech spectral envelopes is presented, which is based on a statistical estimation function. We suggest the usage of a minimal "corrective" bitstream and propose Coding with Side Information (CSI) techniques for an efficient Forward Error Correction (FEC) strategy. The proposed methods are tested on multiple scenarios of missing frames. Objective results indicate that with only 4 bits per lost frame, a spectral distortion reduction of 0.77-1.14 dB is achieved, compared to results obtained by current state-of-the-art estimation methods. Compared to "predictive" estimation methods, the use of jitter buffer as side information and 4 bits per lost frame provide a 42% reduction of spectral distortion for single packet losses, and a 32% reduction for double packet losses. Subjective results indicate that the corrected speech has fewer artifacts.

## 1. INTRODUCTION

In VoIP, a voice packet is useless after it's playback time. A small buffer called "jitter buffer" counteracts small network delays and potential reorderings of the packet sequence. The size of this buffer is actually limited by the acceptable end-to-end transmission delay. A typical jitter size is only 1-2 packets (approximately 20ms-40ms), which results in an increased packet loss rate. Unfortunately, speech codecs were not initially designed to cope with such losses.

The design of speech codecs capable of providing acceptable quality with packet losses in the order of 10% even 20% is desirable. Some researchers propose FEC (Forward Error Correction) and introduce redundancy to the bitstream by replicating the description of each frame [1]. On average, FEC schemes increase network congestion, and may degrade overall performance [2]. On the other hand, the IP packet overhead is relatively large, therefore, doubling the codec bitrate results in a much smaller relative increase of the total bandwidth, especially when high compression codecs are used. The packet overhead can be significantly reduced with header compression techniques. For most codecs though, repeating more than once the speech bitstream, is prohibitively expensive in terms of bitrate. Furthermore, FEC does not benefit from the existence of the jitter buffer, i.e. if the n-th frame/packet is lost and it's neighboring frames are received, the bit rate to encode/repeat the n-th frame does not benefit from the highly correlated information carried by neighboring frames.

Improved PLC (Packet Loss Concealment) schemes claim that a quality improvement will be gained when non-linear estimators are used to predict the spectral envelope of the lost frame, from the spectral envelopes of the previous received frame(s). In [3], a bounded support GMM (Gaussian Mixture Model) is used to provide an enhanced estimation of the lost spectral envelope. In [4] a GMM is used to estimate lost LSF subsets. A Markov Chain prediction is proposed in [5]. All these schemes compare favorably with the simple repetition schemes currently used, but still provide reconstructed spectral envelopes with too high distortion.

We argue that a minimal "corrective" bitstream provides reconstructed LSFs (Line Spectrum Frequencies) with much lower distortion. The "corrective" bitstream can benefit from the already received LSF vectors to encode the lost LSF vectors. In terms of Conditional Rate Distortion theory, this is referred to as the Coding with Side Information (CSI) problem. The lost envelope ($Y$ source) will be coded having the received spectral envelope(s) ($Z$ source) as side information.

Under reasonable assumptions, a CSI scheme can benefit from the mutual information $\mathcal{I}(Z;Y)$ between the two sources [6], even in cases where the optimal estimator cannot [6], [7]. This is a clear advantage over FEC schemes, where the mutual information between the two sources is not used. The CSI scheme can also benefit from the existence of the jitter buffer, and utilize the received spectral envelopes. In a way, CSI is already used in speech coding since the widely used Predictive Coding of LSFs [8] can be seen as a form of CSI: the current LSFs are encoded having the previous LSFs as side information.

In this paper we investigate several CSI methods based on residual coding, for encoding lost LSFs for several cases of received/lost LSF vectors, taking advantage from the existence of the jitter buffer. We suggest the use of the estimator proposed in [9] for regression and residual coding. Furthermore, a VQ-based CSI approach that benefits from the mutual information between the estimation residual and the side information is proposed and evaluated. A direct comparison with estimation shows that there is a significant reduction of spectral distortion in the order of 0.77-1.14 dB, for several cases of lost/received LSFs, with just 4 bits per lost frame. Furthermore, the utilization of the jitter buffer in estimation, offers a significant spectral distortion reduction of 0.35-0.9 dB over the widely used forward prediction (estimation). Our CSI schemes are tuned to handle 1-2 packet losses, using a

This work was supported by the General Secretary of Research Technology Hellas and ICS FORTH, under ARISTEIA grant.

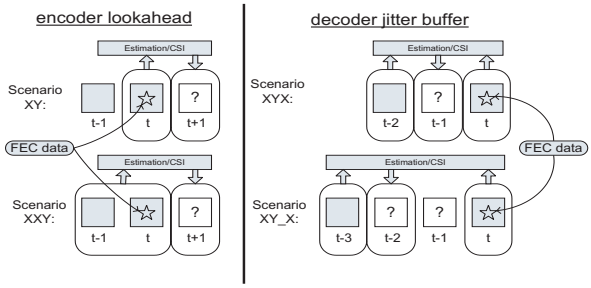


**Fig. 1**. The 4 examined scenarios of lost/received packets using a 0-2 packet jitter buffer.The boxes indicate lost/received packets. A lost packet (questionmark) is estimated or CSI encoded using some of it's neighboring packets. In each scenario, the CSI data -when needed- is stored in the packets with the star.

jitter buffer of 1-2 packets. Section 2 presents the investigated lost/received LSF scenarios. Section 3 presents the methods used to estimate the lost LSFs from the received LSFs and Section 4 introduces the tested CSI schemes. Section 5 states the conducted experiments and Section 6 reports the obtained results.

## 2. RECOVERY SCENARIOS

Let $\vec{x}_t$, $t = \{1, 2, ...\}$ be the sequence of transmitted LSF vectors. We assume that each packet contains 1 LSF vector. We further assume that the decoder has a jitter buffer of 1-2 packets and keeps a history of 1-2 packets. The idea is to use the information in the received packets to recover the information of a lost packet. Clearly, there are many possible combinations of lost/received packets that could be examined. Four possible scenarios, depicted in Figure 1, were selected. Incoming LSF vectors/packets are drawn as boxes. Lost packets contain a questionmark. CSI data are stored in each packet. The CSI data that is going to be used in each scenario under study, is presented by a star inside the boxes. The box with the star is always the last *received* packet $\vec{x}_t$. The two leftmost scenarios require 1 packet lookahead at the encoder, while the rightmost scenarios require a jitter buffer of 1-2 packets.

In scenarios XY, XXY, the LSF vector $\vec{x}_{t+1}$ is lost and one (for XY) or two (for XXY) previous LSF vectors are used as side information. Using more than two past spectral envelopes for estimation does not enhance the estimation performance, as shown in [3]. Scenario XYX considers the case when the current LSF vector is lost, while the next and the previous vectors are received and used as input space information. Scenario XY_X is the case when two consecutive LSF vectors are lost and we wish to recover $\vec{x}_{t-2}$ using the information carried in $\vec{x}_{t-3}$,$\vec{x}_t$.

The other lost vector $\vec{x}_{t-1}$, can be recovered from the reconstructed $\hat{\vec{x}}_{t-2}$ and the received $\vec{x}_t$ applying a technique used in scenario XYX. This way, a decoder can use the XYX CSI bitstream to recover from single packet losses and both XYX, XY_X CSI bitstreams to recover from double packet losses, without retransmitting redundant information for $\vec{x}_{t-1}$. Table 1 shows a brief description of the 4 scenarios. For each scenario, input space $Z$ denotes the known (side) information and output space $Y$ denotes the lost information that is going to be reconstructed.

| Scenario | packet delay | Input Space | Output Space |
|---|---|---|---|
| XY | 1 (lookahead) | $Z_t = [\vec{x}_t]$ | $Y_t = [\vec{x}_{t+1}]$ |
| XXY | 1 (lookahead) | $Z_t = [\vec{x}_{t-1}\vec{x}_t]$ | $Y_t = [\vec{x}_{t+1}]$ |
| XYX | 1 (jitter buffer) | $Z_t = [\vec{x}_{t-2}\vec{x}_t]$ | $Y_t = [\vec{x}_{t-1}]$ |
| XY_X | 2 (jitter buffer) | $Z_t = [\vec{x}_{t-3}\vec{x}_t]$ | $Y_t = [\vec{x}_{t-2}]$ |

**Table 1**. Brief description of the 4 scenarios. Note that the packet delay in the first two scenarios refers to the case of transmitting FEC data.

## 3. ESTIMATION

When no FEC bits can be sent to the decoder, estimation is the best option. The simplest form of estimation is repetition of the previous vector (for scenario XY, XXY) or interpolation (for scenario XYX, XY_X). Another form of estimation is Linear Estimation (*LE*), where the lost information $\vec{y}_t$ is estimated from a linear combination of the elements of the available information $\vec{z}_t$. The linear estimation $\hat{\vec{y}}_t$ is computed according to the formula:

$$\hat{\vec{y}}_t = \Sigma_{yz}\Sigma_{zz}^{-1}\vec{z}_t, \quad \Sigma_{yz} = \frac{1}{N}\sum_{t=1}^{N}\vec{y}_t\vec{z}_t^T, \quad \Sigma_{zz} = \frac{1}{N}\sum_{t=1}^{N}\vec{z}_t\vec{z}_t^T \tag{1}$$

where $N$ is the number of training set vectors. Linear estimation is commonly used in Predictive Vector Quantization [8].

The gaussianity assumptions made by linear estimation is a rather gross approximation of input-output space relationships. Generic models like GMMs are better suited for LSF joint distributions $p(\vec{y}, \vec{z})$. We implemented a GMM based regression function, the GMM Conversion Function [9] (*CF*). The training of *CF* is completed in two stages [9]: The first stage estimates the $Z$-space GMM via Expectation Maximization, and in the second stage a linear system of equations is solved for the $Y$-space means and the cross-covariance matrices.

## 4. CODING WITH SIDE INFORMATION

The optimal performance of CSI and/or estimation is upper bounded by the mutual information between the side information $Z$ and the lost information $Y$. Therefore, a mutual information measurement provides some insight on the number of bits that can be gained from a CSI scheme [6], or an estimator [7]. If the joint pdf of $[Z\ Y]$ is modelled with a GMM, mutual information $\mathcal{I}(\vec{z}, \vec{y})$ can easily be computed with stochastic integration [10], [7] of the mutual information formula:

$$\mathcal{I}(\vec{z}; \vec{y}) \approx \frac{1}{N}\sum_{n=1}^{N}\log\frac{p(\vec{z}_n, \vec{y}_n)}{p(\vec{z}_n)p(\vec{y}_n)} \tag{2}$$

where $\vec{z}_n$, $\vec{y}_n$ are drawn from the joined pdf $p(\vec{z}, \vec{y})$, and $p(\vec{z})$, $p(\vec{y})$ are the marginal distributions of $p(\vec{z}, \vec{y})$. The mutual information measurements in this paper were conducted with GMMs using diagonal covariance matrices, 1024 Gaussians and $10^6$ samples for the Monte Carlo integration.

The simplest form of CSI is residual coding. Let $\hat{\vec{y}}_t$ be the estimation of $\vec{y}_t$. Residual coding uses a form of VQ to encode $\vec{\epsilon}_t = \vec{y}_t - \hat{\vec{y}}_t$. In literature, residual coding is typically made using Linear Estimation [8].

In this paper we suggest to use the *CF* estimator for residual coding. The *CF* estimator capability of modelling complex nonlinear relationships between $Z$ and $Y$, provides a residual $\vec{\epsilon}_t$ that

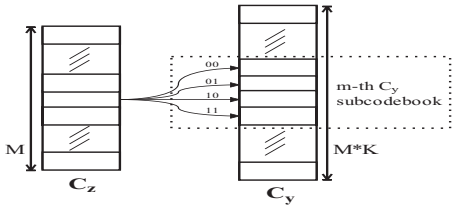


**Fig. 2**. VQ based CSI scheme

is more whitened, compared to the *LE* residual. In our knowledge, until now, nobody has used GMM based estimators like *CF* for residual coding of LSFs.

Even if the (unknown) *optimal* estimator was used, residual coding may not be able to benefit from all the mutual information between $Z$ and $Y$. Our measurements indicate that in scenario XY, the mutual information between the side information $\vec{z}_t$ and *CF* estimation residual $\vec{\epsilon}_{CF,t}$, is 2.51 bits, while the mutual information between $\vec{z}_t$ and $\vec{y}_t$ is 5.85 bits. The mutual information between the *LE* estimation residual and $\vec{z}_t$ is measured to be 2.82 bits. In other words, the *CF* estimation residual has nearly 43% of the initial mutual information between $\vec{z}_t$ and $\vec{y}_t$. Note also, that *CF* residual has less mutual information that *LE* residual, indicating that *CF* provides a better estimation than *LE*. Similar results were also obtained for scenarios XXY, XYX, XY_X.

We attempt to gain the mutual information between the estimation residual and the side information, by using a VQ based scheme for CSI that has been recently presented in [11]. It consists of two linked codebooks, a $Z$-space codebook with M codevectors, and a $Y$-space codebook. Each $Z$-space codevector is linked to a different $Y$-space subcodebook with $K$ entries, as shown in Figure 2. Therefore, the $Y$-space codebook has $M * K$ codevectors. Side information $\vec{z}_t$ is mapped to the nearest $Z$-space codevector, and lost information $\vec{y}_t$ is encoded and decoded according to the selected subcodebook. This scheme will be referred to as CVQ (Conditional Vector Quantization). Note that CVQ will be used to encode the estimation residual, and not $\vec{y}_t$.

We also tested CVQ to directly encode $\vec{y}_t$, but the results were worse than the results obtained from a simple VQ of the linear estimation (*LE*) residual. However, as $M$ increased from 32 to 512, the results were improving, indicating that a much higher $M$ is required for a proper modelling of the input-output space relationship. The removal of a simple rotational relationship between $Y$-space and $Z$-space by *LE* was enough to let CVQ benefit from the (remaining) mutual information.

## 5. EXPERIMENTS

For the scenarios presented in Section 2, two modes will be evaluated for the recovery of lost LSFs:

- estimation mode (no data transmission), using the following estimators:
  - Linear Estimation (*LE*)
  - GMM Conversion Function (*CF*)
- CSI mode (with data transmission), using the following methods:
  - VQ of the *LE* Residual (*VQLE*)
  - VQ of *CF* Residual (*VQCF*)
  - CVQ coding of *LE* residual (*CVQLE*)
  - CVQ coding of *CF* residual (*CVQCF*)

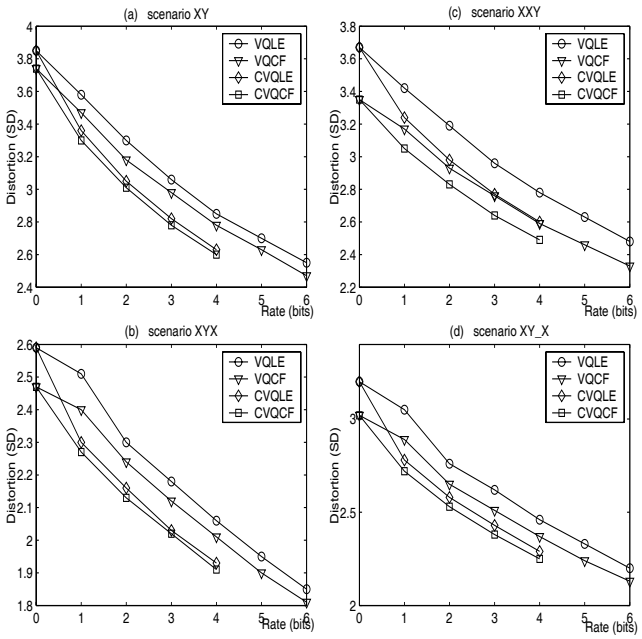


**Fig. 3**. CSI Rate-Distortion curves for each scenario and each CSI method. Note that *CVQLE* and *CVQCF* uses $M$=256.

The experiments were conducted using the whole training set of TIMIT database for training and the whole testing set of TIMIT for testing. A sequence of LSF vectors (with 10 LSFs/frame) was extracted using analysis frames of 25ms at a rate of 50 frames/sec (5 ms overlap). For each scenario, all available $Z$-space and $Y$-space features were collected from the LSF sequence, excluding silent frames. The AR filter was computed from the full narrowband (0-4 kHz) signal with the autocorrelation method using preemphasis ($\mu = 0.95$). The used Spectral Distortion measure is given by:

$$\mathcal{D}(X_t, \tilde{X}_t) = \frac{1}{\pi} \int_0^{\pi} \left( 20 \log_{10} \frac{|X_t(e^{j\omega})|}{|\tilde{X}_t(e^{j\omega})|} \right)^2 d\omega \qquad (3)$$

where $|X_t(e^{j\omega})|$,$|\tilde{X}_t(e^{j\omega})|$ is the original spectrum and the reconstructed spectrum respectively. Simple averaging was used for the evaluation over the test-set.

The linear system that has to be solved for *CF* training [9], is ill-conditioned in scenarios XXY, XYX, XY_X, where $Z$-space has 20 dimensions. A dimensionality reduction via PCA (Principal Component Analysis) to the 18 strongest dimensions was used to avoid the ill-conditioning. This indicates the existence of redundancy in $Z$-space. In all scenarios, the *CF* estimator had 128 Gaussians, and CVQ had $M$=256 input space codevectors. The size of the $Y$-space codebook in CVQ was constricted to have at most 4096 vectors. Therefore, $K = \{2^1, 2^2, 2^3, 2^4\}$.

## 6. RESULTS

The experiment results are shown in Figure 3. Rate-Distortion measurements are plotted for each scenario and each CSI method (*VQLE*, *VQCF*, *CVQLE*, *CVQCF*). Since each CSI method is associated with an estimation method, it is convenient to represent the estimator performance as the performance of the corresponding CSI scheme at the rate of 0 bits/frame (no FEC transmission). This allows a direct comparison of CSI techniques and estimation methods, in terms of distortion.

Regarding estimation methods, *CF* outperforms *LE* in all scenarios, especially in scenario XXY, where 3.35 dB were obtained. These results are similar to those presented in [3]. Having the performance of *CF* in scenario XXY as a reference, jitter buffer provides an improvement of 0.35-0.90 dB when *CF* estimation is used.

In all scenarios, distortion can be significantly reduced with a few bits. Regarding CSI techniques, it is clearly seen that VQ-based residual coding can benefit from a better estimator, i.e. *VQCF* outperforms the widely used *VQLE* at least 0.5 bit, while in scenario XXY the gain is greater than 1 bit. The clear advantage of *VQCF* over *VQLE* in scenario XXY suggests using a "predictive" VQ technique based on *CF* estimation for "transparent" residual coding of LSFs [8]. On the contrary, CVQ-based residual coding is less dependent on the estimator and provides similar performance for both estimators in all scenarios except XXY. Furthermore, CVQ always benefits from the available mutual information between the residual and the side information, providing an improvement of 1 bit over the widely used *VQLE*, and a gain of 0.75-1 bit over *VQCF*.

For single packet losses, just 4 bits/frame of FEC data encoded with *CVQLE* provide a 42% distortion reduction (-1.42 dB) over the best "predictive" estimation (3.35 dB using *CF* in scenario XXY), and a 21% distortion reduction (-0.54 dB) over *CF* estimation in scenario XYX. For double packet losses, Figure 3d shows only the distortion from the reconstruction of the first lost vector. Note, that the recovery of the second lost vector is made from the *reconstructed* first vector as stated in Section 2. However, our measurements showed that when this *cascaded* form of CSI recovery is made, the second lost vector is reconstructed with less distortion than the first lost vector. Therefore, double packet losses can be recovered with at least 25% distortion reduction (-0.75 dB) over *CF* estimation, and at least 32% distortion reduction (-1.10 dB) over the best "predictive" estimation, using *only* 4 additional bits.

Since both CVQ-based methods have the same performance, and *CVQLE* is more simple than *CVQCF*, we chose *CVQLE* for informal subjective testing, assuming a 2 frame (40ms) jitter buffer, and using 4 bits/frame of FEC data for scenario XYX and 4 bits/frame for scenario XY_X. These 8 bits/frame of FEC data were used to recover from 1 or 2 packet losses as stated in Section 2. LSFs were computed from the speech signal, according to Section 5, and speech was inverse filtered using the original AR filter parameters. An amount of 5%-25% losses was introduced to the LSF vector sequence, constricted to generate either 1 or 2 sequential losses. The proposed *CVQLE* methods were compared to simple interpolation. Speech signal was then synthesized from the original excitation and the reconstructed LSFs. Listening tests showed that envelope related artifacts were fewer and milder with *CVQLE*. Sample utterances from this work can be found in:
http://www.ics.forth.gr/~jagiom/icassp2005.

## 7. CONCLUSION

The problem of LSF reconstruction from packet losses in VoIP was addressed. Four scenarios of lost/received spectral envelopes were examined, the first two depend on the past LSF vector(s) to reconstruct the current LSFs, and the last two scenarios depend on past and future LSF vector(s) taking advantage from the existence of jitter buffer. Two estimation methods were examined along with several suggested Coding with Side Information (CSI) schemes, in terms of spectral distortion. We found that a few bits for CSI based Forward Error Correction (FEC) provide a significant distortion reduction, compared to estimation. Furthermore, the utilization of the jitter buffer introduces a clear advantage over "predictive" LSF reconstruction. In future work we will examine CSI methods for recovery and error correction of VoIP data.

## 8. REFERENCES


[1] Roch Lefebvre, Philippe Gournay, and Redwan Salami, "A study of design compromises for speech coders in packet networks," in *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, May 2004.

[2] Eitan Altman, Chadi Barakat, and Victor M. Ramos R., "Queueing analysis of simple fec schemes for ip telephony," in *Proceedings of the IEEE Infocom 2001*, April 2001, vol. 2.

[3] Jonas Lindblom, Jonas Samuelson, and Per Hedelin, "Model based spectrum prediction," in *IEEE Workshop on Speech Coding*, Delaway, USA, 2000.

[4] Martin R., C. Hoelper, and I. Wittke, "Estimation of missing lsf parameters using gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process*, Salt Lake City, USA, 2001.

[5] Kohler M.A. and Yarlagadda R.K., "Markov chain prediction for missing speech frame compensation," in *IEEE Workshp on Speech Coding*, Delavan, USA, 2000.

[6] Robert M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 19, no. 4, July 1973.

[7] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, Institute of Communication Systems and Data Processing (IND), Aachen, Germany, 2002.

[8] J. Skoglund and J. Linden, "Predictive VQ for noisy channel spectrum coding: AR or MA?," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 1351–1354.

[9] Y. Stylianou, O. Cappe, and Eric Moulines, "Continuous propabilistic transform for voice conversion," *IEEE Trans. Speech Audio Processing*, 1998.

[10] 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. Acoust. Speech Sign. Process*, Orlando USA, 2002.

[11] Yannis Agiomyrgiannakis and Yannis Stylianou, "Combined estimation/coding of highband spectral envelopes for speech spectrum expansion," in *ICASSP*, Montreal, Canada, May 2004.