Quality-Complexity Trade-off in Predictive LSF Quantization

Davorka Petrinovic, Davor Petrinovic
davorka, davor.petrinovic@fer.hr

Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia

Abstract

In this paper several techniques are investigated for reduction of complexity and/or improving quality of a line spectrum frequencies (LSF) quantization based on switched prediction (SP) and vector quantization (VQ). For switched prediction, a higher number of prediction matrices is proposed. Quality of the quantized speech is improved by the prediction multi-candidate and delayed decision algorithm. It is shown that quantizers with delayed decision can save up to one bit still having similar or even lower complexity than the baseline quantizers with 2 switched matrices. By efficient implementation of prediction, lower complexity can be achieved through use of prediction matrices with reduced number of non-zero elements. By combining such sparse matrices and multiple prediction candidates, the best quality-complexity compromise quantizers can be obtained as demonstrated by experimental results.

1. Introduction

A combination of switched prediction [1], [2] and either split or multi-stage vector quantization has frequently been used in modern speech coders for LSF quantization. In such memory-based quantization scheme, the prediction of an input LSF vector is calculated first by selecting one prediction matrix among the set of previously designed matrices. The difference between the original and the predicted vector is then vector quantized. In split VQ (SVQ) that is employed in this work, a single vector is divided into few subvectors and each subvector is quantized independently. Bit allocation between switched predictor and VQ codebooks determines the trade-off between quantizer complexity and quality. For the fixed total number of bits per frame, the highest quality is achieved if only one bit is assigned to the predictor [1]. Such quantizer is most frequently used and will be called the baseline quantizer. Nevertheless, for transparent quantization, this bit allocation also results with the highest number of operations since VQ is rather complex.

For applications in which LSF quantization should be less computationally demanding, complexity of the VQ can be reduced by allocating more bits to the predictor. That is the direction followed in this study. Although computationally more efficient, this approach introduces some distortion compared to the baseline, 2 switched matrix case. Therefore, in the first part of this paper, it will be presented how the quality of the SP-SVQ scheme with higher number of switched matrices can be improved by applying so called multi-candidate and delayed decision algorithm. This algorithm (also known as M-L search) represents a very important part in multi-stage vector quantization [3]. Similar to the work presented in [4], it is employed in this study for determining the best prediction matrix among several prediction candidates according to results after vector quantization. It will be shown that compared to the baseline quantizer, proposed quantizers with 4, 5 or 6 bits allocated for predictor result with comparable quality and lower complexity, but can also offer much higher quality if more then 4 candidates are used.

In the second part of the paper, a recently reported approach [5] for additional reduction of the number of operations for the studied quantization scheme will be shortly reviewed. This approach is aimed towards simplifying the prediction part since the prediction may also become an intensive task due to increased number of matrices. Simplification is achieved by replacing a larger number of matrix elements with zeros based on some selection criterion. Certainly, prediction calculation is then performed only on the remaining non-zero elements. Although more efficient, the simplification introduces some additional distortion.

Therefore, in the final part of the paper two independently analyzed approaches are combined in order to find the best quality-complexity trade-off. It will be shown that LSF quantizers with 16, 32 or 64 sparse prediction matrices and delayed decision achieve the same quantization performance as the baseline systems with only 2 full matrices but at less then half of the operations for quantization.

2. Simulation setup

It is evident from introduction that this paper presents several techniques. Each of them will be shortly explained and its role and the obtained results discussed in the sections that follow. The aspects of simulation setup and quantizer evaluation that are common to all will be summarized here.

All of the LSF quantizers were designed based on a speech database containing 20 minutes of speech (119960 LSF vectors), spoken by 2 male and 2 female speakers. Different speech material obtained from 7 male and 3 female speakers (59980 vectors) was used for their evaluation. Both databases were created by sampling at 8kHz and using 10th order robust LPC analysis. The analysis was performed on 25ms speech segments with two different LPC frame rates: $FR = 100$ and 50 frames/s. Switched prediction with $b = 1$ up to 6 bits for predictor (number of matrices $N = 2^b$) was combined with split VQ (4-6 split). The total number of bits per frame, B, was varied from 19 to 23 bits at $FR = 50$ frames/s and from 15 to 19 at $FR = 100$ fr./s. The SVQ codebook sizes for a certain value of B and b were determined based on the minimum spectral distortion.

Quantization performance was evaluated by the average log spectral distortion, $\overline{SD}$, and percentage of the outlier frames with $\overline{SD}$ greater then 2 dB. Quantization complexity, denoted with K, was calculated as the total number of operations (multiplications, additions and comparisons) for prediction and VQ required for quantization of a single input vector. All results presented here were obtained based on vectors from the evaluation speech database.

3. SP-SVQ with delayed decision

In a classical SP-SVQ LSF quantization scheme [1], first order vector linear prediction of the input LSF vector $\mathbf{x}(n)$ is performed first. It is based on a preceding quantized vector $\hat{\mathbf{x}}(n-1)$ and a predictor composed of N switched matrices $\mathbf{A}_i$, $i = 1,...,N$ determined in the design process. Prediction residual $\mathbf{e}(n)$ is found as a difference between the original and the prediction vector and it is then vector quantized. The reconstructed LSF vector $\hat{\mathbf{x}}(n, m_n, j_n)$ is obtained by a similar process described with:

$$\hat{\mathbf{x}}\,(n,m_n,j_n) = \mathbf{A}_m\hat{\mathbf{x}}(n-1,m_{n-1},j_{n-1}) + \hat{\mathbf{e}}_j(n) \qquad (1)$$

where $\mathbf{A}_m$ is a switched matrix chosen as "best" for prediction and $\hat{\mathbf{e}}_j$ represents the "closest" quantized value of the prediction residual, both for the n-th LSF vector.

Index j is a result of minimization of the weighted Euclidean distance (WED) commonly used as distortion measure in VQ. On the other hand, index m is normally obtained from the minimal squared or weighted squared Euclidean norm of the prediction residual (i.e. residual energy) and has no direct relation to the final result after quantization.

Since there is no guarantee that the best prediction residual will produce the best overall result it is reasonable to determine the best prediction according to the combined prediction-quantization result (i.e. delayed decision). To avoid prohibitively complex exhaustive search of all possible combinations for indexes m and j, quantization is performed on a group of residual vectors (called prediction candidates) ranked according to the minimal squared Euclidean norm criterion.

To evaluate this approach, a group of quantizers with different parameters b and number of prediction candidates, M, was designed. Their performance at $FR = 50$ fr./s with $B = 21$ is presented in Figure 1 while quantizers with $B = 16$ ($FR = 100$) are shown in Figure 2. Different lines connect quantizers of the same total number of bits for predictor while each symbol type represents a quantizer of different value for M.

As expected, complexity of quantization is increased if more then one prediction residual ,is vector quantized, but on the other hand performance gain is significant. The relative improvement of quality is inversely proportional to the number of candidates. It is interesting to compare the baseline quantizers with $b = 1$ (marked with circles) to those with higher number of prediction matrices and higher M. Although quantizers with higher b and only one candidate (i.e. classical approach) offer substantial reduction of complexity (except for the quantizer with $b = 6$ at $FR = 100$, Fig. 2), they suffer from certain loss of quality. As can be seen, this can be very efficiently compensated for by increasing the number of candidates to 2 or 4, thus obtaining quantizers with lower distortion and still less computation then the baseline.

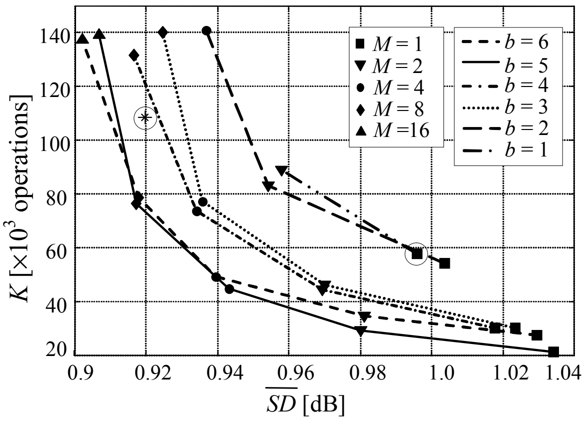

Figure 1. Number of operations vs. $\overline{SD}$ for SP-SVQ quantizers with delayed decision, $B = 21$, $FR = 50$ fr./s

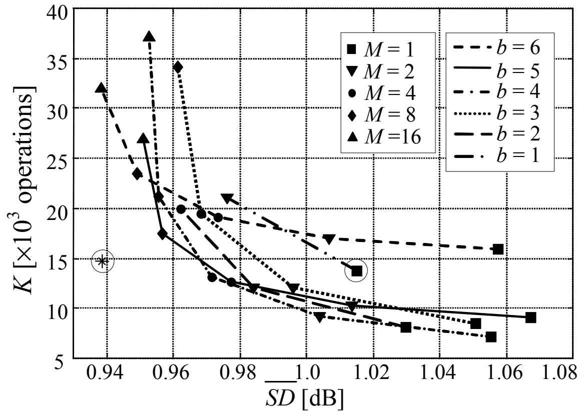

Figure 2. Number of operations vs. $\overline{SD}$ for SP-SVQ quantizers with delayed decision, $B = 16$, $FR = 100$ fr./s

On the other hand, if quantization quality is of primary interest, quantizers with $b = 5$ or 6 bits for predictor and more then 4 candidates represent very good options. For such quantizers the reduction of $\overline{SD}$ compared to the baseline quantizer is close to 0.1 dB, at $FR = 50$.

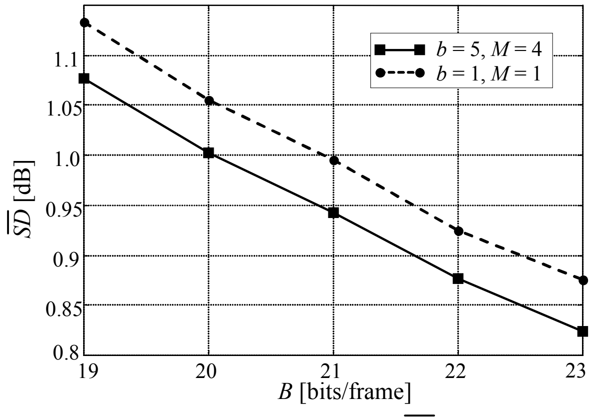

Figure 3. Number of bits per frame vs. $\overline{SD}$ for baseline and quantizers with delayed decision, $FR = 50$ fr./s

Furthermore, if the positions of the baseline quantizers having one more bit per frame ($B = 22$ and $B = 17$) are observed (circled stars), it is obvious that this technique can actually save up one bit of the total bit-rate sometimes even with lower overall complexity.

To demonstrate that the improvement offered by delayed decision algorithm is consistent over all considered bit-rates, the baseline quantizers and those with 4 candidate and $b = 5$ are shown in Figure 3 for $FR = 50$. It can also be noted that all multi-candidate quantizers in Figure 3 have between 10 and 30% lower complexity then the baseline quantizers.

4. Quantizers with sparse predictors

Prior research used as basis for this paper has been focused on computationally efficient SP-SVQ spectrum quantization techniques obtained by modifying the prediction part. Approach that has proved the best in that sense is allocating more bits for predictor and additionally reducing the number of operations for prediction by replacing some nonzero elements in prediction matrices with zeros (i.e. matrices become sparse). This approach will be shortly described next.

Two issues had to be addressed: first, which matrix elements to set to zero and second how to calculate the optimal values of the remaining nonzero elements. An algorithm has been developed [5] for calculating optimal sparse predictors that also incorporates the optimal criterion for element reduction. In the algorithm, no restrictions on the sparse matrix structure (zero-nonzero element arrangement) are made, so each row of any prediction matrix can have an arbitrary number of nonzero elements on arbitrary positions and is therefore treated independently of others. That way prediction matrix rows are actually scalar predictors of the order equal to the number of nonzero elements. Replacing nonzero elements with zeros in any matrix row results with the increase of prediction residual energy since the order of that scalar predictor is reduced and furthermore, its element values are no longer optimal. If the element values are recalculated (i.e. made optimal) [5] for the reduced prediction row, the increase of residual energy would be less. Thus, by the proposed criterion for element reduction, the elements to be replaced with zeros are selected in a way that the increase of the residual energy after re-computation of the prediction matrix row is the least. In a switched predictor scheme, those elements are searched across all rows of all switched matrices.

Design of sparse switched predictors is an iterative procedure resulting with a new set of optimal matrices in each iteration that, in addition, have less and less nonzero elements as iterating progresses. Each iteration is characterized by a number called the element reduction factor, η, defined as the ratio between total number of elements in all N full switched matrices of dimension $k \times k$ and total number of nonzero elements S in all sparse matrices:

$$\eta = N \cdot k^2 / S \qquad (2)$$

In the quantizer design the optimal sparse predictor and the SVQ codebooks are first designed in the open-loop and then refined by a certain number of closed-loop iterations as in [1]. Extensive simulations have been performed and SP-SVQ quantizers with sparse predictors (called sparse quantizers for short) for different values of η, b, B and at both frame rates have been designed. Only the most interesting results will be presented here. Performance of a group of proposed quantizers with element reduction factors η equal to 3, 6 and 10 and with $b = 4, 5$ and 6 is shown in Figure 4. These quantizers are realized at $FR = 100$ fr./s with total of $B = 16$ and 17 bits/frame. Besides these sparse quantizers, full quantizers ($\eta = 1$) with $b = 1$ (circled stars) and $b = 4$, 5 and 6 (triangles) are also shown for comparison since the former result with the highest quality while the latter group illustrates the influence of higher number of switched prediction matrices.

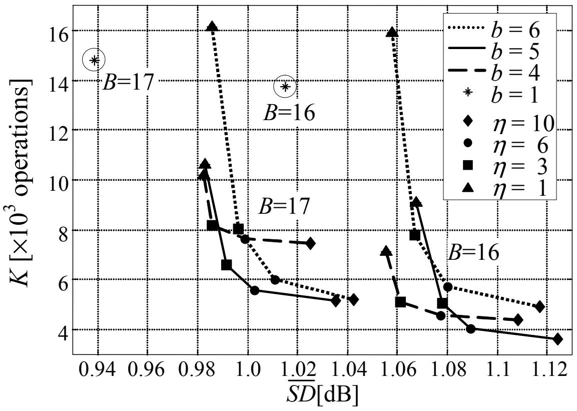

Figure 4. Number of operations vs. $\overline{SD}$ for sparse and full quantizers with different b and η, at $FR = 100$ fr./s

It is obvious from Figure 4 that the increase of the number of (full) matrices (compare stars with triangles) simplifies computation in most cases, but can also make it even more complex ($b = 6$). The increase of distortion is also evident. Technique with sparse matrices reduces K for all values of b, most effectively for $b = 5$ and 6. As for the element reduction factor, it can be noticed that quantizers with $\eta = 3$ (squares) offer the steepest reduction in number of operations relative to the full matrix cases as well as the highest quality of all sparse quantizers. If even lower complexity is required, quantizers with $\eta = 6$ can be used. Those with $\eta = 10$ result with the highest distortion and insignificant decrease of K compared to cases with $\eta = 6$ but may be used for implementations with limited memory capacity.

The choice of the best quantizer concerning b depends on the total number of bits per frame and can be either 4, 5 or 6. Complexity of quantizers with $b < 4$ is higher so they were not further discussed. The achieved reduction of K for sparse quantizers compared to the baseline quantizers with $b = 1$ (stars) also depends on B and is higher for higher values of B. For $FR = 50$ fr./s the increase of the number of matrices by itself greatly reduces the total number of operations for quantization, but sparse quantizers with $\eta = 3$ offer some additional reduction of K (see Figure 5, for $M = 1$) .

5. Sparse SP-SVQ with delayed decision

From discussion in the previous sections it is obvious that delayed decision improves quantization quality on the account of increased complexity. On the other hand, the technique with sparse matrices reduces prediction complexity but introduces some additional distortion. Combining the two approaches comes as a logical solution for achieving the best complexity-quality trade-off.

To evaluate the effectiveness of the proposed technique a large number of LSF quantizers was designed. Figures 5 i 6 present the results obtained for $B = 21$ bits/frame at $FR = 50$ fr./s and $B = 16$ at $FR = 100$ fr./s respectively. Only the quantizers with values of b that result with best complexity-quality trade-off are shown. Others may be less complex but exhibit higher distortion or vice versa.

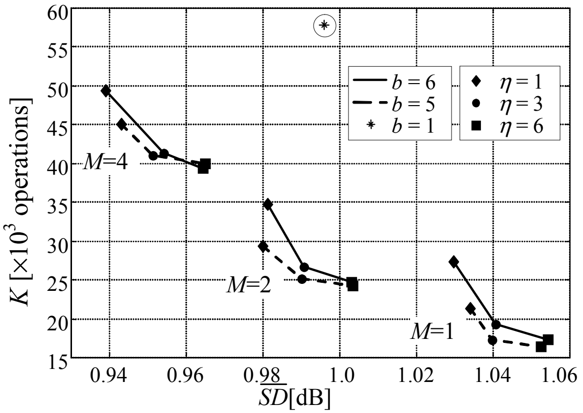

Figure 5. Number of operations vs. $\overline{SD}$ for different quantizers, $B = 21$, at $FR = 50$ fr./s

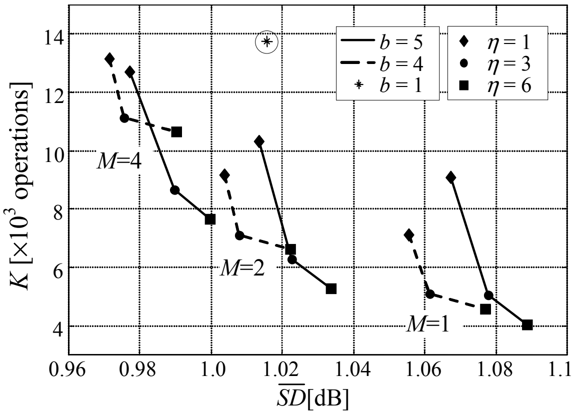

Figure 6. Number of operations vs. $\overline{SD}$ for different quantizers, $B = 16$, at $FR = 100$ fr./s

The sparse and the full quantizers obtained without delayed decision are located at the bottom right side of both figures ($M = 1$). The effect of multiple prediction candidates is obvious: the quantizers move towards lower values of distortion but require somewhat higher number of operations for quantization. Proposed quantizers with $M = 2$ achieve quality comparable to the baseline quantizers using less then half of the number of operations. Quantizers with $M = 4$ are still less complex but offer reduction of $\overline{SD}$ of 0.02-0.05 dB depending on other quantizer parameters.

The percentage of outliers $p_{2\text{dB}}$ obtained for all presented techniques will be discussed based on the results shown in Figure 7. As it was reported previously [1], the increase of the number of matrices increases $p_{2\text{dB}}$. (compare the baseline quantizer marked with a circle to full quantizers, $M = 1$). Technique with sparse prediction matrices does not increase the outliers more then it is a consequence of the increased distortion. Since outliers are perceptually important, the best quantizer should have the lowest $p_{2\text{dB}}$ for any resulting $\overline{SD}$. As can be seen in the figure, sparse quantizers with delayed decision exhibit such behavior while $p_{2\text{dB}}$ of the full multi-candidate quantizers is generally higher.

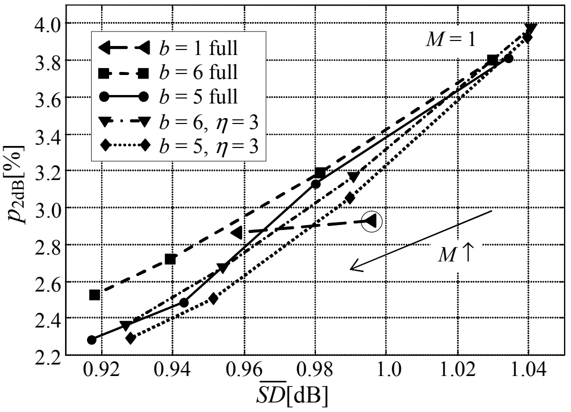

Figure 7. Percentage of outliers vs. $\overline{SD}$ for different quantizers, $B = 21$, at $FR = 50$ fr./s

6. Conclusion

Contrary to the previously reported results that the highest quality of SP-SVQ quantizers can be achieved with only 2 prediction matrices, it is shown in this paper that much better quality can be obtained using 32 or 64 matrices and delayed decision algorithm. If the optimal sparse matrices are used in conjunction with delayed decision, the resulting quantizers achieve the quality comparable to the baseline quantizers but with less then half of complexity.

7. References

[1] S. Wang, E. Paksoy, A. Gersho, "Product code vector quantization of LPC parameters", in *Speech and Audio Coding for Wireless and Network Applications*, editors: B. S. Atal, V. Cuperman, A. Gersho, Kluwer Acad. Pub., pp. 251-258, 1993.

[2] M. Yong, G. Davidson, A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction", *Proc. ICASSP*, vol.1, pp. 402-405, 1988.

[3] W.F. LeBlanc, B. Bhattacharya, S.A. Mahmoud, V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4kb/s speech coding", *IEEE Transaction on Speech and Audio Processing*, vol. 1, no. 4, pp. 373-385, 1993

[4] E. Shlomot, "Delayed decision switched prediction multi-stage LSF quantization", *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, 1995, pp. 45-46

[5] D. Petrinović, D. Petrinović, "Sparse vector linear prediction with near-optimal matrix structures", *Proc. of International. Workshop on Image and Signal Processing and Analysis*, Pula, Croatia, pp. 235-240, 2000.