

HARMONIC POST-PROCESSING OF SPEECH SYNTHESIZED BY STOCHASTIC CODERS

I. M. Trancoso *, J. M. Tribolet *

ABSTRACT

High quality speech coding at medium-to-low bit rates is presently one of the major goals in speech research. Stochastic coding represents an important step towards this objective. Yet, the quality of synthetic speech is still not always good enough. A subjectively important part of the distortion may arise from imperfect reproduction of voiced regions, where the harmonic structure is not so well marked in the synthetic as in the original speech signal. Post-processing of synthetic signals using harmonic modelling arises as a natural solution to reduce this distortion. The disadvantages of this method in terms of additional delay, complexity and dependency on high precision pitch detectors can be well counterbalanced by the higher quality of resynthesized speech signals in voiced regions.

INTRODUCTION

The impact of vector quantization on speech coding cannot be denied. Codebooks have penetrated almost every type of method in the medium-to-low bit rate range, being used to code side parameters such as LPC coefficients, as well as residual information whether in the time or in the frequency domain. One of the many examples, nowadays on the front line of research, is the stochastic or code-excited linear predictive (CELP) coder (ref 1).

This coder derives its first designation from the nature of the codebook of 1024 sequences which is used as an excitation to two recursive time-variant linear filters. Their function is to impose the pitch and formant structures that characterize speech signals. The use of white gaussian innovation sequences for the excitation can be questioned as they do not provide an adequate model for all types of speech sounds. Stop bursts in unvoiced consonants and the onset of voiced regions are two examples of sounds where the statistics of the prediction residual signal do not match the ones chosen for the codebook. In order to take these regions into account in the codebook design one could, for instance, use a subset of sequences in which spikes were randomly added to noise, in the first case, and another subset consisting of single pulses for each sequence, eventually with some background noise, in the

* INESC, R. Alves Redol, 9, 2, 1000 Lisbon and IST, Lisbon, Portugal

second case. The use of disjoint subsets of codewords with different statistics may lead to higher quality schemes, despite the oversimplified models we have suggested, while at the same time making the codebook more amenable to fast search procedures.

The problem of reproducing unvoiced and transition regions, however, is not our major concern in this paper. Here, we are mostly interested in voiced regions, which have been long recognized as playing a dominant role as far as the overall synthetic speech quality of the coder is concerned. In stochastic coders, periodicity is introduced by the pitch or long-term predictor, which is generally robust and efficient. The codeword selection criterion, however, is based on a short-time matching strategy, and is not designed to reinforce the periodic structure of voiced regions. The inter-harmonic distortion that results from this fact can be particularly unpleasant when the pitch predictor fails to do its job (ref 2). In fact, voiced segments are not always very close to periodic and, in these cases, the excitation derived from the codebook becomes relatively more important, as if trying to "correct" the deficiencies of the two combined predictors. A codebook with only 1024 words, however, can only do this to a limited extent. The reduction of this type of distortion is our major goal in this paper. Though arising from different sources, this is not a problem exclusive to stochastic coders and, as in many other cases, the solution that was adopted involves a post-filtering technique.

HARMONIC POST-PROCESSING

Harmonic analysis/synthesis (ref 3) arises as an ideal tool to solve the problem of inter-harmonic distortion. In fact, by computing the amplitudes and the phases of the harmonics of the distorted speech signal and resynthesizing it as a sum of harmonics, one removes a large part of the energy contents in between them. This type of post-processing should take place in the receiver stage but, since stochastic coding is an analysis-by-synthesis method, there is also motivation to include it in the transmitter stage as well, which has the advantage of the memory contribution of the two predictor filters being computed on the basis of post-filtered signals.

Harmonic analysis is restricted to voiced regions. The pitch and voicing information implicit in this analysis may be derived in the receiver, in order to avoid transmitting further information. It may be argued that harmonic post-processing introduces a dependency on these parameters in what was originally a very robust coder, as far as they were concerned. Robustness can be increased by making the voiced/unvoiced decision in the transmitter, where the signal-to-noise ratios of the synthetic signals, without and with

post-filtering, can be compared. A decision of unvoiced could be made whenever the difference between the two ratios exceeds a certain threshold. This scheme, however, calls for the transmission of one extra bit, for every frame that is post-filtered.

In addition to this robustness problem, post-filtering also introduces some more delay and complexity to stochastic coding. Let us first deal with the complexity problem. Harmonic analysis involves only the computation of a short-time Fourier transform and the solution of a system of linear equations obtained by minimization of a m.s.e. criterion. Analysis frames are typically 32 msec long, in order to include several pitch periods. Harmonic synthesis, on the other hand, can assume two major forms: the simplest one consists of an inverse Fourier transform and overlap-add; the more complex one is used for frame spacings larger than 15 or 20 msec, and includes amplitude and phase interpolation schemes in order to cope with fast varying pitch frames. In the context of such a complex coder as is the stochastic one, the additional complexity brought by harmonic post-processing is, therefore, hardly relevant.

The integration of post-filtering in stochastic coders can be done in several ways, depending on the maximum delay feasible in each particular application. The solution that adds virtually no delay consists in, for every 5 msec frame that has been synthesized by stochastic coding, computing the harmonic coefficients in the region that includes some previous frames and the present one, and resynthesizing only the latter. The harmonic analysis, therefore, is not centered in the present frame and hence the determined coefficients are not the most adequate ones. A number of alternative solutions exist that involve more delay but for which, on the other hand, the center of the harmonic analysis is closer to the frame being resynthesized.

The choice of the appropriate frame spacing in harmonic post-filtering is obviously related to the delay. Larger frame spacings call for interpolation schemes and also longer delays. Post-filtering can be implemented with frame spacings as short as one would like but, in practice, there is no need to update harmonic coefficients more frequently than about once every 10 msec.

SIMULATION RESULTS

The simulation results mentioned in this section were not meant to represent the optimum solution for harmonic post-filtering, but only to illustrate the type of performance that can be expected from the combined stochastic/harmonic coder. Our experiments involved a data base consisting of 24 sentences by male and female speakers, amounting to a total of 65 seconds,

approximately. In these experiments, harmonic post-filtering was only included in the receiver and the frame spacing was 15 msec. In informal listening tests, the perceptual speech quality of the resynthesized speech was judged significantly better than without post-filtering. The inter-harmonic distortion is greatly reduced, though at the cost of a slightly muffled quality which can be attributed to the lower energy of the harmonic coefficients derived from the synthetic signal, as compared with the ones derived from the original. Objective snr measurements thus show a very small degradation with post-processing (less than 0.5 dB, on the average). The better performance of the hybrid coder is illustrated in Fig. 1, which shows spectrograms of one of the speech utterances by a female voice. The original is shown in (a). (b) refers to the synthetic signal produced by stochastic coding (unquantized predictor parameters) and (c) to the post-processed signal. The inter-harmonic noise that was so clear in (b) is significantly reduced in the last spectrogram.

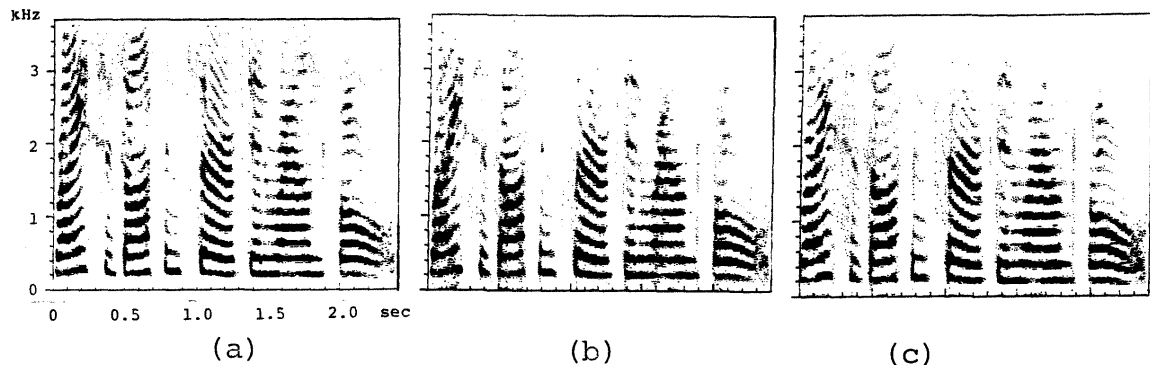


Fig. 1

CONCLUSIONS

Harmonic post-processing can be a solution to the problem of inter-harmonic distortion in stochastic coders, despite introducing extra delay, complexity and some robustness problems which may be dealt with by the transmission of further information. In our opinion, this type of approach should also be viewed as a stepping stone towards a more distant goal: the "ideal" low bit rate, high quality coder, which integrates stochastic and deterministic elements and manages to combine the robustness of present CELP coders with the noise-free performance of harmonic coders in voiced regions.

REFERENCES

1. M. R. Schroeder and B. S. Atal, Proc. Int. Conf. Commun., 1610 (1984)
2. I. M. Trancoso, L. B. Almeida and J. M. Tribolet, Proc. Int. Conf. on Acoustics, Speech and Signal Proc., 1709 (1986)
3. L. B. Almeida and J. M. Tribolet, IEEE Trans. Acoustics, Speech and Signal Proc., vol. 31, 664 (1983)