CONSIDERATIONS IN THE SELECTION OF AN OBJECTIVE MEASURE TO ASSESS THE QUALITY OF SPECTRAL CODING METHODS

H.R. Sadegh Mohammadi* and W.H. Holmes**

* Electrical Eng. Research Center, Jahad Daneshgahi, IUST University, Iran
** Electrical Engineering School, University of NSW, Australia

ABSTRACT - In low rate speech coders based on the linear prediction method, the quality of the synthesized speech is highly affected by the amount of distortion arising from the spectral coding stage. In this study we investigate two basic models for the evaluation of the quality of the short-term spectrum quantization. The advantages and disadvantages of each model are studied. Moreover, the difficulties in comparing the results of different published studies are found to be because of five groups of incompatibilities. We demonstrate the differences between the results of the assessments based on these models for several spectral coding methods using vector quantization.

INTRODUCTION

The spectral coding stage is an important part of any low rate speech coder that uses the linear prediction model. To develop and test a new spectrum quantization method or to compare several spectral coding techniques, it is necessary to assess the quantized spectra and compare them with the original spectral envelope.

Since the human auditory system is only sensitive to acoustic signals in a certain range of frequencies, both the quantized and unquantized short-term spectra are not perceptible in the absence of complete speech signals, which have other attributes apart from the short-term spectrum. Therefore, the quantized spectral envelopes must be converted to speech signals using an appropriate model if a subjective quality measure is to be evaluated. Similarly, objective measures which have been employed for spectral coding evaluation also use such conversions. It is desirable to have quality measures that are sensitive only to spectrum quantization distortion and not to other possible degradations caused by the model that converts the quantized spectral envelope to the speech signal.

This paper deals with the sources of difficulties in comparing the distortions from different spectral coding methods and provides an insight into the factors that should be taken into account in the development of any spectral coding assessment experiment or in the selection of a proper short-term spectrum quantization method.

The structure of this paper is as follows. The next section explains two basic models for the assessment of spectral coding using either objective or subjective measures. It also describes the benefits of each model. Then the sources of incompatibility between various experiments reported by different researchers will be addressed. These incompatibilities are believed to have major effects in biasing the results of quality assessments toward one method or another. Then an experimental simulation based on different basic models using a CELP speech coder is described, including a comparison between several vector quantization methods for spectral coding using two objective measures based on the two basic models. Finally, conclusions are presented.

BASIC MODELS FOR SPECTRAL CODING ASSESSMENT

Various basic models can be applied to convert the quantized short-term spectrum to a frame of speech signal. As will be discussed later, almost all reported results in the published articles about different spectral coding methods use such basic models, even though this is often not very explicit. Here, two basic models are outlined and their advantages are reviewed briefly.

The first model assumes that the residual signal created by filtering the original speech signal with the analysis filter $A(z)$ is then passed through the quantized synthesis filter $\hat{A}^{-1}(z)$ (all-pole filter).

Figure 1 shows the overall structure of this process. In this model, the spectral quantization assessment is performed independently from the effects of any other artifact caused by other stages of a typical speech coding system.

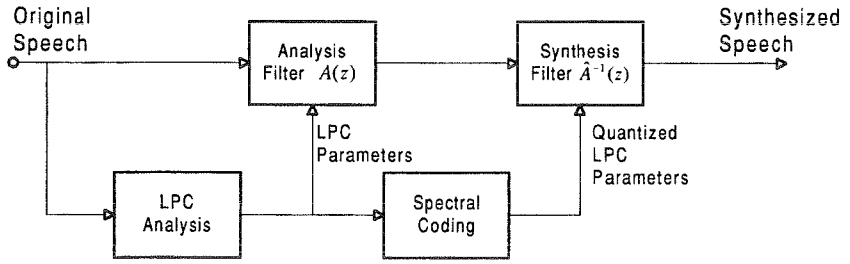

Figure 1. The first basic model for spectral coding assessment

The second model considers that a particular speech encoder has been chosen and that the spectral coding method to be evaluated is embedded in it (Figure 2). In this model, the parameters of the spectral envelope extracted from the original and the synthesized speech signals are compared for evaluation of the degradation due to the spectral quantization method.

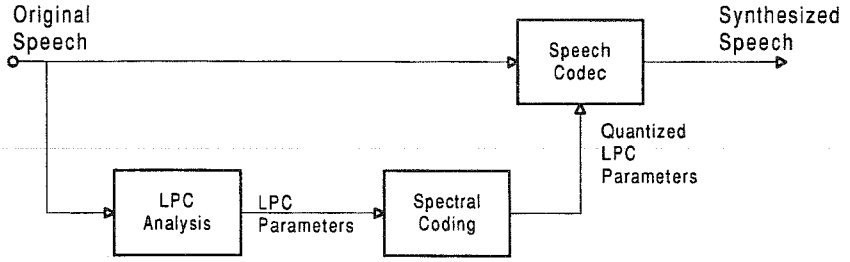

Figure 2. The second basic model for spectral coding assessment

It is noteworthy that in analysis-by-synthesis coders, such as FS1016 CELP (1991), there is no need to have a complete codec as in Figure 2, since the synthesized speech signal is available in the encoder.

Each of these two basic models has its own strengths and weaknesses. The first model can be used for comparison of spectrum quantization methods without any need for implementation or even choice of a complete speech coding algorithm. Therefore the assessment will not be affected by the distortion resulting from other stages of a chosen speech coder. This is not the only advantage of the first model over the other model; its other benefit is its simplicity for evaluating new spectral coding methods without making any assumptions about the rest of the speech coding algorithm.

However, despite its superiority in these respects, the first model has some drawbacks in comparison with the second model. The remarks in the previous paragraph can be interpreted in a different way, which reveals a deficiency of the first model. Thus, with this method it could be concluded that selection of a proper short-term spectrum quantization method for a particular speech coder is independent of the choices for any other blocks in the coder. Therefore, after making a final selection of those blocks, there would be no way to enhance the performance of the spectral coding method using that knowledge. This conclusion does not seem likely, since one may imagine other spectral coders that could benefit from this information.

The next problem of the first model is its basic hypothesis. In this model it is assumed that the residual signal, which is obtained by passing the original speech signal through the analysis filter, is filtered by the quantized synthesis filter. In other words, it is considered that the above residual is the target excitation, which will be encoded in a typical LP-based speech coder. However, as is well known in all standard LP coders, the target residual is calculated by passing the original speech through the quantized analysis filter and not the unquantized one. By contrast, the second model copes with these problems very well. This matter will be discussed further in the following.

INCOMPATIBILITIES IN COMPARISONS OF SPECTRAL CODING METHODS

It is every researcher's ideal to develop an experimental simulation whose results are affected only by the variation of the parameter under study. However, most of the time this may not be possible or at least not feasible, especially in the speech processing field. The outcome of objective quality assessment is not only a function of the particular spectral coding scheme under test, but is also sensitive to other variables.

In the past decade enormous efforts have been dedicated to research on the quantization of the short-term spectrum in low rate speech coders and have resulted in scientific publications that contain the results of simulation experiments. Unfortunately, most of the time these results cannot be compared directly without a great deal of concern about some important aspects of simulation that make it almost impossible to reach a certain conclusion about the superiority or otherwise of one particular method.

Hence, for a precise comparison between the results of independent studies with various spectrum quantization methods, one needs to pay attention to all the procedural differences before deriving any conclusion, even in the situation that the same class of measures is used in all experiments (e.g. objective or subjective measures). These sources of incompatibilities are divided into five groups, namely *Speech Material Effects, Analysis Procedure Effects, Training and Test Condition Effects, Quantization Search Process Effects* and *Quality Assessment Compatibility*, which will be addressed here briefly.

Speech Material Effects

The first source of incompatibility between the results of different studies is the speech variability itself. There is no standard unique set of spoken material that is mandated for spectral coding assessment or even for speech coding itself (regardless of the rationale for such a hypothetical database). Although some particular databases, such as the TIMIT corpus, have been widely used for this purpose, there is still no guarantee that different studies have used the same sentences spoken by similar speakers.

The use of different languages, dialects and utterances in the selection of speech material causes some uncertainty in comparing the results of quality assessments reported in different articles. Most of the investigations use utterances from just one language (normally English), and probably several popular dialects, but a few articles have been published that use a multi-lingual database, see for example Hedelin (1994).

The next issue is the speakers themselves. Speakers are usually selected from different groups of age and gender. It is also common to choose native speakers, except for some particular purposes. Even matters such as emotional condition and the rate of speaking will have some effects on the total speech material and consequently on the evaluation results.

The recording conditions and the equipment used for recording of the speech material are also important in this regard. For example, a codebook that is trained exclusively on clean speech recorded in an anechoic chamber with professional high quality equipment will not be an optimum choice for a system that normally deals with telephone speech with high levels of background noise and other interference (as in mobile phone applications).

In general, speech material for training or testing a national or global standard coder is taken from different possible conditions. However, for most research studies, only clean speech is used, and the same recording conditions apply for collecting the entire speech material. Of course, if an investigation about the variability of recording conditions is considered to be one of the aims of such studies, then different types of recordings should be used, like the research reported in Paliwal & Atal (1993). Usually, each investigation is performed with speech material collected with (or converted to) a unique sampling frequency, normally 8 kHz.

Analysis Procedure Effects

The next group of factors which affect the compatibility of simulations in different articles is related to the LPC analysis procedure used for extraction of the spectral envelope information (regardless of the parameters employed to represent the LP filter).

The size of the analysis frame and its repetition rate (whether or not the frames have any overlap, and also the size of such overlaps if any) will affect the final results. It is well known that inter-frame correlation between the LP parameters of adjacent frames for 30 ms frames is less than for 10 ms frames, and also the dispersal of the spectral envelope extracted from the former frames is wider than for the latter ones. Use of overlapping increases the correlation between adjacent frames, making them more suitable for quantization schemes that exploit inter-frame correlation and provide better assessment results. On the other hand, coding of overlapped frames will reduce the compression ratio of a speech coding algorithm.

The type of window (e.g. rectangular or Hamming) used for spectral analysis, and whether or not a filter (e.g. pre-emphasis filter) is employed in the pre-processing stage, will also change the results of experimental simulations. Other major issues are the linear prediction analysis method and the order of the predictor. For instance, the covariance method provides more accurate information about the spectral envelope than the autocorrelation method, since it does not suffer from boundary effects. Moreover, increasing the order of the predictor also improves the performance of spectral envelope estimation, assuming that the same type of model (e.g. auto-regressive model) is used. However, for spectral coding purposes and with few bits available, it is not always beneficial to increase the order of the predictor beyond some limit.

Training and Test Condition Effects

Even with the same training and test speech databases, and even assuming compatibility in the other four groups of effects considered here, it is still not certain that the results of experiments on two different spectral coding methods can be properly compared. For instance, some of the spectral coding methods are intentionally trained for a target range of channel errors. Obviously, for these types of spectral coders the results of the test may be different from the cases in which no channel errors have been considered in the training.

It should be noted that here we consider only quality assessment of the coding system without channel errors. If the arrangement for reducing the effects of channel errors is only limited to simple schemes without considering channel errors during training, then the evaluated results can be compared without any trouble.. Otherwise, this incompatibility should allowed for in the final conclusions. Some papers consider the quality assessment of spectral coders after introducing the effects of channel error with different probabilities, see for example Paliwal & Atal (1993). In those cases the above considerations should be noted and special attention should also be paid to the compatibility between the probability of channel errors in the training and test stages of the spectral coders under study. In other words, it is necessary to compare different short-term spectrum quantization methods under similar channel error conditions in the training and test stages.

There is another sort of variation that may arise when the number of speech coding stages in a tandem connection known *a priori* and this information is used in the training of the quantization tables or codebooks for spectral coding. To our knowledge this point has not been addressed before and no investigation into it has been performed. Nevertheless, it seems that whether a spectral coder is trained with several speech coders in tandem, or whether it is trained with a single speech coder, will

affect its performance. Therefore, this point should also be noted when comparing the performance of different spectral coding methods.

Quantization Search Process Effects

Given two spectral coding methods with the same overall coding scheme, bit allocations and codebooks, it is still possible to end up with two different quantized spectra if different search strategies are used. For example, if a partial codebook search is employed for reasons of computational saving, there may be increased quantization distortion.

Another issue is the use of different distance measures in the codebook search. To produce a quantized spectral envelope with higher quality, one may use more complex distance measures, e.g. with data-dependent or fixed weightings. This normally increases the computational load of a search algorithm. Moreover, the effectiveness of this arrangement depends on the suitability of the selected measure and weighting.

In some spectral coding methods, such as tree-search VQ, multi-stage VQ and split VQ, it is possible to conduct a search with several survivors. This improves the quality of the quantized spectrum, but the computational complexity might increase several-fold, depending on the number of survivors.

Quality Assessment Compatibility

The compatibility between the applied distortion measures is the final group of effects that is reviewed here. First, it should be assured that the same basic model (either Figure 1 or Figure 2) is used in all experimental simulations.

The next point is to consider exactly the same types of quality measures in different simulations before any comparison. For example, several variants of the spectral distortion measure have been used in the literature under the same names, see Quackenbush *et al.* (1988). Sometimes only speech frames with a certain property are taken into account. For instance, in evaluation of the Segmental Signal to Noise Ratio, researchers may use different thresholds. Therefore, the compatibility from this point of view should also be considered.

COMPUTER SIMULATION

To choose the best spectral coding method among several candidates, various factors should be taken into account. For instance, the complexity and storage costs of the quantization algorithm, the total delay of the coding algorithm and the number of bits assigned to spectrum quantization are all important in making such a decision. However, if these quantities are similar for different spectrum quantization methods, then the spectral coding method that produces the minimum distortion (or equivalently provides maximum quality) will be the optimum choice.

Regardless of the model used for spectral coding assessment, two types of evaluation methods can be applied, namely *subjective quality measures* and *objective quality measures*. No published papers have reported the use of well-known subjective tests, such as mean opinion score (MOS), for spectral coding assessments, but some of them have reported the results of informal listening tests. Although the use of various objective measures has been reported for speech quality assessment, the only one that has been employed extensively for the assessment of spectral coding techniques is the spectral distortion (SD) measure (with small difference in formulations). Moreover, almost all reported investigations have deployed the first basic model (Figure 1) for this purpose.

The other spectral distortion variant, called *synthesized spectral distortion (SSD)* (Sadegh Mohammadi & Holmes, 1994), employs the second basic model (Figure 2) and the same formulation as SD, with the difference that the SSD is calculated from the difference between the original LPC spectrum and the LPC spectrum that is extracted from the synthesized (reconstructed) speech signal by linear prediction analysis. Obviously, for this evaluation it is necessary to import the quantized LP parameters (e.g. LSFs) to a typical speech coder, such as the FS1016 standard CELP coder, but with its spectral coding stage replaced by the spectrum quantization method under test.

Here we use the results of simulation experiments reported by Sadegh Mohammadi & Holmes (1995) to present the differences between the results of quality assessments using two objective measures when all conditions of the experiments are quite similar except for the basic criteria (i.e. SD and SSD). Various vector quantization methods were used in those experiments with trained codebooks for LSF quantization, including unstructured vector quantization (UVQ) and tree-searched VQ (TSVQ) of split LSF vectors of size (3, 3, 4) for representing an entire ten-dimensional LSF vector. In addition, two multi-stage VQs of the entire LSF vector (without splitting) have been used, i.e. three stages (MSVQ3-8) and four stages (MSVQ4-6). All VQs are at bit rates of 24 bits/frame.

The test database includes 69 seconds of speech. Neither the speakers nor the sentences were common to the two databases. The LSFs of the test database were quantized by the various methods, and then used in a simulated speech coder which is similar to FS1016 (apart from the LSF quantization method). Table 1 depicts the objective measures obtained with the various quantization schemes. As the results show the SD indicates UVQ is superior to MSVQ-4 while the SSD predicts that it is inferior.

Table 1. Results of quality assessment

Quantization Method	SD [dB]	SSD [dB]
UVQ	1.46	2.35
MSVQ3-8	1.40	2.30
MSVQ4-6	1.47	2.32
TSVQ	1.65	2.43

CONCLUSIONS

The difficulties in comparing the results of different objective measures for the quality assessment of various spectral coding methods are discussed and several important considerations are described in this regard. Two basic models for such assessments are addressed and an insight into various sources of incompatibilities is obtained. The results of a computer simulation were used to show how these incompatibilities lead to different conclusions for quality assessment.

REFERENCES

Federal standard 1016 (1991) *Telecommunications: analog to digital conversion of radio voice by 4,800 bit/second code excited linear prediction (CELP)*, (National Communications System, Office of Technology and Standards: Washington, DC20305-2010).

Hedelin, P. (1994) "Single stage spectral quantization at 20 bits", *Proc. ICASSP*, vol. 1, pp. I.525-I.528, Apr. 1994.

Paliwal, K.K. & Atal, B.S. (1993) "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. on Speech and Audio Proc.*, vol. 1, no. 1, pp. 3-14, Jan. 1993.

Quackenbush, S.R., Barnwell III, T.P. & Clements, M.A. (1988) *Objective Measures of Speech Quality* (Prentice-Hall: New Jersey).

Sadegh Mohammadi, H.R. & Holmes, W.H. (1994) "Fine-Coarse Split Vector Quantization: An Efficient Method For Spectral Coding", *Proc. of Fifth Australian Intern. Conf. on Speech Sci. and Tech.*, pp. 118-123, Dec. 1994.

Sadegh Mohammadi, H.R. & Holmes, W.H. (1995) "Low cost vector quantization methods for spectral coding in low rate speech coders", *Proc. ICASSP*, vol. 1, pp. 720-723, May 1995.