Proceedings of National Conference on Communications, NCC 2003, Jan 31 –Feb 2, IIT Madras 2003

ROBUSTNESS OF THE MBE VOCODER TO ACOUSTIC BACKGROUND NOISE

Preeti Rao , Milind Pathak** and T.R. Ramamohan***

*Department of Electrical Engineering,
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
(e-mail: prao@ee.iitb.ac.in)

**Central Research Labs.,
Bharat Electronics Limited,
Bangalore 560013

ABSTRACT

I.I.T. Bombay, in collaboration with C.R.L., B.E.L., is working on developing a speech codec to provide communication quality speech at low bit rates. The codec is required to be robust to acoustic background noise and to channel errors. A sinusoidal coder based on the MBE (Multiband Excitation) model has been adopted as the basic speech representation due to its compact parameter set and the relatively robust, available parameter estimation algorithms. This paper reports a study on the influence of acoustic background noise on the performance of an MBE vocoder with the eventual goal of designing a suitable speech enhancement preprocessor for use in low SNR situations.

1. INTRODUCTION

For the compression of narrowband speech at bit rates below 4 kbps, the quality of hybrid waveform coders such as CELP drops dramatically. In this realm, "vocoders", i.e. speech coders based not on waveform matching but rather on a purely parametric description of the signal (usually derived from a speech production model) have been adopted in various applications. Vocoders use the periodic characteristics of voiced speech and the noise-like characteristics of unvoiced speech to achieve compact parametric representations of the speech signal. The most popular vocoders today are: harmonic coders (includes the Sinusoidal Transform coder and the Multiband Excitation coder), Prototype Waveform Interpolation (PWI) coders and LPC-based vocoders (includes Mixed Excitation LP) [1].

The above speech coders are typically based on compact, parametric signal models based on the speech production mechanism. That is, the design of the model, as well as the analysis and synthesis algorithms, are optimized for the characteristics of clean speech rather than for the matching of a general waveform. As a consequence, the presence of acoustic background noise leads to gross inaccuracies in the estimates of the parameters and therefore degrades the performance of such speech codecs far more severely than would be expected from a consideration of the input noisy speech signal alone. To what extent the parameter errors affect the subjective quality of the reconstructed speech output of the codec depends greatly on the signal model, the particular parameters and the parameter estimation algorithm.

Low quality speech is stressful and fatiguing over long listening durations, and is often accompanied by reduced speech intelligibility. It is of interest therefore to investigate methods to improve the performance in noise of communications systems employing low rate speech codecs. One obvious approach to improving the performance of the codec is to apply some form of preprocessing to increase the speech signal-to-noise ratio at the input to the codec as depicted in Fig. 1. Such an approach is simple and convenient in that it does not require any modification of the speech coding algorithm itself.

Recently, significant improvements have been reported when specific speech coders were combined with a speech enhancement preprocessor. Guilmin et al [2] showed that Wiener filter-based noise preprocessing significantly improved the output, in the presence of noise, of a low rate LPC vocoder both in terms of parameter estimates and subjective quality. Earlier Kang and Fransen [3] evaluated spectral subtraction enhancement for LPC-processing of noisy speech and reported dramatic

improvements in subjective quality for speech corrupted with a variety of background noise. Recently a number of different preprocessing schemes were examined for use with the Federal standard 2.4 kbps MELP coder [4]. These studies indicate that the choice and optimization of the speech enhancement preprocessor will be dependent on the speech coding algorithm. Also, it has been noted [5] that this optimization may actually be quite different from that required for simple listening (without coding) to noisy speech. As a first step therefore, it is necessary to study the noise performance of the speech coding algorithm at hand including the influence of noise on the parameters of the speech representation used.

In our work we consider the class of speech coders based on the sinusoidal model of speech representation. Typically, sinusoidal coders operating at low rates have a minimal parameter model for encoding that is based on a mixed spectral representation. Bit-rates in the region of 2 kbps have been achieved by sinusoidal-model based codecs such as the Sinusoidal Transform Codec and the Multiband Excitation (MBE) codec [6],[7]. The MBE vocoder has been shown to be robust to background noise at SNRs above 5 dB [7]. The focus of this paper is a study of the impact of additive background noise on the performance of an MBE speech codec operating at even lower SNRs.

In the next section, we review the MBE speech representation and associated speech analysis method, and discuss its performance for noisy speech.

2. THE MBE MODEL AND ANALYSIS ALGORITHM

In the MBE speech representation, voiced regions are modeled by harmonics of a fundamental frequency, and unvoiced regions by spectrally shaped random noise. The parameters of the MBE speech model consist (for each analysis frame) of the fundamental frequency, voicing decisions (one for each group of 3 harmonics) and the harmonic amplitudes [7]. The voicing information allows the mixing of the harmonic spectrum with a

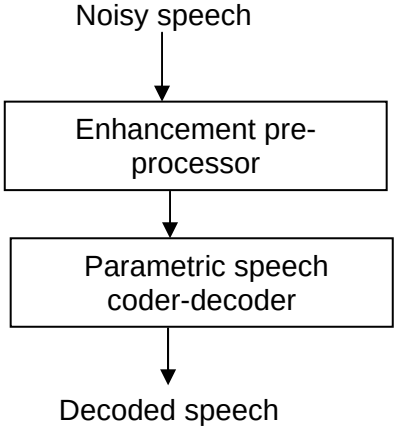

Fig. 1. Improving the noise robustness of a low rate speech codec with a preprocessor.

random noise spectrum in a frequency dependent manner. The phase of harmonics is not transmitted but predicted during synthesis in most low rate coders. Often, to keep the bit rate low, the multiband voicing vector is replaced by a voicing “cut-off” band number, above which band the frequency spectrum is taken to be unvoiced. Fig. 2 depicts the MBE parameters and their relation to the speech signal power spectrum.

As described in [7], we estimate the excitation and system parameters, for each input frame of 20 ms duration, which minimize the distance between the original and synthetic speech spectra by an analysis-by-synthesis (AbS) method. The error distance is first minimized over the fundamental frequency and spectral amplitudes assuming all voiced speech. Once these parameters are estimated, voicing decisions are made based on the closeness of fit between the original and synthetic spectrum for each group of harmonics. A frequency-dependent threshold is applied to this normalized error to get a voicing decision error for each band [6]. The multiband voicing vector is then replaced by a single voicing cut-off band number obtained by a suitable filtering of the binary decisions. We see, therefore, that the synthesized speech quality of the MBE speech coder depends greatly on the accuracy of the pitch estimate since both spectral amplitudes and voicing decision are based on AbS matching of a synthetic spectrum, based on the estimated pitch, with the input spectrum. Gross pitch errors due to the selection of pitch period multiples are minimized by a process that favours lower submultiples of pitch. Dynamic pitch tracking is used to improve pitch estimation in noise by reducing gross pitch errors via imposing smoothness constraints on the estimated pitch across frames. Several past and

future frames are searched jointly in order to find the pitch track with the minimum error [7]. In our implementation we have used two frame look-back but a single frame look-ahead in order to keep computational complexity low.

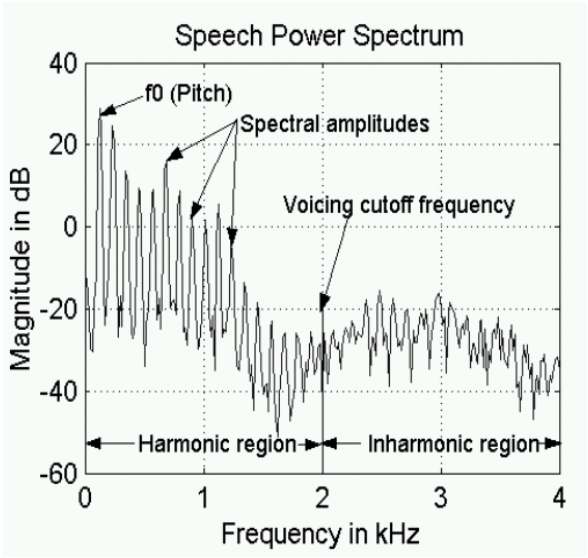

Fig. 2 The parameters of the MBE speech model

3. INFLUENCE OF ADDITIVE NOISE

While the quantisation of the sinusoidal and noise parameters varies greatly among codecs depending on the actual target bit rate, it turns out that the parameter set, and associated analysis and synthesis procedures, are essentially the same for all low rate sinusoidal coders. We therefore confine ourselves at first to studying the effect of noise on a combination of only the main functional blocks, namely analysis and synthesis modules (henceforth referred to as the MBE speech codec). All our simulations are based on the "white noise" sample file from the SPIB database [8].

We studied the influence of additive white noise at various SNRs on the parameters obtained by the MBE analysis of noisy speech. An 8 kHz sampled speech file consisting of the concatenation of 30 sentences uttered by 15 male and 15 female speakers was constructed using sentences from the TIMIT database [9] among others. This file contained about 30 % silent frames (a typical telephone conversation has 60% silence and non-speech sound) but otherwise fully tested the system in terms of a wide range of voice types and content. The SNR is computed according to ITU standard [10] based on the r.m.s. "active speech level" and r.m.s. noise level. The active speech level is estimated by leaving out silence and idle segments but including grammatical/structural pauses (i.e. those within 300 ms). Reference parameters obtained from the MBE analysis of the clean speech are used to determine parameter errors that occur after the addition of noise. A pitch estimate is deemed correct if its value does not differ from the reference by more than 5 % [2]. Fig. 3 shows the behaviour of percentage pitch errors in speech frames for speech in additive white noise at various SNRs. We see that while the pitch tracker is useful in keeping down pitch errors at high SNRs, the pitch estimate degrades gradually as SNR decreases. The increase in pitch errors with decreasing SNR is also borne out by informal listening. Another aspect of the perceived speech quality that is evident from informal listening is the change in the nature of the background noise as reproduced by the codec. This is an important determinant of overall signal quality. While at high SNRs, the output sound continues to be a good reproduction of the input, there is a turning point below which the background noise is marked by an annoying, intermittent buzziness resulting in significant overall quality degradation. On closer examination, his was found to be due to the occurrence of intermittent pitch structure from a marked increase of unvoiced-to-voiced errors in the non-speech (noise only) regions. Fig.3 also shows the percentage of voicing errors in the non-speech frames versus SNR. We note a sharp increase in % errors as SNR drops below 8 dB. This can be explained by the increased variance of the background noise at lower SNRs leading to larger fluctuations in the normalized error between the estimated and synthesized spectra computed during MBE analysis. That the percentage voicing error remains more or less constant below the "breakdown" SNR may be explained by the fact that in the MBE analysis, the error threshold itself is adapted according to the local signal energy [6].

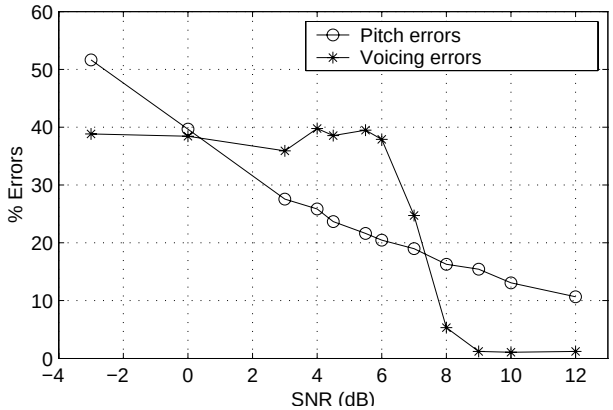

Fig. 3. % Pitch errors in speech frames and % voicing errors in non-speech frames for noisy speech at various SNRs

Table 1 shows the extent of parameter errors during speech frames (that is speech plus noise is present). We see that voicing errors (defined as occurring whenever voicing cut-off frequency deviates by more than 500 Hz) are high and tend to increase with decreasing SNR. The log spectral distortion (S.D.) is measured over the 0-3 kHz band and is seen to increase with decreasing SNR. The voicing errors and S.D. point to the loss of important speech spectral cues at low SNRs.

SNR (dB)	Parameter errors		
	pitch	voicing	S.D.
	(%)	(%)	(dB)
+3	27	75	6.8
0	39	78	7.7
-3	50	84	8.5

Table 1. Parameter errors during speech frames for noisy speech at various SNRs

4. EVALUATION OF SPEECH QUALITY

The performance of a speech codec under noisy conditions may be expressed in terms of subjective measures such as quality and intelligibility that reflect how the output signal is perceived by listeners. The quality attribute is related to the pleasantness of the sound or how much effort is required on behalf of the listeners in order to understand the message. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted by listeners from the given signal. A given signal may be of high quality but low intelligibility, and vice versa. Hence, the two measures are relatively independent of each other [11].

To assess the extent of degradation of performance of the speech coder in noisy conditions compared with the clean speech condition, it is necessary to obtain some form of quantitative estimate of the speech quality and intelligibility. Subjective testing procedures may be used for this purpose. Evaluation using additive white noise was carried out at 3 low SNRs: -3 dB, 0 dB and 3 dB (SNRs computed according to [10]). To avoid listener familiarity with a specific noise sample, segments of the noise file to be added to the sentences were chosen randomly.

Subjective tests to evaluate the intelligibility and quality of spectrally enhanced noisy speech before and after MBE modeling were carried out. Intelligibility was measured using the Modified Rhyme Test [12]. There are 50 sets of 6 single-syllable words available to test consonant intelligibility. Vowel intelligibility was included by adding one more set: *had, hid, hod, hud, head, hee*d. The listener hears one word (randomly chosen) from each set in the carrier phrase "Would you write xxx" And so on, 51 times. The 6 possible words are presented on the test sheet to the listener who indicates his opinion of the word uttered by selecting it. One speaker and six listeners were used in the subjective test. Percentage correct responses were scored for each listener and averaged across listeners. Although additional speakers would have been desirable to get reliable measurements of intelligibility, in the present study only comparisons were of interest.

The results appear in Table 2. We see that the intelligibility of noisy speech drops sharply after coding (MBE analysis and synthesis) at all SNRs. We note that while the codec reduces the intelligibility of clean speech (SNR = infinity) by only 3 points it degrades the intelligibility as well as quality of noisy speech much more. The decrease in intelligibility appears to be correlated with the increase in the parameter errors of Table 1.

SNR (dB)	Intelligibility (%)	
	Input	Output
Infinity	94	91
+3	63	53
0	57	46
-3	46	42

Table 2. Percentage intelligibility of noisy input speech and coded output speech at various SNRs.

5. CONCLUSIONS AND FUTURE WORK

Our studies so far indicate that the loss in speech quality and intelligibility of MBE-coded speech in high acoustic background noise is related to the increased errors in analysis parameters. While the MBE analysis algorithm has an inherent noise robustness, there is a strong need for added noise pre-processing at very low SNRs such as considered here. The understanding obtained of the behaviour of parameter errors with SNR can aid in the design of an effective noisy speech preprocessor. Optimisation of the preprocessor can be based on obtaining the maximum reduction in parameter errors.

So far, we have not considered any broadband noise other than white noise. We plan to study the performance of the speech codec under more realistic noise conditions including moving vehicle noise and babble. Speech enhancement algorithms will be reviewed and investigated for possible application in a noisy speech preprocessor for the low bit rate MBE speech codec.

REFERENCES

[1] Voice Compression and Communications, Hanzo, Sommerville and Woodard, IEEE Press, 2001.

[2] Guilmin G., Le Bouquin-Keanns R., and Gournay P., “Study of the influence of noise pre-processing on the performance of a low bit rate parametric speech coder,” in *Proc. Europ.Conf. on speech Comm. and Tech*., Sep. 1999, vol. 5, pp. 2367–2370.

[3] Kang G.S. and Fransen L.J., “Quality improvement of lpc-processed noisy speech by using spectral subtraction,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 6, pp. 939–942, Jun 1989.

[4] Tarun Agarwal, “Pre-processing of noisy speech for voice coders,” M.S. thesis, Department of Electrical and Com-puter Engineering, McGill University, Montreal, Canada, Jan 2002.

[5] Collura J.S., “Speech enhancment and coding in harsh acoustic noise environment,” in *Proc. IEEE Workshop on Speech Coding*, May 1999, pp. 162–164.

[6] Kondoz A.M., “Multi-band excitation speech coder,” in *Digitalspeech coding for low bit rate communications systems*. 1994, John Wiley.

[7] Griffin D.W.and Lim J.S, “Multiband excitation vocoder,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, Aug 1988.

[8] Signal Processing Information Base, “Noise data,” http://spib.rice.edu/spib/select noise.html, Jan 2001.

[9] N. Institute of Standards, “The TIMIT cdrom,” 1989.

[10] “Objective measurement of active speech level,” Mar 1993, ITU-T Recommendation P.56.

[11] Ephraim Y., “Statistical-model-based speech enhancement systems,” *Proc. of the IEE*E, vol. 80, no. 10, pp. 1527–1555, Oct 1992.

[12] Meyer Sound Laborataries, “Statistical measures of speech intelligibility,” http://www.meyersound.com/support/papers /speech/section3.htm, Oct 2001.