Accepted for publication in Proceedings of 7^{th} International Conference on Spoken Language Processing, ICSLP2002, September 16-20, Denver, Colorado, USA 2002

SPECTRAL ENHANCEMENT PREPROCESSING FOR THE HNM CODING OF NOISY SPEECH

Gautam Moharir, Pushkar Patwardhan and Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology, Bombay, India 400076
Email:prao@ee.iitb.ac.in

ABSTRACT

Low rate coders based on the harmonic-noise model are sensitive to acoustic background noise at low SNRs due to the increase in parameter errors from the analysis of noisy speech. We investigate the use of spectral subtraction enhancement preprocessing on the performance of the sinusoidal model based codec both by objective assessment of parameter errors and the subjective testing of output speech quality and intelligibility. We find that while for noisy speech enhancement, improving speech quality is often accompanied by a decrease in intelligibility, in the context of coding, significant combined improvements are obtained when the speech coder is combined with a speech enhancement preprocessor.

1. INTRODUCTION

Low bit rate codecs for narrowband speech are typically based on compact signal models with characteristics tuned to clean speech signals. As a consequence, the presence of acoustic background noise usually degrades the performance of such speech codecs far more than would be expected from a consideration of the input noisy speech signal alone. A low signal-to-noise ratio at the input can lead to errors in the estimates of model parameters. How these affect the subjective quality of the reconstructed speech depends to a great extent on the signal model, the specific 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. Such an approach is simple 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 [1] 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 [2] 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 [3]. 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 [4] that this optimization may actually be quite different from that required for simple listening (without coding) to noisy speech.

In this 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 [5],[6]. The Harmonic plus Noise Model (HNM), recently popular in speech synthesis applications [7], is closely related to the MBE and has the potential to result in even lower bit rates due to a reduction in transmitted information. The focus of this paper is a study of the impact of a spectral subtraction-based speech enhancement method on the performance of a sinusoidal-model based codec (HNM codec) operating at low SNRs. The performance of such a system 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 [8].

In the next section, we review the HNM speech representation and associated speech analysis method, and discuss its performance for noisy speech. The spectral subtraction method of speech enhancement is reviewed next followed by a study of its performance as a preprocessor for the HNM based codec.

2. HNM SPEECH REPRESENTATION AND INFLUENCE OF ADDITIVE NOISE

In the HNM speech representation (and its precursor, the MBE model), 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 magnitudes [9]. The voicing information allows the mixing of the harmonic

This work has been supported in part by a research grant from D.R.D.O., Govt. of India.

spectrum with a random noise spectrum in a frequency dependent manner. The phase of harmonics is not transmitted but predicted in most low rate coders. The HNM representation is identical to the MBE, except that the multiband voicing vector is replaced by a voicing "cut-off" band number, above which band the frequency spectrum is taken to be unvoiced. 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 to studying the effect of noise and of speech preprocessing on a combination of only the main functional blocks, namely analysis and synthesis modules (henceforth referred to as the HNM speech codec). All our simulations are based on the "white noise" sample file from the SPIB database [10].

2.1. HNM analysis algorithm

Similar to [9], 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/HNM 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 [9]. In our implementation we have used two frame look-back but a single frame look-ahead in order to keep computational complexity low.

2.2. Influence of noise on pitch and voicing

We studied the influence of additive white noise at various SNRs on the parameters obtained by HNM 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 [11] 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 [12] 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 HNM 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% [1].

Figure 1 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 a significant overall quality degradation. On closer examination, this 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. Figure 1 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 HNM analysis. That the percentage voicing error remains more or less constant below the "breakdown" SNR may be explained by the fact that in the HNM analysis, the error threshold itself is adapted according to the local signal energy [6].

3. SPECTRAL SUBTRACTION SPEECH ENHANCEMENT

Spectral subtraction is a family of frequency-domain noise reduction techniques based on subtracting the estimated short-term amplitude spectrum of the noise from the short-term amplitude spectrum of noisy speech. Speech enhancement by spectral subtraction consists essentially of the three processing steps: short-term spectral analysis, noise suppression, and short-term spectral synthesis. The spectral subtraction method of Berouti et al [13] is used in this work. It is described by the following equation:

$$\begin{aligned} \mid \hat{S}_w(\omega) \mid^2 &= \mid X_w(\omega) \mid^2 - \alpha \mid \hat{D}_w(\omega) \mid^2 \\ &if \mid \hat{S}_w(\omega) \mid^2 > \beta \mid \hat{D}_w(\omega) \mid^2; \\ &= \beta \mid \hat{D}_w(\omega) \mid^2 otherwise \end{aligned} \tag{1}$$

where α and β are "subtraction factor" and "spectral floor" respectively, and lie within the range $\alpha \geq 1.0$ and $0 \leq \beta \leq 1.0$. $\hat{S}_w(\omega)$,$X_w(\omega)$ and $\hat{D}_w(\omega)$ are spectral estimates of clean speech, noisy speech and the noise respectively. α is varied from frame to frame as:

$$\alpha = \alpha_o + s \times SNR \tag{2}$$

where SNR is the aposteriori speech-to-noise ratio, and α_o is the value of α at SNR = 0dB and s is a chosen constant. The noisy speech and the noise power spectra are computed using DFTs with 75% overlapping, Hamming windowed frames of duration 30 ms.

The noise estimate is updated during non-speech periods [14]. The enhanced speech is obtained by synthesis from the modified magnitude spectra combined with the original estimated phases. Eq. (1) represents a modified version of the basic spectral subtraction method [15] in that an overestimate (determined by α) of the noise spectrum is subtracted and the resultant spectral components are prevented from going below a spectral floor (determined by β).

The enhanced speech is characterized by a speech quality that can be described as speech distortion plus residual noise. The residual noise consists of i) the remaining background noise which has the same perceptual characteristics as the original noise, and ii) musical tone artifacts due to the presence of random, short duration peaks in the spectrum [13]. β trades off musical noise for background noise. Increasing α_o leads to increasing oversubtraction of the noise and therefore reduces both background and musical noise but at high values, increasing distortion of speech is perceived due to loss of speech spectral amplitudes.

4. PERFORMANCE EVALUATION OF THE SPECTRAL SUBTRACTION PREPROCESSOR

In the speech coder preprocessing application of this paper, the performance of the Spectral Enhancement System (SES) must be judged by the output speech quality of the speech codec. It is possible that the SES is good in this context even if its own output (the enhanced speech) is inferior in some sense to its input. Therefore subjective tests were carried out to evaluate the performance of the SES itself, the speech codec alone and the combination of the two. Three different configurations of the SES were evaluated: E1: $\alpha_o = 3, \beta = 0.005$; E2: $\alpha_o = 5, \beta = 0.001$; E3: $\alpha_o = 7, \beta = 0.001$. Evaluation using additive white noise was carried out at 3 low SNRs: -3 dB, 0 dB and 3 dB (SNRs computed according to [12]). To fully test the system, and to avoid listener familiarity with a specific noise sample, segments of the noise file to be added to the sentences were chosen randomly.

4.1. Subjective tests

Subjective tests to evaluate the intelligibility and quality of spectrally enhanced noisy speech before and after sinusoidal modeling were carried out. Intelligibility was measured using the Modified Rhyme Test [16]. 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, heed*. The listener hears one word from each set in the carrier phrase "Would you write ..." And so on, 51 times. The 6 possible words are presented on the test sheet. One speaker and six listeners were used. Percentage correct responses are 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 are of interest. Quality rankings were carried out by means of A-B comparison tests [3]for each SNR using a set of 10 different sentences spoken by 5 male and 5 female speakers. Quality ranking was relative within each SNR condition. On clean speech, we observed intelligibility to be equal to 94% before coding and 91% after coding. Results for noisy speech appear in Tables 1 and 2.

4.2. Objective evaluation

Parameters are estimated for noisy speech input to the codec with and without enhancement preprocessing. The objective evaluation was carried out using the set of 30 sentences described in Section 2.2. White noise at each of the three SNRs was added to the sentences. Reference analysis parameters were obtained by HNM analysis of the clean speech. Parameter errors were determined for the noisy speech by comparison with the reference parameters of the corresponding frames.

The error measures were defined in a manner so that they could be reasonably related to the subjective test measures. The errors computed are: pitch (% correct), pitch multiples (% errors), voicing errors (%) in non speech frames, (%) correct voicing (cut-off frequency deviation < 500 Hz) in speech frames and log spectral distortion (S.D.)(0-3kHz band) in speech frames where there are no pitch errors. Once again, a "pitch error" occurs only when the pitch frequency deviation is > 5 %.

SNR (dB)	Configuration	Subjective tests	
		Quality rank	Intelligibility (%)
	noisy	3	63
3	E2	2	54
	E3	1	54
	noisy	3	57
0	E1	2	51
	E2	1	50
	noisy	4	46
-3	E1	3	50
	E2	2	54
	E3	1	45

Table 1. Quality ranking and % intelligibility for spectral subtraction system with different configurations

SNR (dB)	Conf.	Qual. rank	Int. %	Pitch		Voicing		S.D. (dB)
				% Pitch corr.	% Pitch multi. err.	% Vuv err. (non-sp.) frames	% Vuv corr. (sp.) frames	
	noisy	3	53	73	7	40	25	6.8
3	E2	2	54	76	3	10	31	5.5
	E3	1	54	74	4	3	29	6.1
	noisy	3	46	61	11	40	22	7.7
0	E1	2	52	71	4	41	29	4.9
	E2	1	50	68	4	15	25	5.7
	noisy	4	42	50	13	41	16	8.5
-3	E1	3	51	64	6	43	23	5.2
	E2	2	51	60	6	17	20	5.8
	E3	1	46	59	6	2	19	6.1

Table 2. Subjective and objective measures for HNM codec speech output. (Int. : Intelligibility; multi. : Multiple; non-sp. : Non-speech; sp. : Speech; err. : Error; corr. : Correct; S.D.: Log Spectral Distortion)

5. DISCUSSION AND CONCLUSIONS

Examining only noisy speech data (Table 1) we see that for decreasing SNR, the intelligibility of noisy speech decreases first gradually and then rapidly. This is in line also with the observations of [17]. We see that the subjective quality improves with SES processing, with some configurations better than others. For different SES parameter settings, different types of degradations are perceived in the SES output. As expected, the musical noise is

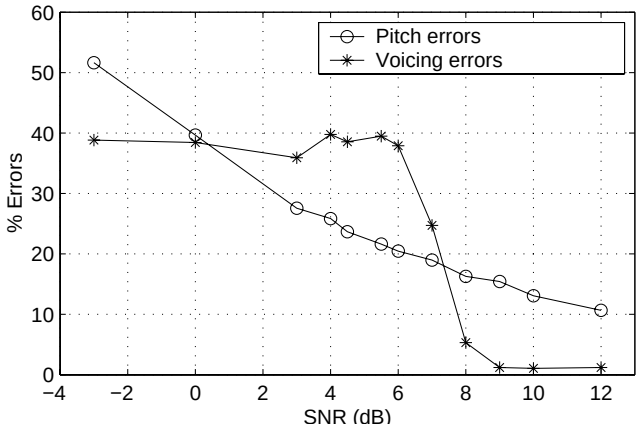

Fig. 1. % Pitch errors in speech frames, and % voicing errors in non speech frames for HNM coded noisy speech

high in configuration E2 where α and β are low. We see that the improvement in overall quality is accompanied by a decrease in intelligibility except at the very low SNR of -3 dB. At this SNR trying to increase quality further by increasing α (E3), though, leads to a large reduction in intelligibility due to the increased speech distortion. Table 2 shows the subjective results of the codec output with and without SES preprocessing. Table 2 indicates that the subjective quality of noisy speech always improves with SES processing. Comparing with Table 1, we see that the intelligibility of noisy speech drops sharply after coding (HNM analysis and synthesis) at all SNRs. We note that while the codec reduces the intelligibility of clean speech by only 3 points it degrades the intelligibility as well as quality of noisy speech much more. While preprocessing has little impact on intelligibility at 3 dB SNR, it improves the intelligibility at lower SNRs of 0 and -3 dB significantly.

The objective error measures support the subjective data. That is, we see from Table 2 that quality rankings correlate well with pitch errors and background voicing errors; and that intelligibility losses can be explained by reduced correct percentage voicing (elimination of an important acoustic cue) in speech frames and by the increased average S.D. We note that preprocessing with E2 exhibits somewhat high voicing errors in the non-speech frames. This may be attributed to the presence of musical noise.

In summary, for sinusoidal model based speech codecs at very low SNRs, increased parameter errors (pitch and voicing) lead to the gross degradation of speech quality. Speech enhancement preprocessing based on spectral subtraction is a promising approach to improving codec performance in white noise. This is reflected in subjective tests as well as by appropriate error measures applied to sinusoidal model parameters. Further work is needed to verify these conclusions for other types of noise background such as moving vehicle and babble noises.

6. REFERENCES

[1] 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.

[2] 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.

[3] Tarun Agarwal, "Pre-processing of noisy speech for voice coders," M.S. thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada, Jan 2002.

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

[5] Spanias A.S., "Speech coding: A tutorial review," *Proc. of the IEEE*, vol. 82, no. 10, pp. 1541–1584, Oct 1994.

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

[7] Yannis Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jan 2001.

[8] Ephraim Y., "Statistical-model-based speech enhancement systems," *Proc. of the IEEE*, vol. 80, no. 10, pp. 1527–1555, Oct 1992.

[9] 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.

[10] Signal Processing Information Base, "Noise data," http://spib.rice.edu/spib/select_noise.html, Jan 2001.

[11] N. Institute of Standards, "The TIMIT cdrom," 1989.

[12] "Objective measurement of active speech level," Mar 1993, ITU-T Recommendation P.56.

[13] Berouti L.and Schwartz R. and Makhoul J., "Enhancment of speech corrupted by acoustic noise," in *Proc. of the IEEE conference on Acoutsic, Speech and Signal Processing*, Apr 1979, pp. 208–211.

[14] Gautam Moharir, "Spectral subtraction method for speech enhancement," M.S. thesis, Department of Electrical Engineering, I.I.T. Bombay, Mumbai, India, Jan 2002.

[15] Boll S.F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, Apr 1979.

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

[17] Lim J.S., "Enhancement and bandwidth compression of noisy speech," *Proc. of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[18] Moharir G., Patwardhan P., and Rao P., "ICSLP-02 demo. samples," http://www.ee.iitb.ac.in/˜prao/speechcoding/ icslp02/icslp02res.htm, Mar 2002.