2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A New Dual-Channel Speech Enhancement Technique with Application to CELP Coding in Noise*

S. Nandkumar, John H. L. Hansen, Robert J. Stets
Department of Electrical Engineering
Duke University, Durham NC 27706

ABSTRACT

In this study, low bit-rate speech coding in noisy acoustic environments is addressed. Speech coding based on various forms of linear predictive coding is known to be sensitive to additive background noise. This study investigates in detail a new auditory constrained enhancement (ACE) technique as a front-end enhancer for a 4800 bps CELP coder operating in noisy environments. A novel, dual-channel, auditory based constrained iterative enhancement scheme is developed and shown to improve quality over all classes of speech. Performance of the proposed enhancement-coding tandem is evaluated using objective speech quality measures for nine background noise conditions over a subset of the NIST/TIMIT database. Average objective quality measures indicate that CELP maintains the level of distortion for speech coded in white Gaussian noise, but further degrades speech coded in non-stationary colored noise. A significant quality improvement for the tandem is shown, which is consistent over all speech classes and most phonemes.

10.21437/ICSLP.1992-173

1 INTRODUCTION

The objective of speech coding is to exploit redundancies introduced in the analog acoustic signal during human speech production in order to encode speech at low data rates. Studies have demonstrated however, the deterioration in performance for speech coders operating at data rates from 2.4 to 16 kbps under background noise conditions [1, 2, 4, 7, 8]. Subjective speech quality tests (DAM, DRT) and objective speech quality measures have both been used to illustrate this performance degradation. Low data rate coders such as standard LPC-10 or ADPCM have been shown to further degrade speech already corrupted by background noise [7, 8]. More recently, LPC based coders with complex excitation models such as CELP [4] and RPE-LTP [9] have been shown to at best, retain the input background noise level at their output. It has been suggested that vocoder performance can be improved by employing speech enhancement prior to coding [10]. However, few studies have presented detailed performance evaluation of such enhancement-coding tandem schemes operating in noisy backgrounds. In addition, it is not clear how sensitive alternative excitation schemes such as multi-pulse, stochastic excitation, or code-book excitation methods are to varying levels and types of additive background noise, when used in conjunction with speech enhancement.

Traditional speech enhancement schemes focus primarily on cancelling additive background noise, with a secondary goal of enhancing speech quality. As such, they at times introduce artifacts which could cause further degradation of coded speech in a tandem application. A new auditory based constrained iterative enhancement (ACE) algorithm in a dual-channel scenario is proposed for front-end enhancement. This enhancement approach takes advantage of both auditory and perceptual properties of the corrupted speech signal in order to improve quality, especially in low-energy and transition regions of speech. In this study, objective speech quality measures which have been shown to be correlated with subjective speech quality [11] are used to evaluate performance of i) a stand-alone 4800 bps CELP coding scheme and ii) the ACE-CELP tandem as shown in Figure 1, for three types of background noise. This paper is organized as follows. The CELP coder will be briefly discussed in Sec.2, and the enhancement algorithm presented in Sec.3. Results are discussed in detail in Sec.4, and conclusions presented in Sec.5.

*This work sponsored in part by National Science Foundation, Grant No.IRI-9010536.

2 THE CELP SPEECH CODER

The CELP coding scheme used in this study is the DoD 4800 bps standard (proposed federal standard 1016) as developed in [4]. CELP coding was first introduced by Schroeder and Atal [5], and subsequent developments resulted over several studies [1, 2, 3, 4]. CELP encoding is a frame-oriented analysis-by-synthesis scheme, which exploits short-time formant characteristics and long-term excitation characteristics of speech in order to provide efficient coding. CELP analysis consists of the following three CELP synthesis functions in reverse order. First, short-term linear predictive analysis is performed to model the spectral envelope, resulting in vocal-tract characterization based on ten line spectral pair (LSP) parameters. Second, long-term speech periodicity is modeled by a vector-quantized adaptive codebook (pitch VQ). Third, error from short-term linear prediction and the pitch VQ is modelled by a vector-quantized stochastic codebook. Figure 1 shows a schematic diagram of the CELP analysis and synthesis functions.

Excitation for the short-term prediction filter is formed by a combination of vectors from the adaptive and stochastic codebooks. During CELP analysis, error between synthesized and input speech is perceptually weighted to drive an error minimization search procedure for the two codebook entries. The adaptive codebook contains a history of past excitation signals, and a pitch delay indexes the codeword containing the best block of excitation from the past for use in the present. This codebook contains 256 codewords with 128 integer and 128 non-integer delays. After removing the short-term and pitch structure of the speech signal, the remaining residual can be approximated by a Gaussian probability distribution. Though it has been shown that similar coder performance results if the Gaussian distribution is replaced by uniform or Laplace distributions [6], this effect has not been studied for cases of white or colored background noise. The stochastic codebook which models this residual, contains 512 codewords which are ternary level quantized samples from a zero-mean, unit variance Gaussian sequence. CELP parameters to be transmitted include the stochastic codebook index and gain, the adaptive codebook index and gain, and 10 LSP parameters. The CELP coder provides outstanding quality of speech in noise-free conditions due to its advanced excitation model. Though successful for varying speaker populations in simulated studies, performance can deteriorate in the presence of background noise. Improvement in codebook search and design as well as post-filtering have resulted in CELP coders which at best maintain the level of background noise without further degradation [2, 4]. Post-filtering is used to decrease the perceivable noise in the coded speech signal at the expense of signal distortion. In this study, the postfilter is disabled for all objective quality assessments in noisy environments due to its influence on spectral characteristics.

3 PROPOSED ENHANCEMENT

Speech enhancement is considered in a dual-channel scenario, consisting of a primary microphone which provides a speech signal degraded by additive background noise, and a reference microphone which samples background noise along with a speech component called cross-talk. The reference microphone is positioned so as to obtain a noise measurement which is correlated with the primary channel noise and to minimize cross-talk. The observed signals at the two microphones, $y_1(t)$ and $y_2(t)$, are given by

$$\begin{aligned} y_1(t) &= s(t) + \mathcal{H}_1\{d(t)\} = s(t) + d'(t) \\ y_2(t) &= d(t) + \mathcal{H}_2\{s(t)\} = d(t) + s'(t). \end{aligned} \tag{1}$$

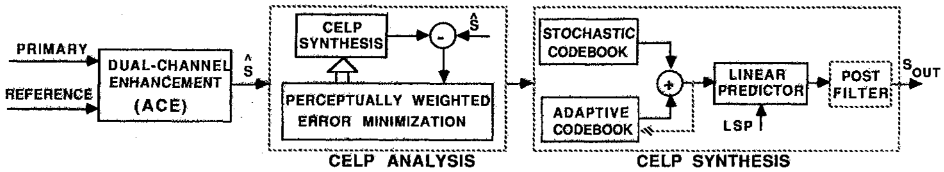

Figure 1: Schematic diagram of the ACE-CELP tandem.

Enhancement in this scenario is formulated as a maximum likelihood estimation (MLE) problem. If the observed signals $\{y_1, y_2\}$ are considered *incomplete*, then a simplified iterative approach to maximum likelihood estimation is given by the Expectation-Maximization (EM) algorithm [14]. In our approach, the unknown signals $\{s, s', d, d'\}$ are chosen to be the *complete* data and the EM algorithm is derived in the frequency domain for the above scenario. The result is a two-step sequential Wiener filtering scheme as shown in Figure 2. In the E-step, cross-talk in the reference channel is first cancelled using a frequency-domain Wiener filter, and the resulting reference channel noise estimate is filtered and used in a primary channel Wiener filter to obtain an estimate of desired speech. In the M-step, the spectral parameters of speech and noise along with the frequency-dependent correlation functions $\{H_1(\omega), H_2(\omega)\}$ are updated in such a way as to maximize the likelihood of the *complete* data estimates obtained in the E-step.

The algorithm iterates between the E-step and the M-step, and spectral parameters of speech for the first iteration are obtained from the noisy observation. In this case, the unconstrained dual-channel algorithm is seen to attain best possible quality in 2-3 iterations, which is not acceptable over all regions of speech [12]. The same observation has been demonstrated for single channel unconstrained Wiener filtering [15]. Limited quality improvement is seen for low-energy and transition regions of speech, resulting in limited overall quality, even though a mathematical error criterion is minimized via Wiener filtering. This provides a motivation for constraints based on auditory and perceptual properties of speech so as to reduce the impact of residual noise on speech parameters. The goal of such constraints is to allow the EM iterations to attain improved quality levels over all regions of speech. A brief discussion of the proposed constraints follows.

Evidence from auditory psychophysics suggests that the ear perceives speech along a non-linear scale in the frequency-domain. It is also known that critical band filtering of the speech spectrum is one possible method by which the human ear filters speech before converting it into a neural representation. An auditory based spectrum made up of critical band energies is closely related to perceptually important properties of speech. A detailed flowchart of the auditory based constraints is shown in Figure 3. The critical-band phenomenon is approximated along the *mel* scale by nineteen triangular band-pass filters with center frequencies linear upto 1 kHz and logarithmic at higher frequencies (sampling rate is 8 kHz). A Discrete Cosine Transform (DCT) of the logarithm of filter-band energies gives 9 mel-cepstral coefficients C_i. The first mel-cepstral parameter C_1 represents the energy balance between low and high frequencies, while higher mel-cepstral coefficients reflect increasing spectral detail. The behavior of each C_i is investigated for different speech sounds during the unconstrained EM iterations. Based upon the resulting observations, constraints are developed over time and iteration on individual C_i.

Constraints over time are adapted to changing speech sections differentiated as voiced, transitional, and unvoiced (v/t/uv) sections. This technique enables application of differing types and degrees of constraints suited to the particular speech section. The goal is to facilitate the best possible enhancement for a given speech section. Boundaries of such sections are provided using an adaptive boundary detector, which has been shown to perform reliably at higher signal-to-noise ratios [16]. For very high levels of noise however, performance of the boundary detector deteriorates. To remedy this, a two-pass enhancement scheme is proposed. In the first pass, unconstrained enhancement is performed to increase SNR, which may not improve quality but improves performance of the boundary detector. In the second pass, the new constrained iterative enhancement algorithm employs the detected speech boundaries to enhance quality of the original noisy signal.

Next, let us consider the specific constraints applied over the three speech sections. Speech sections classified as voiced possess mel-cepstral coefficients which have smooth variations from frame to frame in noise-free speech. In order to preserve these variations for mel-cepstral parameters during iterative enhancement, three types of constraints are applied. First, three frame median filtering is performed to remove any outliers which could result from a single noisy frame. Second, a polynomial is fitted using a least squares technique, over all frames in a given voiced section in order to smooth variation of the parameters over time. Third, the mean and variance values of the parameters over each section for a given iteration are linearly interpolated to values closer to the mean and variance values from the previous iteration. This last step was performed based on the observation that mean and variance values of mel-cepstral coefficients over voiced sections diverge from their noise-free values during unconstrained EM iterations. It has been shown [12, 13] that the above constraints do ensure that mean and variance values, and variation of the parameters over a given speech section, move towards those for noise-free original speech sections, as the constrained iterations proceed. An example of constraining the mean of C_1 for iteration n is given by

$$C'_{1(n)} = C_{1(n)} - \bar{C}_{1(n)} + \left[\frac{\bar{C}_{1(n)} - \bar{C}_{1(n-1)}}{k} + 1 \right] \bar{C}_{1(n-1)} \qquad (2)$$

where $\bar{C}_1$ is the mean of C_1 over all frames in a given voiced section, and k is an interpolation factor determined experimentally. For unvoiced sections, constraints on the mean are applied only over iterations. Median filtering and functional fitting over time are not performed in order to allow time variations to take their normal course. During transitional sections, due to the highly varying nature of speech spectrum, the behavior of mel-cepstral parameters is difficult to characterize. Hence, direct constraints are not applied during transitional sections, however the auditory spectrum transformation is performed to maintain continuity over the utterance. The constrained mel-cepstral parameters from all speech sections are then inverse transformed (IDCT), amplitude and frequency unwarped, compensated for triangular band-pass filtering, and linearly interpolated in order to obtain a sub-band smoothed and constrained linear speech spectrum for use in the primary channel Wiener filter. Evaluation of the proposed enhancement algorithm and the enhancement-coding tandem are discussed in the following section.

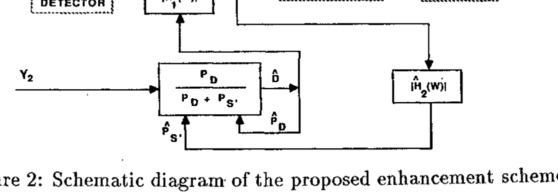

Figure 2: Schematic diagram of the proposed enhancement scheme.

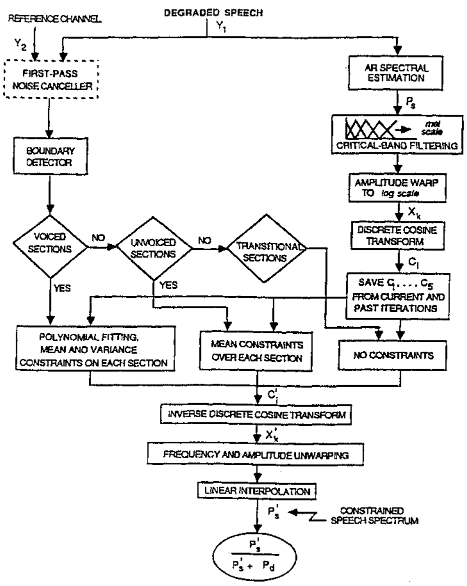

Figure 3: Algorithm flowchart of the proposed Constraints.

4 RESULTS AND EVALUATION

In this study, signal-to-noise ratio (SNR) is defined as the ratio of signal energy to noise energy in the primary channel, given in decibels (dB). Observations for three noise cases were simulated using digital samples of white Gaussian noise, aircraft cockpit noise from a C130 transport, and computer cooling fan noise. Evaluations are presented both for CELP coding in noise and the ACE-CELP tandem in four areas which include, i) overall objective quality measures across a subset of utterances from the NIST/TIMIT database, ii) distribution of the objective measures over all processed frames, iii) quality improvement over individual phonemes using objective measures, iv) speech spectrograms over a section of the processed utterances. Evaluations were performed using three objective quality measures [11], Itakura-Saito (IS) log-likelihood measure, log-area-ratio measure, and the Klatt measure. Here, only IS quality measures are presented, but performance as evaluated by each of the above quality measures was comparable.

Overall IS quality measures (mean and standard deviation values) over 20 sentences from the NIST/TIMIT database for three SNR values and the three noise cases, are presented in Table 1. It is seen that CELP causes further spectral distortion in both non-stationary colored noise cases, and maintains the distortion level in white Gaussian noise. The ACE algorithm as well as the ACE-CELP tandem result in significant quality improvement for white Gaussian noise and aircraft cockpit noise cases, especially at SNR levels of 10 and 20 dB. Finally, the ACE-CELP tandem is shown to improve both mean and standard deviation of the IS quality measure, indicating consistent quality improvement in severely degraded regions of speech. Quality improvement is also considerable for the computer fan noise relative to the amount of distortion CELP introduces, though the average quality measures for enhanced speech (especially at 5 and 10 dB) indicate persistent distortion. It is also noted that CELP performs very well in noise-free conditions with an average IS quality measure of 0.370, an almost imperceptible spectral distortion, which confirms several earlier studies.

Next, frame-to-frame quality measure histograms are presented in Figure 4, for CELP coding in noise-free, noisy (AWGN), and with front-end ACE scheme, over all 20 utterances ($\approx$ 7000 frames of speech). Histogram concentration about the mean indicates that CELP performs consistently over all regions of speech in noise-free conditions (4-i). CELP performance degrades rapidly over some frames of speech in noisy conditions (4-ii), and the ACE-CELP tandem is seen to improve performance consistency (4-iii).

ITAKURA-SAITO OVERALL SPEECH QUALITY MEASURE								
Noise-free CELP			$\bar{m} = 0.370$			$\sigma = 0.411$		
SNR	DEG		DEG-CELP		ACE		ACE-CELP	
dB	$\bar{m}$	σ	$\bar{m}$	σ	$\bar{m}$	σ	$\bar{m}$	σ
Additive White Gaussian Noise								
20.0	1.197	2.095	1.183	1.896	0.550	0.667	0.762	0.642
10.0	2.566	4.046	2.446	4.179	0.948	1.007	1.074	1.027
5.0	3.408	5.006	3.296	5.227	1.252	1.405	1.302	1.308
Aircraft Cockpit Noise								
20.0	1.354	3.872	2.138	5.360	0.579	2.215	0.742	2.184
10.0	3.388	8.214	4.835	11.27	1.594	5.457	1.606	5.270
5.0	4.894	11.30	6.411	13.76	2.467	7.835	2.415	7.720
Computer Cooling Fan Noise								
20.0	3.480	9.136	4.768	12.16	1.399	6.135	1.543	7.167
10.0	6.587	13.55	8.496	17.09	3.484	10.93	3.327	10.67
5.0	8.499	15.60	10.71	19.43	5.082	13.11	4.792	12.98

Table 1: Global IS quality measures over 20 sentences for i) DEG (between original and degraded), ii) DEG-CELP (between original and CELP coded in additive noise, iii) ACE (between original and enhanced using auditory constraints), iv) ACE-CELP (between original and output of ACE-CELP tandem).

It has been shown [12], that dual-channel ACE performs reliably in low to moderate levels of cross-talk. For the ACE-CELP tandem evaluations presented above, cross-talk level was assumed to be zero. An example of performance in the presence of cross-talk is illustrated in the following IS quality measures for a single utterance : 3.629 for CELP in 5 dB of white Gaussian noise, 1.183 for the ACE-CELP tandem without cross-talk, 1.475 for the ACE-CELP tandem with a moderate cross-talk-to-noise ratio of -15 dB, which indicates good performance for the ACE-CELP tandem in the presence of cross-talk. The enhancement scheme was also tested as a post-processor (i.e. enhancement was performed after CELP resynthesis in white Gaussian noise situations in an attempt to remove residual noise). This is necessary in cases where only the noisy coded speech is available. There is quality improvement for ACE as a post-processor with an IS quality measure of 1.87 as compared to 3.629 for noisy CELP at a SNR of 5 dB, but spectral distortion is higher compared to 1.183 for the ACE-CELP tandem. This may be due in part to the lack of a reliable noise reference channel. Further study is necessary in order to establish ACE as a post-processor.

Quality measures over individual phonemes in Table 2, and speech spectrograms in Figure 5, are presented to further illustrate the improvement obtained with the ACE-CELP tandem. The ACE-CELP tandem was tested on 100 sentences from the NIST/TIMIT data-base for the case of AWGN at an SNR value of 5 dB. Frame-to-frame IS quality measures were classified over individual phonemes for CELP

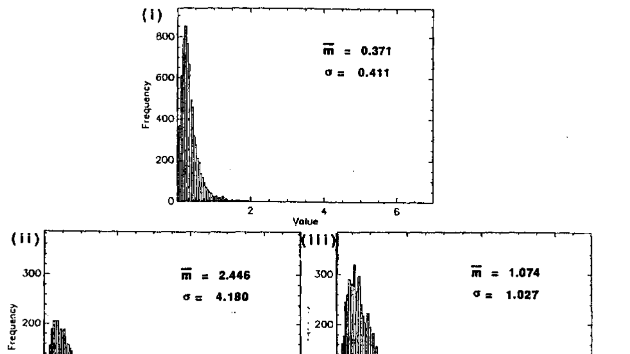

Figure 4: Histograms of frame-to-frame IS quality measures for i) CELP in noise-free ii) CELP in AWGN, SNR = 10 dB iii) ACE-CELP tandem.

without enhancement and with front-end enhancement. Quality improvement is seen over virtually all phonemes, with particular improvement over phonemes such as stops, liquids, nasals, and glides which are highly sensitive to background noise. Quality improvement is seen to be consistent over all processed sentences. Next, speech spectrograms for the word "popularity" for the cases of (i) noise-free original, (ii) CELP in noise-free, (iii) CELP in degraded, and (iv) ACE-CELP tandem are presented in Figure 5. CELP performs very well in noise-free conditions, accurately representing formant location, bandwidth, and movement over time; but CELP for degraded speech is seen to misrepresent the majority of spectral characteristics over the displayed speech section. The ACE-CELP tandem is seen to restore formant tracks and bandwidths at low and mid frequencies, thereby confirming quality improvement. Informal listening tests confirmed that the ACE-CELP tandem reduces most of the background noise without introducing any artifacts.

5 CONCLUSIONS

A new dual-channel iterative speech enhancement algorithm has been proposed and evaluated as a front-end enhancer for CELP coding in noisy environments. Constraints based on an auditory spectrum are developed and applied during iterations to take advantage of auditory and perceptual properties of speech. Auditory constrained enhancement (ACE) results in improved quality, which is consistent over a large set of sentences from the NIST/TIMIT database. CELP coding is evaluated for three cases of background noise using objective quality measures, both as stand-alone and with front-end enhancement. Overall speech quality is shown to deteriorate for CELP coding in noise, with significant quality loss in non-stationary colored noise. Overall quality and quality over individual phonemes indicate consistent improvement using the ACE-CELP tandem. Improvement is reflected in a mean shift for frame-to-frame Itakura-Saito quality measure of 2.44 to 1.07, with a corresponding decrease in standard deviation from 4.18 to 1.02, for an SNR value of 10 dB. Speech spectrograms and informal listening tests confirm the quality improvement obtained using the ACE-CELP tandem.

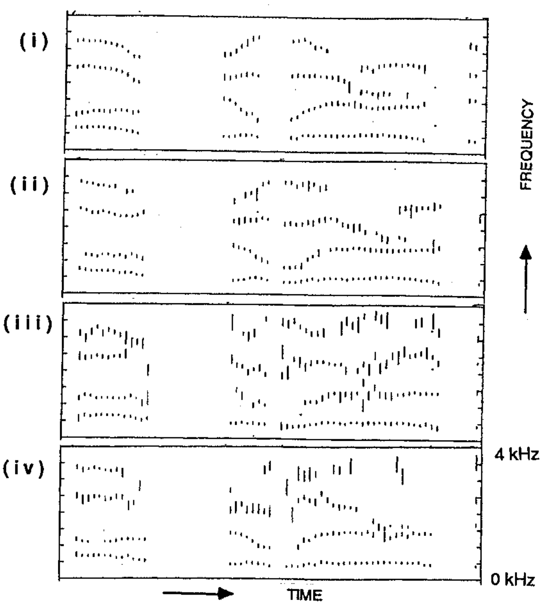

Figure 5: Speech spectrograms for i) original noise-free ii) CELP in noise-free iii) CELP in AWGN, SNR = 10 dB iv) ACE-CELP tandem, for the word "popularity".

[1]We wish to thank Dr. D. Pallett of NIST for providing data, and DoD for providing their release of CELP.

Itakura-Saito Likelihood Measure							
Phon -eme	DEG CELP	ACE CELP	Frame Count	Phon -eme	DEG CELP	ACE CELP	Frame Count
Vowels							
/aa/	3.552	1.493	1339	/ae/	1.805	1.262	977
/axr/	11.29	3.757	594	/ix/	3.708	1.529	1043
/ah/	2.867	1.150	625	/oy/	4.300	1.347	171
Nasals							
/n/	8.480	3.124	1153	/m/	8.502	2.695	683
Stops							
/p/	2.963	1.283	508	/t/	1.948	1.362	542
/b/	2.938	1.033	135	/d/	1.679	1.060	186
Fricatives							
/sh/	1.206	2.035	673	/dh/	3.791	1.288	270
Glides							
/r/	10.78	2.901	747	/y/	2.271	1.675	318
Liquids							
/w/	6.964	2.061	289	/l/	4.913	1.657	1079
Overall DEG-CELP		*Overall ACE-CELP*			*Frame Count*		
3.436		1.653			36006		

Table 2: IS quality measures for some phonemes over a 100 sentence subset from the NIST/TIMIT database, SNR = 5.0 dB

References

[1] D. Kemp, R. Sueda, T. Tremain, "An evaluation of 4800 bps Voice Coders," *Proc. 1989 IEEE ICASSP*, pp. 200-203, May 1989.

[2] R. Cox, W. Kleijn, P. Kroon, "Robust CELP Coders for Noisy background and Noisy channels," *Proc. 1989 IEEE ICASSP*, pp. 739-742, May 1989.

[3] J. Campbell,V. Welch, T. Tremain, "An Expandable Error-Protected 4800 bps CELP coder," *Proc. 1989 IEEE ICASSP*, pp. 735-738, May 1989.

[4] J. Campbell, T. Tremain, V. Welch, "The DOD 4.8 kbps Standard (Proposed Federal Standard 1016)," *Advances in Speech Coding*, pp. 122-133, Kluwer Academic Publishers, 1991.

[5] M. Schroeder, B. Atal, "Code-Excited Linear Prediction (CELP): high quality speech at very low bit rates," *Proc. 1985 IEEE ICASSP*, pp. 937-940, March 1985.

[6] P. Kroon, B. Atal, "Quantization procedures for 4.8 kbps CELP coders," *Proc. 1987 IEEE ICASSP*, pp. 1650-1654, April 1987.

[7] C. Teacher, D. Coulter, "Performance of LPC Vocoders in a Noisy Environment," *Proc. 1979 IEEE ICASSP*, pp. 216-219, April 1979.

[8] J. H. L. Hansen, S. Nandkumar, "Speech Quality Assessment of Real-Time Speech Coding Algorithms for varying Noise and Language Conditions," *Proc. 1992 EUSIPCO*, pp. S6.L-7.1-4, Belgium, August 1992.

[9] P. Kroon, E. F. Deprettere, R. J. Sluyter, "Regular Pulse Excitation-A Novel Approach to Effective and Efficient Multipulse Coding of Speech," *IEEE Trans. on ASSP*, pp.1054-1063, Oct.1986.

[10] J. S. Lim, A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. of IEEE*, pp. 1586-1604, Dec. 1979.

[11] S. R. Quackenbush, T. P. Barnwell, M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, 1988.

[12] S. Nandkumar, J.H.L. Hansen, "Dual-Channel Speech Enhancement with Auditory Spectrum based Constraints," *Proc. 1992 IEEE ICASSP*, Vol.I, pp. 297-300, March 1992.

[13] S. Nandkumar, J.H.L. Hansen, "Dual-Channel Speech Enhancement with Constraints based on an Auditory Spectrum,"submitted to *IEEE Trans. on Sig. Proc.*, Feb. 1992.

[14] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, Ser.39, pp.1-38, 1977.

[15] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Transactions on Sig. Proc.*, vol.39, no.4, pp.795-805, April 1991.

[16] J. H. L. Hansen, "A New Speech Enhancement Algorithm employing Acoustic Endpoint Detection and Morphological based Spectral Constraints," *Proc. 1991 IEEE ICASSP*, pp. 901-904, May 1991.