IV. Conclusion

In this work, it is shown that the actual energy of analysis frames should be taken into account for interpolation. The required approximation of the sample autocorrelation function can be implemented by multiplying the autocorrelation coefficients with the frame energy and interpolating this function (ACF interpolation). ACF interpolation outperformed LSP interpolation in a subjective test, contrasting the objective results.

The main reason for the discrepancy between subjective and objective results is that the largest outliers occur in low energy parts of segments with rapidly changing energy and it turned out that these do not have much influence on the subjective quality.

References

[1] F. Itakura, "Line spectral representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, p. S35, 1975.

[2] V. R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 309–321, 1975.

[3] A. H. Gray and J. D. Markel, "Quantization and bit allocation in speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 459–473, 1976.

[4] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[5] B. S. Atal, R. V. Cox, and P. Kroon, "Spectral quantization and interpolation for CELP coders," in *Proc. Int. Conf. ICASSP*, 1989, pp. 69–72.

[6] T. Umezaki and F. Itakura, "Analysis of time fluctuating characteristics of linear predictive coefficients," in *Proc. Int. Conf. ICASSP*, 1986, pp. 1257–1260.

[7] M. Yong, "A new LPC interpolation technique for CELP coders," *IEEE Trans. Commun.*, vol. 42, pp. 34–38, 1994.

[8] K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," in *Proc. Int. Conf. EUROSPEECH*, 1995, pp. 1029–1032.

[9] H. B. Choi, W. T. K. Wong, B. M. G. Cheetham, and C. C. Goodyear, "Interpolation of spectral information for low bit rate speech coding," in *Proc. Int. Conf. EUROSPEECH*, 1995, pp. 1033–1036.

[10] J. S. Erkelens and P. M. T. Broersen, "Interpolation of autoregressive processes at discontinuities: Application to LPC based speech coding," in *Proc. Int. Conf. EUSIPCO*, 1994, pp. 935–938.

[11] R. Hagen, E. Paksoy, and A. Gersho, "Variable rate spectral quantization for phonetically classified CELP coding," in *Proc. Int. Conf. ICASSP*, 1995, pp. 748–751.

[12] I. A. Atkinson, A. M. Kondoz, and B. G. Evans, "1.6 kbit/s LP vocoder using time envelope," *Electron. Lett.*, vol. 31, pp. 517–519, 1995.

[13] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," in *Proc. IEEE*, 1985, vol. 73, pp. 1551–1588.

[14] J. S. Erkelens and P. M. T. Broersen, "Quantization of the LPC model with the reconstruction error distortion measure," in *Proc. Int. Conf. EUSIPCO*, 1996, pp. 1677–1680.

[15] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 3–14, 1993.

[16] J. S. Erkelens and P. M. T. Broersen, "Bias propagation in the autocorrelation method of linear prediction," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 116–119, 1997.

[17] J. S. Erkelens, *Autoregressive Modeling for Speech Coding: Estimation, Interpolation and Quantization*. Delft, The Netherlands: Delft Univ. Press, 1996.

An Improved (Auto:I, LSP:T) Constrained Iterative Speech Enhancement for Colored Noise Environments

Bryan L. Pellom and John H. L. Hansen

***Abstract*—In this correspondence we illustrate how the (Auto:I, LSP:T) constrained iterative speech enhancement algorithm can be extended to provide improved performance in colored noise environments. The modified algorithm, referred to here as *noise adaptive* (*Auto:I, LSP:T*), operates on subbanded signal components in which the terminating iteration is adjusted based on the *a posteriori* estimate of the signal-to-noise ratio (SNR) in each signal subband. The enhanced speech is formulated as a combined estimate from individual signal subband estimators. The algorithm is shown to improve objective speech quality in additive noise environments over the traditional constrained iterative (Auto:I, LSP:T) enhancement formulation.**

I. Introduction

There are numerous areas where it is necessary to enhance the quality of speech that has been degraded by background distortion. Some of these environments include aircraft cockpits, automobile interiors for hands-free cellular, and voice communications using mobile telephone. Speech enhancement under these conditions can be considered successful if it i) suppresses perceptual background noise and ii) either preserves or enhances perceived speech quality. As voice technology continues to mature, greater interest and demand is placed on using voice-based speech algorithms in diverse, adverse, environmental conditions. It is suggested that the success of advancing speech research in the fields of speaker verification, language identification, and automatic speech recognition could be improved by incorporating front-end speech enhancement algorithms [1].

A number of speech enhancement algorithms have been proposed in the past. A survey can be found in [2], as well as an overview of statistical based approaches in [3]. Several enhancement approaches have been proposed using improved signal-to-noise ratio (SNR) characterization [4], linear and nonlinear spectral subtraction [5], [6], and Wiener filtering [7]. Traditional speech enhancement methods are based on optimizing mathematical criteria, which in general are not always well correlated with speech perception. Several recent methods have also considered auditory processing information [8], [9], and constrained iterative methods using various levels of speech class knowledge [10]–[12].

In this study, we focus on an extension to a previously proposed constrained iterative speech enhancement algorithm termed (Auto:I, LSP:T)[1] [10] (described briefly in Section II). Basically, this method employs spectral constraints on the input speech feature sequence across time and iterations to ensure more natural

Manuscript received February 26, 1997; revised February 26, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-Claude Junqua.

The authors are with the Department of Electrical Engineering, Robust Speech Processing Laboratory, Duke University, Durham, NC 27708-0291 USA (e-mail: jhlh@ee.duke.edu).

Publisher Item Identifier S 1063-6676(98)07793-1.

[1] The term (Auto:I, LSP:T) formulated in [10] is derived from the notion that spectral constraints are applied across iterations (I) to the speech autocorrelation lags as well as across time (T) to the speech line spectrum pair (LSP) parameters. For simplicity, (Auto:I, LSP:T) will be referred to as *Auto-LSP* throughout this work.

sounding enhanced speech with little processing artifacts. The constraints are applied based on speech production ideas from estimated broad phoneme classes. Since the method employs an iterative Wiener filter, the proper terminating iteration must be obtained from prior simulation in the desired noise conditions. A revised class-directed (CD-Auto-LSP) algorithm employed a noisy trained hidden Markov model recognizer to classify input phoneme classes, so that a class dependent terminating iteration could be applied [11]. This resulted in improved speech quality consistency for speech degraded with white Gaussian noise (WGN) from the TIMIT data base. Other constrained iterative methods (ACE-I, ACE-II) have been proposed by Nandkumar and Hansen [9], [12] which address colored noise using a dual-channel framework with various auditory processing constraints such as critical-band filtering, intensity-to-loudness conversion, and lateral neural inhibition. While previous single-channel methods such Auto-LSP and CD-Auto-LSP have been successful in white noise environments, their constraints have not been specifically formulated to address the changing structure of colored background noise. Methods such as ACE and adaptive noise canceling [13] address this via a second reference channel. In this study, we propose to reformulate the manner by which spectral constraints are applied within the Auto-LSP enhancement algorithm to specifically address the nonuniform impact colored noise will have on degraded speech. As such, when background noise levels are high, constraints will be tightened, especially in regions where smooth spectral transitions should take place (i.e., voiced transitions from vowels to semivowels). For portions of the frequency domain where the SNR is high, spectral constraints will be either relaxed or disabled, since such constraints could alter the natural spectral structure of speech in these clean regions. This paper is organized as follows. In Section II, we present details of the Auto-LSP enhancement algorithm. Next, the noise adaptive Auto-LSP enhancement algorithm is proposed in Section III, followed by algorithm evaluations in Section IV. Finally, we draw conclusions in Section V.

II. Auto-LSP Enhancement

The constrained iterative Auto-LSP enhancement approach is based upon extensions to the two-step maximum *a posteriori* (MAP) estimation of the all-pole speech parameters and noise-free speech formulated by Lim and Oppenheim [7]. In the unconstrained MAP estimation procedure, the ℓth frame of speech is modeled by a set of all-pole linear predictive parameters $\vec{a}_\ell$ and gain g_ℓ. The estimation process iterates between two sequential MAP estimations. For the ith algorithm iteration, the all-pole speech model parameters $\hat{\vec{a}}_\ell^{(i)}$ are first obtained from the estimated noise-free speech at the $(i-1)$th iteration, $\hat{\vec{S}}_\ell^{(i-1)}$. In the second step, a MAP estimate of the noise-free speech is obtained by applying a noncausal Wiener filter to $\hat{\vec{S}}_\ell^{(i-1)}$. Here, the frequency domain filter is constructed using the all-pole model spectrum described by $\hat{\vec{a}}_\ell^{(i)}$ as an estimate of the noise-free speech power spectrum. The estimation process at the ith iteration can be described by

$$\mathrm{MAX}\, p\big(\hat{\vec{a}}_\ell^{(i)} \,\big|\, \hat{\vec{S}}_\ell^{(i-1)}, g_\ell\big) \text{ which gives } \hat{\vec{a}}_\ell^{(i)} \tag{1}$$

$$\mathrm{MAX}\, p\big(\hat{\vec{S}}_\ell^{(i)} \,\big|\, \hat{\vec{a}}_\ell^{(i)}, \hat{\vec{S}}_\ell^{(i-1)}, g_\ell\big) \text{ which gives } \hat{\vec{S}}_\ell^{(i)} \tag{2}$$

where $\hat{\vec{S}}_\ell^{(0)}$ represents the original noise-corrupted frame of speech. The two-step procedure is repeated until an *a priori* terminating criterion is satisfied.

In the constrained iterative approach [10], spectral constraints are applied between MAP estimation steps in order to ensure 1) stability of the all-pole model, 2) that it possess speech-like characteristics (e.g., natural formant bandwidths), and 3) to provide frame-to-frame continuity in vocal tract characteristics. In particular, two types of spectral constraints known as *interframe* and *intraframe* constraints are applied to the speech spectrum during the iterative all-pole parameter estimation. Interframe constraints are applied over time to the LSP position and difference parameters in order to reduce frame-to-frame pole jitter and to ensure that the enhanced speech has speech-like characteristics. For the jth LSP position parameter computed from the ℓth frame on the ith iteration, $p_\ell^{(i)}(j)$, the spectral constraint is implemented by smoothing over an adaptive triangular base of support of width $2N(j)+1$ frames,

$$\hat{p}_\ell^{(i)}(j) = \sum_{k=-N(j)}^{N(j)} H(E_\ell, j) \cdot \left[1 - \frac{|k|}{W(E_\ell, j)}\right] \cdot p_{\ell+k}^{(i)}(j) \qquad \forall j = 1, \cdots, 5 \tag{3}$$

where $H(\cdot)$ and $W(\cdot)$ represents the smoothing window height and width which are dependent upon both frame energy E_ℓ and LSP parameter index j. In addition to LSP position parameter smoothing, constraints are applied to the LSP difference parameters in order to ensure that the pole locations do not drift too close to the unit circle causing unnatural formant bandwidths in the enhanced speech.

The second type of constraint, known as intraframe constraints, are applied across iterations to the autocorrelation parameters in order to control the rate of improved estimation for phoneme sections less sensitive to noise. This relaxation constraint is implemented by estimating the kth autocorrelation lag as a weighted combination of the kth lag from M previous iterations. Specifically

$$R_\ell^{(i)}[k] = \sum_{m=0}^{M} \psi_m R_\ell^{(i-m)}[k] \tag{4}$$

with the condition that $\sum_{m=0}^{M} \psi_m = 1$.

The constrained iterative enhancement algorithm was formulated using an additive white Gaussian noise (WGN) assumption. As such, the method has been shown to be successful in WGN environments, with some improvement for colored noise sources as well. In WGN environments, the incorporation of spectral constraints was shown to provide a more consistent terminating iteration and improved objective speech quality over the unconstrained iterative enhancement method [7].

III. Noise Adaptive Auto-LSP Enhancement

In many real-world settings, such as aircraft cockpit or automobile environments, the spectral content of the degrading noise is not flat, but rather concentrated within a small portion of the frequency spectrum. This may result in only a localized degradation of speech quality over a finite frequency interval. Furthermore, due to the time-varying nature of speech, the local SNR across both time and frequency may differ dramatically from frame-to-frame. In the Auto-LSP formulation described in Section II, inter- and intraframe spectral constraints are applied to the speech signal at each iteration regardless of the spectral content of the noise. In low-frequency distortions, such as automobile highway noise, it is undesirable to apply spectral smoothing constraints to regions of high frequency, since this can reduce the quality of the high SNR spectral components. In theory, spectral based speech constraints should be selectively applied only to regions of the speech signal which have been corrupted by noise. In other words, either a soft-decision or hard-decision is needed to determine when constraints should be applied.

As a consequence, we propose an extension to the Auto-LSP enhancement algorithm for colored noise environments by considering the decomposition of the estimated enhanced speech signal into a set of Q frequency subbands. Here, we assume that the degrading noise will impact each subband differently and hence, the terminating iteration should be appropriately adjusted for each time-frequency partition. By reducing the terminating iteration in spectral regions of high SNR, spectral smoothing is reduced and speech quality is maintained. In a similar manner, by increasing the terminating iteration in spectral regions of low SNR, noise attenuation can be improved. Hence, selecting an appropriate terminating iteration based on the presence of noise in each signal subband provides a better compromise between signal distortion and noise attenuation.

In the proposed framework, we consider the speech signal as being comprised of a set of Q frequency bands which uniformly partition the linear frequency scale. The speech signal $s(n)$ can be expressed as the sum of individual subband components

$$s(n) = \sum_{k=1}^{Q} s(n;k) = \sum_{k=1}^{Q} \sum_{m=0}^{M_k - 1} h(m;k) s(n-m) \tag{5}$$

where $s(n;k)$ represents the time-domain output of the kth filter. Although in this formulation we assume a uniform bank of bandpass filters, other filterbank decompositions such as those based on models of auditory perception could also be used [9], [12]. Using frame-oriented processing of the subband filtered speech $s(n;k)$, the algorithm is summarized as follows (n: sample value, ℓ: frame index, i: iteration, k: frequency band).

1. Initialization:

   a) Decompose the ℓth degraded speech frame, $s_\ell(n)$, into subband signal components $s_\ell(n;k)$. Compute the signal energy in each subband component

   $$E_\ell(k) = \sum_n s_\ell^2(n;k).$$

   b) Estimate average noise energy, $\hat{E}_{\text{noise}}(k)$, in each subband from N most recent frames classified as noise-only (silence) segments

   $$\hat{E}_{\text{noise}}(k) = \frac{1}{N} \sum_{j=1}^{N} E_{nf(j)}(k)$$

   where $nf(j)$ represents the index of the jth most recent frame of noise-only activity.

   c) Compute an estimate of the *a posteriori* SNR (in dB) for each signal subband

   $$\text{SNR}_\ell(k) = 10 \log_{10} \left(\frac{E_\ell(k)}{\hat{E}_{\text{noise}}(k)} - 1 \right)$$

   where the local SNR in each time-frequency band is constrained to range from -5 to 25 dB.

   d) Assign a terminating iteration, $\text{ITER}_\ell(k)$ to each signal subband "k" and frame "ℓ" based on the local SNR estimate in each band

   $$\text{ITER}_\ell(k) = \text{int}\left\{ (\text{ITER}_{\max} - \text{ITER}_{\min}) \times \left(\frac{\text{SNR}_{\max} - \text{SNR}_\ell(k)}{\text{SNR}_{\max} - \text{SNR}_{\min}} \right) \right\} + \text{ITER}_{\min}$$

   where $\text{int}\{\cdot\}$ rounds to the closest integer, $\text{SNR}_{\max} = 25$ dB and $\text{SNR}_{\min} = -5$ dB. $\text{ITER}_{\max}$ and $\text{ITER}_{\min}$ represent the maximum and minimum terminating iteration allowed in each signal subband.

2. Iterative Estimation:

   a) Obtain enhanced speech frame from the ith iteration, $\hat{s}_\ell^{(i)}(n)$, from Auto-LSP.

   b) Decompose $\hat{s}_\ell^{(i)}(n)$ into Q subband components. If the terminating iteration for the current subband component equals the current iteration $(\text{ITER}_\ell(k) = i)$, then retain the kth subband component as a final estimate for the current subband.

   c) Repeat (a) to obtain estimate for the $(i+1)$th iteration until terminating iteration, $\text{ITER}_{\max}$, is reached.

3. Signal Reconstruction:

   a) For each frame, sum the retained subband components from step 2 and recover the enhanced speech frame.

   $$\hat{s}_\ell(n) = \sum_{k=1}^{Q} \hat{s}_\ell(n;k)$$

   b) Recover final enhanced speech signal using standard overlap and add procedure.

In summary, an estimate of the local *a posteriori* SNR is computed on a frame-by-frame basis in each signal subband in order to select a local terminating iteration. For real-time enhancement applications, the noise energy in each signal subband (and noise power spectral estimate) can be updated during periods of silence or speaker pause. Consequently, local SNR estimates will in general depend on the most recent estimate of the noise energy corrupting each subband. In this work, we consider a linear relationship between the local SNR estimate (measured in dB) and terminating iteration selection and constrain the amount of iterations to range between $\text{ITER}_{\min}$ to $\text{ITER}_{\max}$ within each signal subband. A reasonable value for $\text{ITER}_{\min}$ is one and a reasonable value for $\text{ITER}_{\max}$ is between 4 and 7. In general, the specific choice of either parameter will depend on global SNR characteristics of the observed noise-corrupted speech. We will refer to the proposed algorithm as *noise adaptive* Auto-LSP due to the adaptation of the terminating iteration based on the presence of noise in each time-frequency signal component. An overall block diagram of the proposed algorithm is illustrated in Fig. 1.

IV. Algorithm Evaluations

A. Evaluation Data Base and Noise Sources

In order to examine the effectiveness of the proposed algorithm in a variety of additive noise environments, ten additive noises summarized in Table I were used for evaluation.[2] Aircraft cockpit, automobile highway, and helicopter fly-by noise are slowly varying low-frequency distortions. Large city, city in the rain, and large crowd noise exhibit slowly varying spectral characteristics. IBM PS-2 cooling fan noise is primarily a stationary low-frequency distortion, while that of the Sun 4/330 Workstation is primarily a stationary higher-frequency distortion. Furthermore, the cooling fan spectra include a prominent spectral peak due to the rotation of the fan blades (approximately 305 Hz for IBM PS-2 cooling fan and 3075 Hz for Sun cooling fan noise).

[2] The same noise sources were used for speech recognition evaluations in [1] and can be obtained from the web address http://www.ee.duke.edu/Research/Speech/rspl_software.html.

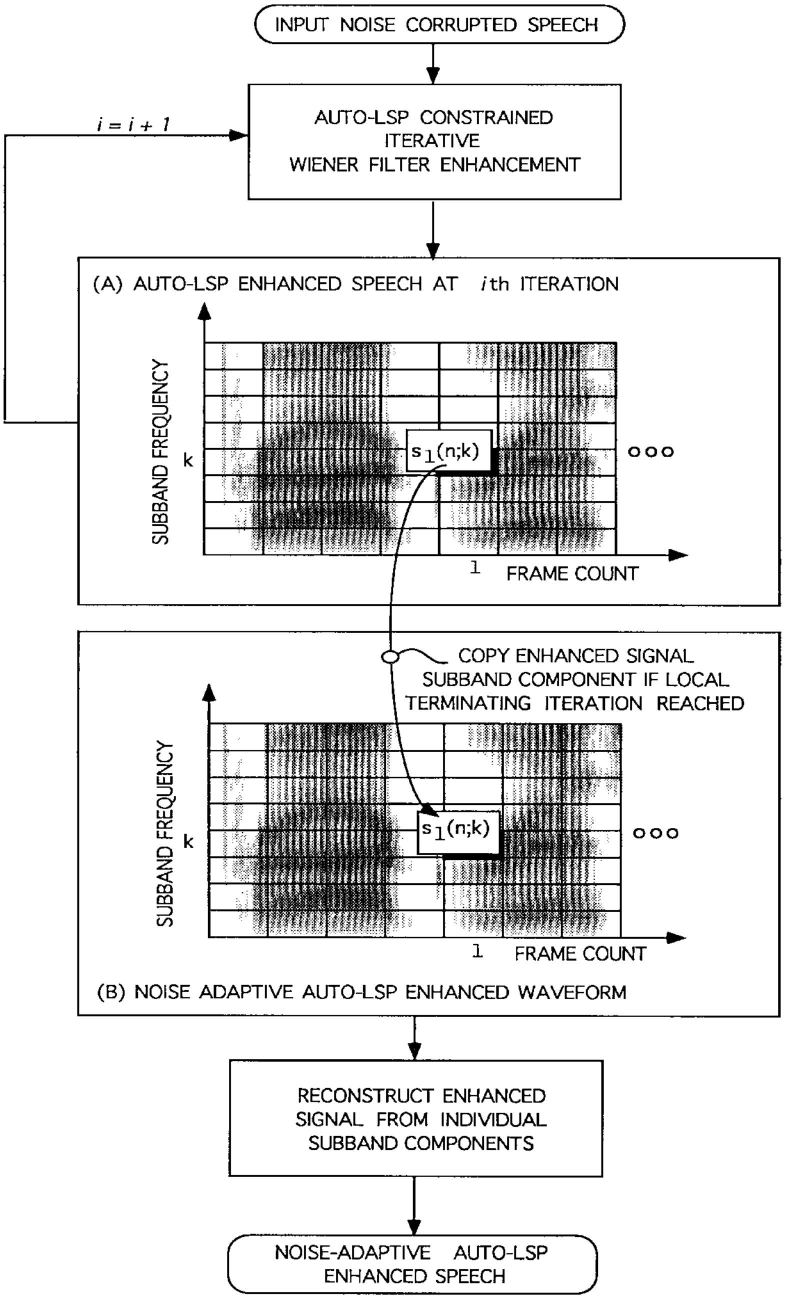
INPUT NOISE CORRUPTED SPEECH
$i = i + 1$
AUTO-LSP CONSTRAINED ITERATIVE WIENER FILTER ENHANCEMENT
(A) AUTO-LSP ENHANCED SPEECH AT ith ITERATION
SUBBAND FREQUENCY
k
$s_1(n;k)$
ooo
1 FRAME COUNT
COPY ENHANCED SIGNAL SUBBAND COMPONENT IF LOCAL TERMINATING ITERATION REACHED
SUBBAND FREQUENCY
k
$s_1(n;k)$
ooo
1 FRAME COUNT
(B) NOISE ADAPTIVE AUTO-LSP ENHANCED WAVEFORM
RECONSTRUCT ENHANCED SIGNAL FROM INDIVIDUAL SUBBAND COMPONENTS
NOISE-ADAPTIVE AUTO-LSP ENHANCED SPEECH

Fig. 1. Noise adaptive constrained iterative speech enhancement.

TABLE I
ADDITIVE NOISES CONSIDERED FOR ENHANCEMENT EVALUATION

Noise	*Description*
Aircraft Cockpit	Noise recorded from the cockpit of a C130 transport plane
Automobile Highway	Noise recorded in a car traveling 95 Km/hour
Helicopter Fly-By	Noise recorded as a helicopter flew overhead
Large City	Noise recorded on the streets of a large city
City Rain	Noise recorded during a rainstorm
Large Crowd	Noise recorded from a crowded auditorium
PS-2 Cooling Fan	Noise recorded from the cooling fan of an IBM PS-2/80 computer
Sun Cooling Fan	Noise recorded from the cooling fan of a Sun 4/330 Workstation
Flat Communication	Noise recorded from a flat communications channel
White Gaussian	Computer generated white Gaussian noise

TABLE II

OBJECTIVE SPEECH QUALITY VERSUS SNR FOR ORIGINAL DEGRADED SPEECH (100 8 kHz SAMPLED TIMIT SENTENCES WITH ADDITIVE NOISE), ENHANCED SPEECH PROCESSED WITH AUTO-LSP AND THE PROPOSED NOISE ADAPTIVE AUTO-LSP ALGORITHM

Itakura-Saito Likelihood Measure under 10 Environmental Noises

Noise Type	*Global SNR (dB)*	*Original Degraded*	*Enhanced Auto-LSP*	*Enhanced Noise Adaptive*
Aircraft Cockpit	15	1.76	0.72	0.58
	10	2.94	1.24	1.03
	5	4.36	2.03	1.73
Automobile Highway	15	4.08	1.84	1.39
	10	6.34	2.71	2.22
	5	8.37	4.01	3.39
Helicopter Fly-By	15	1.66	1.31	0.93
	10	2.57	1.88	1.42
	5	3.80	2.75	2.22
Large City	15	0.75	0.61	0.56
	10	1.17	1.02	0.94
	5	1.70	1.63	1.49
City Rain	15	0.70	0.69	0.65
	10	1.09	1.12	1.06
	5	1.55	1.74	1.61
Large Crowd	15	0.71	0.72	0.66
	10	1.11	1.15	1.06
	5	1.60	1.78	1.65
PS-2 Cooling Fan	15	2.79	1.55	1.24
	10	4.20	2.06	1.75
	5	5.94	2.82	2.53
Sun Cooling Fan	15	1.04	0.66	0.61
	10	1.63	1.06	0.98
	5	2.34	1.66	1.54
Flat Communication	15	1.00	0.73	0.69
	10	1.54	1.16	1.10
	5	2.17	1.74	1.66
White Gaussian	15	1.78	1.21	1.11
	10	2.67	1.92	1.76
	5	3.68	2.87	2.63

B. Evaluation Method

The proposed noise adaptive Auto-LSP enhancement algorithm was evaluated by adding a controlled level of noise to 100 sentences extracted from an 8 kHz lowpass filtered version of the TIMIT data base. For each noise type, global SNR's of 5, 10, and 15 dB were considered. In this study, objective speech measures [14] were used for algorithm evaluation. For each degraded utterance, the Itakura–Saito (IS) likelihood measure was calculated before and after enhancement processing. The frame-based IS likelihood measure for a (clean) reference frame x_ϕ and (noisy) test frame x_d is given by

$$d_{\mathrm{IS}}(x_\phi, x_d) = \int_{-\pi}^{+\pi} [e^{V(\theta)} - V(\theta) - 1] \frac{d\theta}{2\pi} \tag{6}$$

where

$$V(\theta) = \log\left(\frac{\sigma_\phi^2}{|A_\phi(e^{j\theta})|^2}\right) - \log\left(\frac{\sigma_d^2}{|A_d(e^{j\theta})|^2}\right). \tag{7}$$

Here, $A_d(e^{j\theta})$ and $A_\phi(e^{j\theta})$ represent the linear prediction analysis filters for the (noisy) test frame x_d and (clean) reference frame x_ϕ. A measure of global sentence quality was then determined by computing the average of the frame-based measures across speech-only sections of each utterance.

For the noise adaptive approach, a total of eight signal subband components that uniformly partitioned the linear frequency scale were utilized. Furthermore, the terminating iteration in each signal subband was constrained to range from one to four iterations. The Auto-LSP algorithm was terminated at the fourth iteration. This was found to provide the best overall objective speech quality during informal experimentation using several additive noise sources. During enhancement processing, the noise power spectrum was estimated from the first 880 samples (110 ms) of silence at the beginning of each utterance. Note that a one-time estimate of the noise was used since each TIMIT utterance contains approximately 3 s of speech activity with little or no pause between words.

C. Evaluation Results

Results of the algorithm evaluations are summarized in Table II. Here, the IS likelihood measure for the original degraded speech, enhanced speech processed using traditional Auto-LSP, and enhanced speech processed using the proposed noise adaptive Auto-LSP algorithm is shown. Considering SNR's ranging from 5 to 15 dB, we see that both enhancement approaches reduce spectral distortion and improve objective speech quality (i.e., reduced IS measures after processing reflect less spectral mismatch). For example, the mean IS measure for speech degraded with aircraft cockpit noise at 10 dB SNR is 2.94 before enhancement, 1.24 after Auto-LSP enhancement, and further reduced to 1.03 using the proposed noise adaptive Auto-LSP algorithm. Furthermore, we see that the difference in IS measures between speech processed using Auto-LSP and the proposed algorithm is most dramatic for colored noises while less dramatic for noises that are almost spectrally flat. This can be partially attributed to the ability of the proposed algorithm to adaptively

TABLE III
OBJECTIVE SPEECH QUALITY VERSUS BROAD PHONEME CLASSIFICATION. HERE, 100 TIMIT SENTENCES WERE DEGRADED WITH ADDITIVE AIRCRAFT COCKPIT NOISE (10 dB SNR) AND SUBSEQUENTLY ENHANCED USING AUTO-LSP AND NOISE ADAPTIVE AUTO-LSP

Objective Quality versus Broad Phoneme Classification
100 TIMIT sentences, Aircraft Cockpit Noise, 10 dB SNR

Sound	*Itakura-Saito Likelihood Measure*			
Type	*Degraded*	*Auto-LSP*	*Noise Adaptive*	*% of frames*
Silence	6.07	4.78	4.11	14.0%
Vowel	0.45	0.17	0.16	37.5%
Nasal	1.38	0.88	0.70	6.3%
Stop	3.90	2.06	1.60	18.3%
Fricative	9.91	3.44	2.99	14.6%
Liquids and Glides	1.45	0.66	0.57	9.3%
Voiced + Unvoiced	2.95	1.24	1.03	86.0%
Total	3.39	1.72	1.45	100.0%

TABLE IV
OBJECTIVE SPEECH QUALITY VERSUS BROAD PHONEME CLASSIFICATION. HERE, 100 TIMIT SENTENCES WERE DEGRADED WITH ADDITIVE AUTOMOBILE HIGHWAY NOISE (10 dB SNR) AND SUBSEQUENTLY ENHANCED USING AUTO-LSP AND NOISE ADAPTIVE AUTO-LSP

Objective Quality versus Broad Phoneme Classification
100 TIMIT sentences, Automobile Highway Noise, 10 dB SNR

Sound	*Itakura-Saito Likelihood Measure*			
Type	*Degraded*	*Auto-LSP*	*Noise Adaptive*	*% of frames*
Silence	24.11	14.15	14.85	14.0%
Vowel	4.34	1.96	1.27	37.5%
Nasal	3.19	1.64	1.06	6.3%
Stop	8.19	3.52	3.50	18.3%
Fricative	11.62	4.54	3.92	14.6%
Liquids and Glides	4.61	1.92	1.57	9.3%
Voiced + Unvoiced	6.37	2.71	2.22	86.0%
Total	8.85	4.28	3.93	100.0%

adjust the final terminating iteration based on local SNR estimates obtained in each time-frequency partition. In addition, the terminating iteration adjustment ensures a relaxation of the spectral smoothing constraints in regions where the noise corruption is not significant. More important, however, we note that the proposed algorithm leads to improved objective speech quality over the original Auto-LSP formulation for all noises and SNR's examined.

It is interesting to point out that the noise adaptive Auto-LSP algorithm leads to further improvements in objective speech quality for the case of white Gaussian noise. Here the mean IS measure for 10 dB was 2.67 for the original degraded test set, 1.92 for the Auto-LSP enhanced, and 1.76 for speech enhanced by the proposed algorithm. This is not surprising, since Auto-LSP applies a fixed terminating iteration to all speech frames. Hence, by adapting the terminating iteration per time-frequency subband, the algorithm is better able to adapt to the time-varying nature of the speech signal by reducing the terminating iteration in regions containing negligible noise corruption while at the same time increasing the terminating iteration in regions of significant noise corruption. We also found that both algorithms provided little or no improvement for "city rain" noise and "large crowd" noise. However, this can be attributed to both the nonstationarity of the background noise as well as the fact that a one-time estimate of the noise was used across each sentence in this set of experiments.

Tables III and IV illustrate specific improvements in objective speech quality for broad speech classifications in aircraft cockpit and automobile highway noise conditions. . In each noise condition, the proposed noise adaptive algorithm further improves objective quality over the traditional Auto-LSP formulation for each broad speech class. For example, the mean IS measure for stop consonants was reduced from 3.90 for the original degraded to 2.06 for the Auto-LSP enhanced speech. The noise adaptive algorithm further reduces this measure to 1.60. In general, the proposed algorithm provides the most improvement for speech classes such as stops and fricatives. However, for automobile highway noise, there is also a substantial improvement for vowel sections (e.g., the average IS is further reduced from 1.96 to 1.27 after processing with the proposed algorithm).

V. CONCLUSION

The original formulation of the constrained iterative Auto-LSP enhancement algorithm proposed by Hansen and Clements [10] focused on additive WGN interference. In such conditions, the application of spectral constraints to the LSP parameters and autocorrelation lags of the degraded speech was shown to provide improved speech quality and a more consistent terminating criteria. In colored noise conditions, such as aircraft cockpit and automobile highway environments, the Auto-LSP algorithm does not provide as much improvement in speech quality, since spectral constraints are applied to the entire frequency spectrum regardless of the localized nature of the noise.

In this correspondence, we have formulated a noise adaptive Auto-LSP enhancement algorithm to provide improved objective speech quality in colored noise environments. In the proposed algorithm, we considered the enhanced waveform as being composed of a sum of it's individual subband signal estimators. By adapting the terminating iteration for each time-frequency partition, the proposed algorithm was shown to provide a better compromise between signal distortion and noise attenuation. We considered ten additive noise sources ranging from highly colored (e.g., automobile highway noise) to completely flat (e.g., white Gaussian noise) and demonstrated that the proposed extension to the original constrained iterative algorithm improves objective speech quality over a wide range of SNR's.

References

[1] J. H. L. Hansen and L. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 169–184, May 1995.

[2] J. Deller, J. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, 1992.

[4] L. Arslan, A. McCree, and V. Viswanathan, "New methods for adaptive noise suppression," in *Proc. 1995 IEEE ICASSP*, pp. 812–815.

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

[6] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, 1992.

[7] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197–210, 1978.

[8] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing*, vol. 39, pp. 1943–1954, 1991.

[9] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints based on an auditory spectrum," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 22–34, Jan. 1995.

[10] J. H. L. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, Apr. 1991.

[11] J. H. L. Hansen and L. Arslan, "Markov model based phoneme class partitioning for improved constrained iterative speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 98–104, Jan. 1995.

[12] J. H. L. Hansen and S. Nandkumar, "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," *J. Acoust. Soc. Amer.*, vol. 97, pp. 3833–3849, June 1995.

[13] W. A. Harrison, J. S. Lim, and E. Singer, "A New application of adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 21–27, Feb. 1986.

[14] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

Improving Performance of Spectral Subtraction in Speech Recognition Using a Model for Additive Noise

Nestor Becerra Yoma, Fergus R. McInnes, and Mervyn A. Jack

***Abstract*—This correspondence addresses the problem of speech recognition with signals corrupted by additive noise at moderate signal-to-noise ratio (SNR). A model for additive noise is presented and used to compute the uncertainty about the hidden clean signal so as to weight the estimation provided by spectral subtraction. Weighted DTW and Viterbi (HMM) algorithms are tested, and the results show that weighting the information along the signal can substantially increase the performance of spectral subtraction, an easily implemented technique, even with a poor estimation for noise and without using any information about the speaker. It is also shown that the weighting procedure can reduce the error rate when cepstral mean normalization is also used to cancel the convolutional noise.**

***Index Terms*— Additive noise, cepstral mean normalization, convolutional noise, speech recognition, spectral subtraction, weighted matching algorithms.**

I. Introduction

In [1], a model for additive noise using infinite impulse response (IIR) filters was proposed and used to compute the uncertainty or variance related to the spectral subtraction (SS) process to weight the DP algorithms. However, most recognizers use hidden Markov model (HMM) structure, and the use of a discrete Fourier transform (DFT) filterbank is desirable because it makes the system less vulnerable to the convolutional distortion. The contributions of this paper concern:

1) a model for additive noise for the case of DFT filters;
2) a weighting procedure applicable to dynamic time warping (DTW) and HMM with SS;
3) comparison between weighted matching algorithms;
4) improvement of SS performance in terms of error rate and dependence on the threshold parameter;
5) improvement of SS combined with cepstral mean normalization (CMN) to cancel additive and convolutional noise.

The approach covered in this work has not been found in the literature and seems to be generic and interesting from the practical applications point of view.

II. Model for Additive Noise Using DFT Filters

Given that $s(i), n(i)$, and $x(i)$ are the clean speech, the noise and the resulting noisy signal, respectively, the additiveness condition in the temporal domain may be set as

$$x(i) = s(i) + n(i). \tag{1}$$

In the results presented in this correspondence, the signal was processed by 14 DFT mel filters. If $S(k), N(k)$, and $X(k)$ correspond to the fast Fourier transform (FFT) of $s(i), n(i)$, and $x(i)$ at the

Manuscript received April 2, 1997; revised December 18, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal. The work of N. B. Yoma was supported by a grant from CNP, Brasilia, Brazil.

N. B. Yoma is with DECOM/FEEC/UNICAMP, 13083-970 Campinas, SP, Brazil (e-mail: nestor@decom.fee.unicamp.br).

F. R. McInnes and M. A. Jack are with the Centre for Communication Interface Research, University of Edinburgh, Edinburgh EH1 1HN, U.K.

Publisher Item Identifier S 1063-6676(98)07788-8.