WAVELET-BASED NOISE REDUCTION

Nathaniel A. Whitmal[1], *Janet C. Rutledge*[1], *and Jonathan Cohen*[2]

[1]Department of EECS, Northwestern University, Evanston, IL 60208-3118
[2]Department of Mathematics, DePaul University, Chicago, IL, 60614

ABSTRACT

A novel method for enhancement of noisy speech is presented. Frames of speech samples are split into low and high frequency bands and projected onto a library of bases consisting of local trigonometric functions and wavelet packets. Coefficients thought to represent only the speech are selected by means of the MDL criterion, and used to synthesize an estimate of the original speech. A tracking algorithm uses MDL values to choose between the MDL processor and alternate processors which reject audible artifacts. Preliminary results indicate that the new algorithm may be useful in applications requiring a single-microphone noise reduction system for speech.

1. INTRODUCTION

Most listeners (particularly those with hearing impairments) have difficulty understanding speech in the presence of noise. Much of this difficulty may be attributed to masking of consonants, which often resemble short-duration bursts of random noise. Numerous signal processing algorithms have been proposed to address this problem. Several of these algorithms have difficulty distinguishing between noise and consonants, and consequently remove both. Furthermore, inaccurate estimates of the noise (which is often assumed to be stationary) can cause some algorithms to create audible artifacts which further mask consonants. The objective of this study was to develop a new approach capable of accurately distinguishing between speech and noise.

2. THE MDL CRITERION

The new approach uses the Minimum Description Length (or MDL) criterion recently applied by Saito [1] to reduce additive white Gaussian noise in digitized image and geophysical signals. The description length, defined as the length (in bits) of the theoretical binary codeword required to describe both a noisy signal and a model thereof, is expressed as

$$L(x, m, k, \theta_{m,k}, \sigma_N^2) = L(x|m, k, \theta_{m,k}, \sigma_N^2) + L(\theta_{m,k}, \sigma_N^2|m, k) + L(m, k).$$

This work was supported in part by the National Science Foundation under grant BCS-9110247, by the Buehler Center on Aging, and by a DePaul University Summer Research Grant.

where $\theta_{m,k}$, the model of the signal, is constructed with k members of orthonormal basis m [2]. Given a library containing M varieties of orthonormal bases (i.e., wavelet packets and local trigonometric functions) with minimum information cost [3], Saito's algorithm selects the basis and coefficients providing optimum compression of the signal and rejection of the noise (which compresses poorly in every basis). Assuming equal probability of basis selection, the approximate minimum description length (AMDL) is given by k^* coefficients in basis m^* such that

$$\tilde{L}(m^*, k^*) = \min_{\substack{0<k<\frac{N}{2} \\ 1\le m\le M}} \left(\frac{3k}{2} \log_2 N + \frac{N}{2} \log \|(I - \Omega^{(k)}) W_m^T x\|^2 \right),$$

where $W_m^T x$ is the transform matrix and $\Omega^{(k)}$ is a rank-k matrix preserving the largest k coefficients. The algorithm was successfully demonstrated on both geophysical data and digitized images.

3. ADAPTIVE MULTI-BAND MDL

3.1. Limitations of MDL for speech processing

Application of the MDL algorithm to speech enhancement confirms its capability for robust, autonomous calibration; a desirable property lacking in previous approaches. However, like earlier approaches, the algorithm tends to remove consonants in the presence of noise, and imposes mild distortion on speech (particularly consonants) in the absence of noise. Furthermore, for the short frame lengths appropriate to real-time processing of speech, the additive noise tends to compress efficiently onto a few basis elements. These retained coefficients produce audible artifacts similar to those produced by previous approaches.

3.2. Multi-band MDL

Several modifications are proposed for adaptation of the MDL algorithm to use with speech. First, the incoming signal is split into two bands: a low-frequency band dominated by vowels and nasal consonants, and a high-frequency band dominated by fricative consonants and plosive bursts. The MDL algorithm is then separately applied to the low and high frequency signals. The split-band approach allows salient features of consonants to be reproduced accurately, reducing distortion produced by the original algorithm in the absence of noise. For comparison, spectrograms of the

0-7803-2431-5/95 $4.00

noiseless sentence "That hose can wash her feet" are shown below in Figures 1, 2, and 3.

Unfortunately, the multi-band method described above shows reduced capability for noise reduction, largely because the noise is also filtered. Under these conditions, proper selection of coefficient vectors requires inverse transform and inverse filtering operations to be performed on the noise estimate for every value of m and k. Although the structure of the inverse filtering matrix may be exploited to reduce computation time, operations of order $O(N^3)$ are still required for each of $\frac{NM}{2}$ coefficient vectors. A more efficient approach uses a power-symmetric quadrature mirror filter bank, which maintains the lack of correlation in the noise required by the original MDL. This allows use of multi-band MDL without any need for inverse filtering.

3.3. Adaptive processing for consonants

Additional modifications are motivated by a relationship between changes in AMDL values and changes in the envelope of the speech waveform. Observed AMDL values have a lower bound dependent on the minimum amplitude of the speech signal, and provide reliable indication of whether the signal is above or below the noise floor. An example of MDL statistics for noisy speech is shown below in Figure 4.

Inspection of Figure 4 indicates that at (or below) the noise floor, the dominant coefficients (i.e., the ones retained) may be attributed to noise. As a result, coefficient retention is unaffected by the presence of the low-level speech, and application of MDL produces audible artifacts.

When the signal is below the noise floor, a tracking algorithm which monitors AMDL values is used to adaptively disable MDL processing at high frequencies in favor of alternative processing that reduces audible artifacts. Here, a form of power spectrum subtraction using local trigonometric bases is employed. A running average of spectra derived from discarded coefficients in the local trigonometric basis is used to construct an estimate of the noise.

4. PRELIMINARY RESULTS

A preliminary comparison of the capabilities of original and modified MDL approaches was conducted. An utterance of the sentence, "That hose can wash her feet," was sampled at 8 kHz, digitized to 16 bits, and added to each of three white Gaussian noise sequences to produce waveforms with overall SNRs of 0, 5, and 10 dB. Successive frames of the speech signals (256 samples, 50% overlap) were processed by each of three algorithms: original MDL, multi-band MDL with QMFs, and adaptive multi-band MDL with QMFs. RMS levels of the /o/ phoneme in "hose" and the closure preceding /t/ in "feet" were used to obtain relative measures of signal-to-noise ratio for each of the three methods. (For sentences with 0, 5, and 10 dB average SNRs, vowel-to-silence SNRs were 8.77, 13.78, and 18.75 dB respectively.) The observed SNR increases are presented below in Table 1.

At all noise levels, the proposed algorithm reduces the "musical noise" produced by the original MDL algorithm. This difference is reflected in the higher SNRs of the proposed algorithm. Informal listening indicates that the proposed algorithm improves perception of consonants at

Method	Vowel-to-Silence SNRs		
Original	18.8	13.8	8.8
MDL	29.8	23.9	17.7
Multi-band MDL	31.0	23.2	17.1
Adaptive MDL	32.6	25.7	20.2

Table 1. Signal-to-Noise Ratio Improvements (in dB)

vowel-to-silence ratios of 18.8 dB SNR. At lower SNRs, where the original algorithm produces substantial amounts of "musical noise," spectral subtraction tends merely to attenuate the consonants. The severity of the attenuation is likely due to the spectrum of the noise, which, being flat, is substantially higher in level than the spectrum of the consonants at high-frequencies.

5. DEVELOPMENT OF AN EXTENDED APPROACH

5.1. Reduction of colored noise

The preliminary investigations cited above point to two areas where improvement is required. The first concerns the spectrum of the interfering noise. In practice, the interference encountered in speech communication systems is generally not white or Gaussian. Unfortunately, the present formulation of MDL has not been extended to colored or non-Gaussian noise. Saito [4] reports a method of removing colored noise (with known parameters) which relies on a pattern classification method using local discrimination bases (LDBs). The method selects for every frame a set of basis functions which optimizes classification of the frame as either "noise" of known spectrum or "signal+noise."

Given a compatible statistical model of speech, it might be possible to estimate the coefficients of the noise signal through conventional parameter estimation methods. Hence, a statistical model of speech, suitable for use in parameter estimation, would first be developed. Standard estimation methods would then be evaluated for use in estimating the noise spectrum; this estimate, in turn, would be used in an LDB-based approach to enhance the speech. Should this approach prove infeasible, a second LDB-based approach using a library of noise classes (one for each type of noise) could be implemented.

5.2. Reduction of residual noise

A second area of improvement concerns the residual "musical" noise produced by retention of unwanted coefficients. This residual noise tends to increase in level as the window length decreases. To illustrate the dependence of residual noise on window length, an additive white Gaussian noise signal (variance: 1000) was processed by MDL using four different window lengths (64, 128, 256, and 512 samples). The average percentage of retained noise energy is shown in Figure 5 as a function of window length. The figure shows that the relative energy in the coefficients decreases as window length increases. This characteristic creates a challenge for real-time processing of speech, in which short window lengths are generally required.

Berger, Coifman, and Goldberg [5] have recently proposed an approach for reducing both residual noise and Gibbs-effect artifacts caused by the truncation of the

wavelet series. Their algorithm averages together denoised versions of the noisy speech, with each version delayed by a small number of samples before processing (and shifted back to its origin after processing). The method exploits the fact that the wavelet-packet and local cosine transforms are not shift-invariant, and that best-basis representations of the noise will generally be less regular than the representations of the signal.

Figures 6 and 7 compare the performance of the original MDL algorithm with that of an MDL algorithm using the shift-denoise-average method at 10 dB SNR. Here, the signal of Figure 6 is combined with a second version shifted forward by four samples, denoised, and then shifted back to its original position. The algorithm does significantly reduce the level of audible artifacts; however, it also has the effect of reducing the level of some consonants. In the example given above, the phoneme /t/ is almost completely removed by the latter processor.

A more rigorous description of residual noise production (using the theory of order statistics) may prove helpful in reducing the residual noise.

6. CONCLUSION

This study has focused on application of digital signal processing methods to the enhancement of speech in additive noise. Existing methods, which generally exploit statistical and/or spatial information for enhancing speech, often improve the SNR without improving intelligibility.

A novel method for enhancement of noisy speech has been proposed. This method exploits the fact that noise compresses less efficiently than speech onto wavelet-packet and local cosine bases. The approach separates the noisy speech into two bands, and selects the basis which describes the signal in each band with minimum information cost. The MDL criterion is used to reduce noise by selectively discarding low-level components in each minimum-entropy basis. At low SNRs, a tracking algorithm adaptively disables MDL processing in favor of alternative processing which reduces residual noise components. Preliminary results indicate that the new method may be useful in applications requiring a single-microphone noise reduction system for speech.

One speech enhancement system that has been shown to improve intelligibility under certain test conditions is the two-microphone LMS filter-based noise cancelling system evaluated by Levitt, *et al.*[6] for application in digital hearing aids. In many circumstances, however, implementing a two-microphone system is infeasible. The adaptive MDL system reviewed in this paper is intended for use in applications where two-microphone systems cannot be implemented.

It is conceivable, however, that an LMS filter-based system could be implemented with one microphone, given a suitable noise reference. The adaptive MDL system constructs such a reference estimate of the noise spectrum. Currently, the estimate is used by the spectral subtraction algorithm. It may also be possible to use this noise estimate in lieu of the reference noise signal required by the two-channel adaptive-filter based system. Future work will include implementation and evaluation of a two-microphone adaptive-filter system using the noise estimate of multi-band MDL as a reference signal.

A second motivating factor for this research was the development of a noise reduction preprocessor for recruitment of loudness compensation. Future work will include evaluation of adaptive multi-band MDL as a pre-processor for the wavelet-based compensation method of Drake, Rutledge, and Cohen [7]. This work will culminate in evaluation of the intelligibility and subjective quality of system output for normal-hearing and hearing-impaired listeners in noise.

7. ACKNOWLEDGMENT

The authors wish to thank R. Coifman for several useful conversations regarding this work.

REFERENCES

[1] N. Saito, "Simultaneous Noise Suppression and Signal Compression using a Library of Orthonormal Bases and the Minimum Description Length Criterion," in Wavelets in Geophysics, E. Foufoula- Georgiou and P. Kumar (eds.): Academic Press, Inc., 1994.

[2] J. Rissanen, *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[3] R. Coifman, and V. Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Trans. Inf. Theory,* Vol. 38, pp. 713-718, 1992.

[4] N. Saito, *Local Feature Extraction and its Application Using a Library of Bases,* unpublished Ph.D. dissertation, Yale University, 1994.

[5] J. Berger, R. Coifman, and M. Goldberg, "Removing noise from music using local trigonometric bases and wavelet packets," *Jour. Audio Eng. Soc.*, Dec., 1994, pp. 808-818.

[6] H. Levitt, M. Bakke, J. Kates, A. Neuman, T. Schwander, and M. Weiss, "Signal processing for hearing impairment," *Scand. Audiol.*, suppl. 38, pp. 7-19, 1993.

[7] L. A. Drake, J. C. Rutledge, and J. Cohen, "Wavelet analysis in recruitment of loudness compensation," *IEEE Trans. Signal Proc.*, vol. 41, no. 12, pp. 3306-3312, 1993.

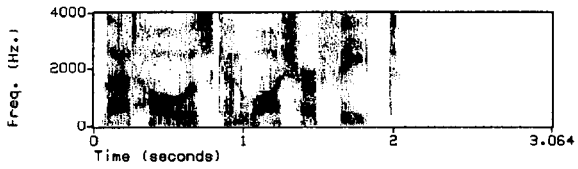

Figure 1. Input to MDL

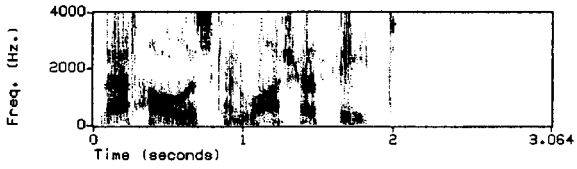

Figure 2. Output from original MDL

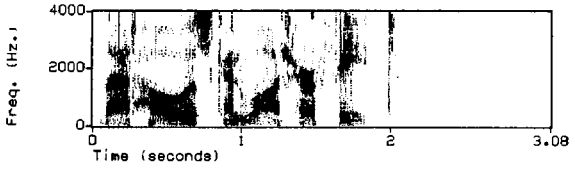

Figure 3. Output from multi-band MDL

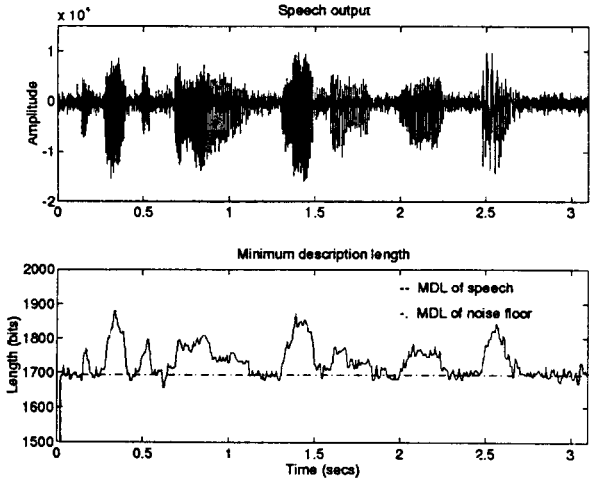

Figure 4. MDL statistics for noisy speech

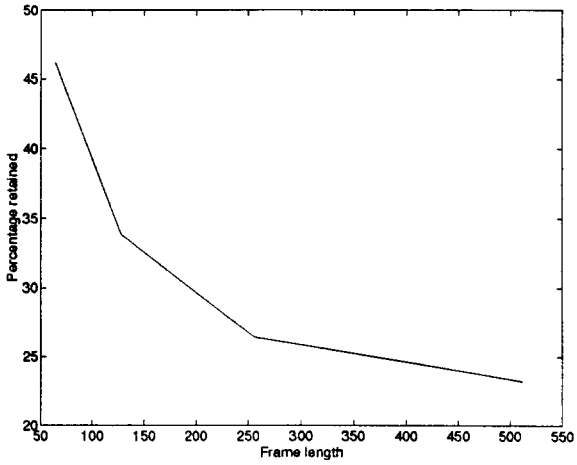

Figure 5. Noise retention vs. frame length

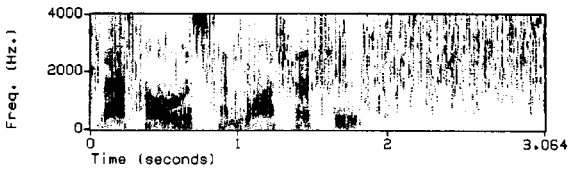

Figure 6. Output of MDL (10 dB SNR)

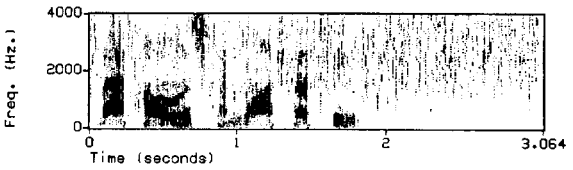

Figure 7. Output of MDL with shift-denoise-averaging (10 dB SNR)

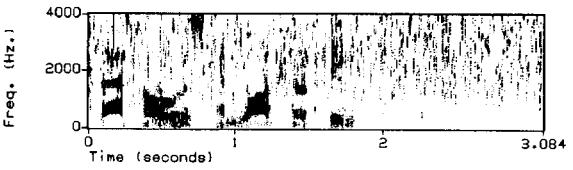

Figure 8. Output of multi-band MDL (10 dB SNR)