AUDIO SIGNAL NOISE REDUCTION USING MULTI–RESOLUTION SINUSOIDAL MODELING

David V. Anderson and Mark A. Clements

Center for Signal and Image Processing,
School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, USA
dva@ece.gatech.edu

ABSTRACT

The sinusoidal transform (ST) provides a sparse representation for speech signals by utilizing several psychoacoustic phenomena. It is well suited to applications in signal enhancement because the signal is represented in a parametric manner that is easy to manipulate. The multi–resolution sinusoidal transform (MRST) has the additional advantage that it is both particularly well suited to typical speech signals and well matched to the human auditory system [1]. The currently reported work discusses the removal of noise from a noisy signal by applying an adaptive Wiener filter to the MRST parameters and then conditioning the parameters to eliminate "musical noise." In informal tests MRST based noise reduction was found to reduce background noise significantly better than traditional Wiener filtering and to virtually eliminate the "musical noise" often associated with Wiener filtering.

1. INTRODUCTION

Speech enhancement by noise removal has received much attention lately because of its application to new, or newly popular, technologies such as hands free telephony. Traditional methods for the removal of stationary noise include spectral subtraction and Wiener filtering. The methods are popular because they are fairly straightforward to implement, simple, and effective at removing background noise. However, they tend to introduce a distortion, often called "musical noise,"[1] that is often more annoying than the noise that they remove.

Many modified forms of spectral subtraction and Wiener filtering have been suggested which address the "musical noise" problem. These include perceptual based techniques, which attempt to place noise below the threshold of audibility [2]; smoothing techniques, which represent the spectral subtraction process or Wiener filtering as a time–varying filter which is smoothed both in time and frequency [3]; and methods which attempt to find optimal spectral subtraction parameters [4, 5].

This work utilizes the multi-resolution sinusoidal transform (MRST) to obtain signal parameters which are then processed using a modified Wiener filter algorithm. The resulting parameters are then conditioned to remove any remaining "musical noise" artifacts. The MRST was chosen because it is a perceptually motivated transform and it yields a parametric representation of an audio signal that is easily modified [1].

2. MRST BACKGROUND

The sinusoidal transform, originally developed by Quatieri and McAulay, represents a signal as a sum of discrete time–varying sinusoids

$$x(t) = \sum_{k=0}^{N(t)} A_k(t) \cos(\theta_k(t)) \tag{1}$$

where $\theta_k(t) = \omega_k(t) + \phi(t)$ is a continuously varying phase [6]. In practice the parameters $A_k(t)$, $N(t)$, and $\phi_k(t)$ are estimated every 5-20 msec from the peaks in the DFT spectra of the signal. Intermediate values are obtained by interpolation.

The multi–resolution sinusoidal transform (MRST) represents a signal as a sum of discrete time–varying sinusoids of different lengths [1]. The signal is still represented as in equation 1 but the analysis method is changed. Parameters associated with high frequencies are updated frequently using short DFT windows. Parameters associated with lower frequencies are updated less frequently and are calculated using long DFT windows for more accurate frequency estimation.

[1] Musical noise is so termed because it consists of short tones at random frequencies.

The MRST directly exploits several psychoacoustic properties, that of masking and also frequency resolution. By picking only the peaks smaller signal components, that would be masked by the nearby peaks, are removed as part of the analysis process. By using a multi–resolution approach, frequencies that the ear can more closely discern are accurately determined, while the higher frequencies that are not as accurately discerned may be calculated using shorter windows and better time resolution.

The MRST analysis in this work was based on an three level octave band decomposition of the speech signal prior to the parameter estimation. The frame updates for the lowest frequency band occurred every 10 msec and used a 30 msec analysis window.

3. WIENER FILTERING

A common model for a noisy signal, $x(k)$, is a signal, $s(k)$, plus additive noise, $n(k)$, that is uncorrelated with the signal

$$x(k) = s(k) + n(k). \tag{2}$$

If the noise is also stationary then the power spectra of the signal and noise add

$$P_x(\omega) = P_s(\omega) + P_n(\omega) \tag{3}$$

Spectral subtraction attempts to recover the signal by estimating $P_n(\omega)$ and subtracting it from $P_x(\omega)$. The signal estimate, $\hat{s}(k)$ is constructed from $\hat{P}_s(\omega) = P_x(\omega) - \hat{P}_n(\omega)$ using the phase from the noisy signal. Common variations include subtracting $\gamma\hat{P}_n(\omega)$ or using the magnitude of the spectra instead of the power spectra. When $\gamma > 1$ this is called oversubtracting and it eliminates noise more effectively at the expense of some distortion in the speech.

The Wiener filter, $H_w(\omega)$, is the filter which minimizes $\sum_k |s(k) - \hat{s}(k)|^2$ for $\hat{S}(\omega) = H_w(\omega)X(\omega)$. The Wiener filter is given by

$$H_w(\omega) = \left[\frac{P_s(\omega)}{P_s(\omega) + P_n(\omega)}\right]. \tag{4}$$

This is not possible to implement in general since it is IIR, non-causal, and $P_s(\omega)$ and $P_n(\omega)$ are not usually known. However, it is possible to implement Wiener filtering on a frame by frame basis given estimates, $\hat{P}_s(\omega; m)$ and $\hat{P}_n(\omega; m)$, of the speech and noise PSDs respectively. The resulting filter is given by

$$H_w(\omega; m) = \left[\frac{\hat{P}_s(\omega; m)}{\hat{P}_s(\omega; m) + \hat{P}_n(\omega; m)}\right] \tag{5}$$

where m is the frame index.

3.1. Spectra Estimation

Perhaps the key factor in effective speech enhancement is determining when speech is present and when only noise is present. We used a simple but fairly accurate voice activity detector (VAD) that used frame energy and spectral deviance to determine voice activity. The frame energy is compared to a minimum frame energy that is leaky (i.e., the minimum frame energy increases over time to allow for changing signal conditions). The spectral deviance is found by first picking 32 points of a smoothed version of the signal spectrum, $\tilde{X}(\omega_l; m)$, $l = 0, ..., 31$. These points are compared to 32 points representing a special noise spectrum, $\tilde{N}(\omega_l; m)$ [7]. $\tilde{N}(\omega_l; m+1)$ is then updated by averaging it with $\tilde{X}(\omega_l; m)$ but only allowing a slight decrease and an even smaller increase from $\tilde{N}(\omega_l; m)$. This causes $\tilde{N}(\omega_l; m)$ to tend toward an estimate of the noise spectrum [3]. Speech is assumed present if:

1. the signal exceeded a minimum energy level by 10 dB, or
2. $\sum_{l=0}^{31} \left|\tilde{X}(\omega_l; m) - \tilde{N}(\omega_l; m)\right|$ exceeds the minimum RMS level by 8 dB.

The noise spectrum, $N(\omega; m)$ is estimated by averaging the signal spectra over time when no speech is present. The speech spectrum, $S(\omega; m)$ is estimated using spectral subtraction with an over subtraction factor, $\gamma = 2$. Alternative methods include estimating the speech spectrum by searching for harmonic sinusoids among the MRST parameters or using LPC to model the noisy signal and iteratively modeling and filtering until a relatively clean LPC representation is obtained. A constrained LPC method of estimating the speech spectrum has been used with excellent results [8].

3.2. Smoothed Wiener Filtering

A major source of distortion when performing Wiener filtering on real signals is rapid fluctuation of $H_w(\omega; m)$ between frames. This is caused by inaccuracies in the estimates of $\hat{P}_s(\omega; m)$ and $\hat{P}_n(\omega; m)$ produces small anomalies in $\hat{S}(\omega; m)$ which result in the "musical" artifacts described above. This problem can be largely eliminated by smoothing $H_w(\omega; m)$ over time when no speech is present. When speech is present $H_w(\omega; m)$ must be allowed to change rapidly or a reverberant effect is introduced as a time averaged spectra is imposed upon subsequent frames.

4. MRST ENHANCEMENT IMPLEMENTATION

The Wiener filter scheme described above was implemented within the context of the MRST with a few enhancements. First, the signal estimates were updated independently within each frequency band. Second, after filtering, any sinusoids within a band that were less than 20 dB below the largest sinusoid were eliminated. Finally, it was possible to remove residual musical noise by eliminating all sinusoids in any frequency band that contained fewer than three sinusoids.

Spectrum estimates and speech–plus–noise/noise decisions were made during analysis. However, these decisions can also be made based only on the sinusoidal parameters and methods exist for picking speech signals out of noise based only on the sinusoidal parameters [9, 10].

5. OBSERVATIONS

Several sentences were used in informal testing by several experienced listeners. For each sentence colored and white noise was added to yield segmental SNRs of 10, 5, and -5 dB. The colored noise was a recording of noisy electrical equipment including computer fans and 60 Hz hum with a high harmonic content. All signals were processed with the smoothed Wiener filter described above, a modified spectral subtraction algorithm, and the MRST based Wiener filter using four levels for the MRST.

In some respects comparison between the various methods is difficult because the types of distortion were different for each method. Moreover, it is possible to trade-off the amount and type of distorion vs. the amount of noise reduction by tuning parameters such as a limit on the maximum allowed attenuation. Thus, there are many possible operating points at which the several methods of noise reduction may be compared. This problem was addressed in a sub-optimal way by fixing the attenuation limit on the MRST based Wiener filter to 15 dB and trying several different limits on the regular smoothed Wiener filter.

In all cases the spectral subtraction produced the worst sounding signal—musical noise was the main problem but spectral subtraction also did a poor job of removing the background noise.[2]

The smoothed Wiener filter produced output that was judged perceptually superior to the noisy signal in all cases when the attenuation was limited to 10 dB or 15 dB. However, when the attenuation limit was removed, the artifacts became more noticeable and the processed speech was not consistently preferred over the noisy speech.

The MRST based Wiener filter achieved greater noise reduction in all cases, even over the smoothed Wiener filter with no attenuation limit. This was attributed to the fact that the MRST only retains the largest spectral components during analysis; therefore, much of the noise is eliminated before the Wiener filter is even applied. During the "silent" periods between phrases, there was little or no perceivable background noise remaining and almost no detectable musical noise was present. However, the output of the MRST based filter had some artifacts associated with the process of modeling noisy signals with the MRST. When the noise power is comparable to the signal power, the peak-picking can yield peaks whose frequency values have been perturbed in addition to the noisy amplitude. The Wiener filter corrects for noise in the amplitude but not bad frequency estimates. Also, the MRST based method produced some slight tonal artifacts when modeling breathy and non-voiced speech. This is due to the MRST implementation itself and not the noise reduction algorithm.

6. CONCLUSIONS

The results show that MRST provides an excellent framework for signal enhancement and specifically noise reduction. Background noise was suppressed significantly more with the MRST based Wiener filtering than with the Wiener filter applied directly to the noisy signal. It was also possible to nearly eliminate "musical noise" artifacts using the MRST because they are easily identified in the MRST parameter set. An added benefit of the noise reduction algorithm is that the processed signal is representable by fewer MRST parameters than the original and may therefore be more easily compressed.

The excellent performance of the MRST may be explained in in several ways:

1. The MRST picks perceptually significant parameters to model the audio signal; this tends to remove some noise components in a manner similar to soft threshold wavelet based denoising.

2. Higher frequency spectral updates occur more frequently giving a smooth estimate of the signal and noise spectra while still tracking the fast changes that occur in that portion of the speech spectrum. The frequency resolution that is sacrificed is not as important in the high frequencies

[2] More background noise can be removed by using oversubtraction of the noise spectrum but that results in more artifacts and some parts of the speech signal being removed.

while the time resolution gained is very important for modeling the signal.

3. The longer analysis window lengths associated with the lower frequencies enable the MRST to more accurately determine the corresponding sinusoidal parameters even when noise is present. Here the frequency resolution is important perceptually but the time resolution is not as critical because the low frequency components of the speech signal do not change as rapidly.

The ability of the MRST to model signals well, in a perceptually significant manner make it suitable to a variety of applications. Additional applications for which the MRST may be well suited include:

- signal conditioning for the hearing impaired [11, 12],
- time–scale modification of speech,
- speech coding for mid to high bit-rates,
- and automatic speech recognition and ASR signal conditioning.

7. REFERENCES

[1] David V. Anderson. Speech analysis and coding using a multi–resolution sinusoidal transform. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 1037–1040, May 1996.

[2] Nathalie Virag. Speech enhancement based on masking properties of the auditory system. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, volume 1, pages 796–799, May 1995.

[3] Levent Arslan, Alan McCree, and Vishu Viswanathan. New methods for adaptive noise suppression. In ICASSP, volume 1, pages 812–815, MAY 1995.

[4] Peter Händel. Low-distortion spectral subtraction for speech enhancement. In 4^{th} European Conference on Speech Communication and Technology, volume 2, pages 1549–1552, Madrid, Spain, September 1995.

[5] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2):113–120, April 1979.

[6] Robert J. McAulay and Thomas F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(4):744–754, August 1986.

[7] N. R. Garner, P. A. Barrett, D. M. Howard, and A. M Tyrell. Robust noise detection for speech detection and enhancement. Electronics Letters, 33(4):270–271, 13 February 1997.

[8] J. H. Hansen and M. A. Clements. Constrained iterative speech enhancement. IEEE Transactions on Acoustics, Speech, and Signal Processing, February 1991.

[9] R. J. McAulay and T. F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pages 249–252, April 1990.

[10] T. F. Quatieri and R. G. Danisewicz. An approach to co-channel talker interference suppression using a sinusoidal model for speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1):56–69, January 1990.

[11] Douglas M. Chabries, David V. Anderson, Thomas G. Stockham, Jr., and Richard W. Christiansen. Application of a human auditory model to loudness perception and hearing. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, volume 5, pages 3527–3530, May 1995.

[12] David V. Anderson, Richard W. Harris, and Douglas M. Chabries. Evaluation of a hearing compensation algorithm. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, volume 5, pages 3531–3533, May 1995.