LPC-Based, Temporal-Lateral Noise Estimation Evaluated on the AURORA Corpus

Nicholas W. D. Evans and John S. Mason
Speech and Image Research Group, Department of Electrical and Electronic Engineering
University of Wales Swansea, UK
email: {eeevansn, j.s.d.mason}@swansea.ac.uk, web: http://eegalilee.swan.ac.uk

ABSTRACT
This paper addresses the problem of noise estimation in the context of front-end speech enhancement for automatic speech recognition. A recently proposed approach uses harmonic analysis of degraded speech to detect regions in the frequency spectrum where reliable noise estimates may be sought. In this paper an analogous LPC-derived spectrum is used to locate low energy regions between resonant peaks where noise is deemed to dominate and provide more accurate estimates of noise. The relative ease with which the noise estimation process is implemented in real-time is of note. Evaluation is performed on the AURORA 2 corpus. Automatic speech recognition experiments are reported using the proposed noise estimation approach in a spectral subtraction framework. Results show an average relative performance improvement over the ETSI baseline of 26% is achieved with the proposed approach.

KEY WORDS
Noise Estimation, Speech Enhancement, Automatic Speech Recognition

1 Introduction

Reliable noise estimation remains a challenging problem in many speech enhancement and noise compensation tasks. The early spectral subtraction approach, proposed originally by Boll in 1979 [1] became the foundation for much research into both noise compensation and speech enhancement for automatic speech recognition (ASR) [2]. Spectral subtraction and its many derivatives aim to decouple the noise and speech components given the observed degraded speech, $d(t)$, assuming the noise, $n(t)$, and speech, $s(t)$, are uncorrelated and additive:

$$d(t) = s(t) + n(t) \tag{1}$$

With spectral subtraction the emphasis is placed on noise estimation and how the estimates may be best subtracted from the contaminated spectrum:

$$|\hat{S}(\omega_k, t_0)|^2 = |D(\omega_k, t_0)|^2 - |\hat{N}(\omega_k, t_0)|^2 \tag{2}$$

where $|\hat{S}(\omega_k, t_0)|^2$, $|D(\omega_k, t_0)|^2$ and $|\hat{N}(\omega_k, t_0)|^2$ are the estimated clean speech, degraded speech and estimated noise spectra for frequency ω_k at time t_0 respectively. In the early schemes the noise estimates come from noise only periods employing speech, non-speech detection to identify these intervals. It stands to reason that incorporating information on speech characteristics should lead to more accurate estimation of the underlying noise signal. Probabilistic models of speech and noise have been shown to be successful as a means of speech enhancement, for example [3, 4].

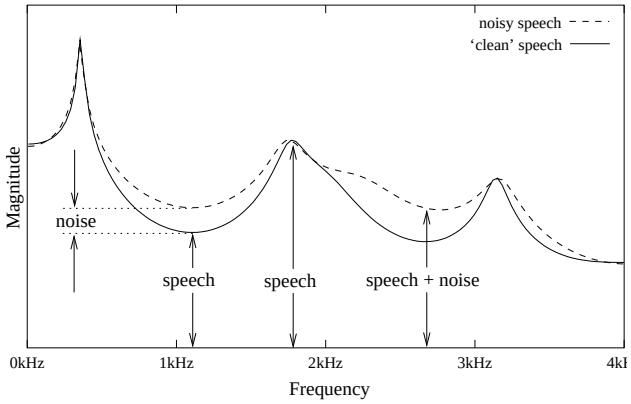

Figure 1. Two LPC spectrums with (dashed line) and without (solid line) car noise from the AURORA database added at 10dB SNR. (window period = 32ms, sampling rate = 8kHz).

However, spectral subtraction remains a widely employed and successful basis for practical systems. Given its success an obvious question is how to integrate knowledge of both speech and noise into this proven approach. In the original approach of Boll [1] the noise estimate is obtained on a frame-by-frame basis immediately prior to the speech interval. More recent adaptations of the original approach update noise statistics both in non-speech *and* speech intervals. Some examples include Stahl et al [5], Martin [6], Arslan et al [7], Doblinger [8] and Hirsch and Ehrlicher [9]. Of particular interest here are the histogram-based [9] and quantile-based [5] approaches where quasi-instantaneous noise estimates are obtained, based on the assumption that even during speech intervals, not all frequencies are permanently occupied by speech.

Very recently, Ealey et al [10] proposed a new approach of tracking non-stationary noise during speech intervals based on harmonic analysis. For any given frequency where speech is deemed to be present, instantaneous noise estimates are taken from lateral frequencies where noise is deemed to dominate. This idea is implemented by time-frequency quantile-based noise estimation (TF-QBNE) [11] and is also the basis of the idea presented in this paper. A spectrum derived through linear predictive coding (LPC) is utilised to classify regions of the spectrogram in both time and frequency, where reliable estimates of the underlying noise signal may be sought. This principle is illustrated in Figure 1 where the spectrums of a single frame are illustrated with noise (dashed line) and without noise (solid line). The utilisation of the LPC spectrum is a direct method of introducing speech information during voiced speech intervals. Resonant peaks are assumed to be associated with speech and the intermediate troughs assumed to be dominated by noise; these troughs provide an indication of frequencies that can in turn provide estimates of noise. The LPC analysis provides a smooth spectral envelope, without the pitch related harmonic structure evident in the discrete Fourier transform (DFT) spectrum. In this respect the approach is complimentary to the harmonic tunnelling approach [10] since in the latter case the very same pitch harmonics avoided here by the use of LPC are actually utilised in the goal to separate noise and speech components along the frequency axis. LPC troughs are referred to as instantaneous and lateral: *instantaneous* because they occur at the same time (within the same frame as the noise compensation process of spectral subtraction) and *lateral* because they are used not only for subtraction at their own frequency bins, but also used to estimate the noise at the adjacent resonant peaks, making the assumption that lateral noise estimates are relatively flat or even linear across the spectrum. A key feature of the approach is in its real-time implementation.

The remainder of this paper is organised as follows. In Section 2 noise estimation from the temporal characteristics of the degraded speech is described. In Section 3 noise estimation based on observations of lateral frequency noise statistics is introduced. ASR experiments are described in Section 4. Experimental results are presented in Section 5 and conclusions in Section 6.

2 Temporal Noise Estimation

Quantile-based noise estimation (QBNE) proposed by Stahl et al [5] is an extension of the histogram approach presented by Hirsch and Ehrlicher [9]. The main advantage of the quantile-based approach is that an explicit speech, non-speech detector is not required. Noise estimates are continually updated during both non-speech *and* speech periods from frequency-dependent, temporal statistics of the degraded speech signal. There are relatively few parameters to optimise and all parameters specific to the quantile are independent of absolute signal levels.

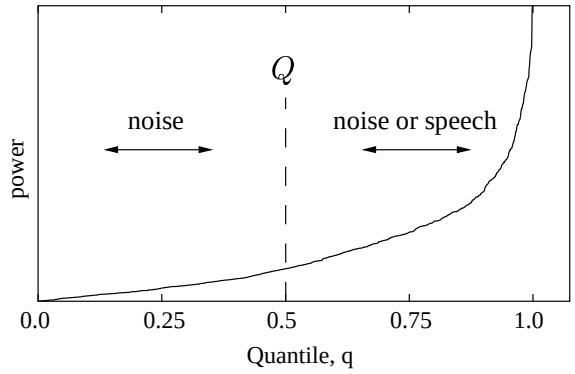

Figure 2. Quantile based-noise estimation. Values at Q are assumed to provide reliable estimates of the instantaneous noise.

For each frequency ω_k over some period, T, the power at that frequency in each frame is placed in a first-in-first-out buffer and the buffer is numerically sorted. The noise estimate is then taken as the middle or median value of the buffer. The QBNE noise estimate, $|\hat{N}_q(\omega_k, t_0)|^2$ at frequency ω_k and time t_0 is defined as:

$$|\hat{N}_q(\omega_k, t_0)|^2 = |D_{\frac{n+1}{2}}(\omega_k)|^2, \text{ assuming } n \text{ is odd} \quad (3)$$

where $|D(\omega_k)|^2$ is a numerically sorted buffer of length n containing values of $|D(\omega_k, t)|^2$ where $t_0 - \frac{T}{2} < t < t_0 + \frac{T}{2}$. Note that when speech is absent $D(\omega_k, t_0)$ becomes $N(\omega_k, t_0)$. The process is continuous and newer instantaneous values replace the oldest in the buffer. Taking the median of the distribution as the noise estimate for each frequency has proven to provide a reasonable estimate of the noise and is as good as the mean estimate used in the conventional implementation of spectral subtraction, even when the speech intervals are hand-labelled [12]. The assumption is that entries in the quantile to the right of Q may be attributed to noisy speech or high energy noise. Entries at Q are assumed to have come from speech gaps and to provide a reliable estimate of the noise. This is illustrated in Figure 2.

However, there remain significant differences between the noise estimate and the actual instantaneous values since the quantile is constructed over the short-term period, T. This is illustrated in Figure 3 where the actual noise energy (solid line) and quantile-based estimate (dashed line) are shown for the quantile at 500Hz.

3 Lateral Noise Estimation

The QBNE approach makes an estimate of the noise for each frequency, ω_k, based solely on local noise statistics at the *same* frequency. Herein lies the reasoning behind approaches such as harmonic tunnelling [10], TF-QBNE [11] and LPC-based, temporal-lateral noise estima-

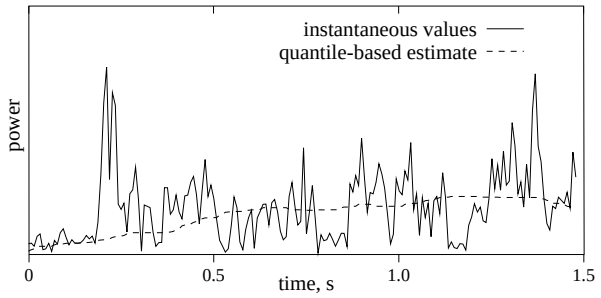

Figure 3. An illustration of the differences between the instantaneous noise values and quantile-based noise estimate at 500Hz for car noise from the AURORA 2 database (frame period = 32ms).

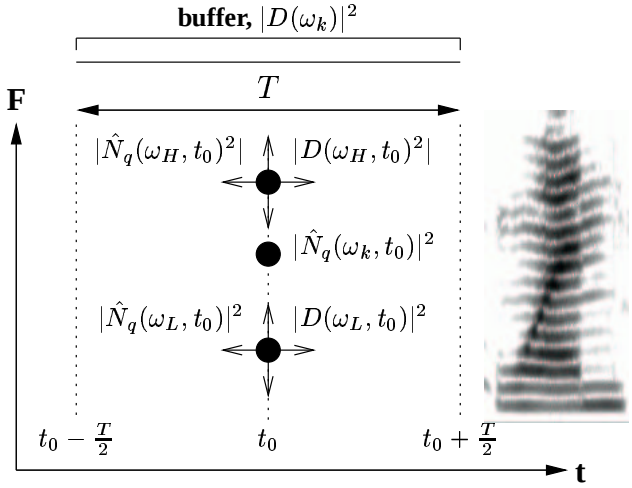

Figure 4. A conceptual diagram of LPC-based temporal-lateral noise estimation.

tion. The noise estimate may become more accurate by injecting knowledge of the degraded speech gained through LPC analysis and contributing instantaneous noise statistics from neighbouring or lateral frequencies to the noise estimation process.

In the harmonic tunnelling approach of Ealey et at [10] harmonic analysis is used to determine frequencies where reliable estimates of the underlying noise signal may be obtained. The instantaneous values are then used in the noise subtraction process for those frequencies and also contribute to estimates of the noise at other frequencies where speech energy is assumed to dominate and hence no instantaneous measures of the noise are available.

In the mixed decision-based approach of Cho et al [4] noise estimates are constructed from two separate sources. First, the estimate from outside of the speech regions termed the hard-decision based approach and second, a measure based on a probabilistic model of the degraded signal that is taken from both non-speech and speech periods, termed the soft-decision based approach. The two approaches are combined to provide the noise estimate.

In the approach presented here, LPC analysis is used to determine regions of the spectrogram that may provide instantaneous noise statistics, exactly analogous to the harmonic tunnelling approach [10]. There are then five sources of information available: the quantile-based estimate, $|\hat{N}_q(\omega_k, t_0)|^2$, the two lateral quantile-based estimates, $|\hat{N}_q(\omega_H, t_0)|^2$ and $|\hat{N}_q(\omega_L, t_0)|^2$, and the lateral instantaneous signal powers, $|D(\omega_H, t_0)|^2$ and $|D(\omega_L, t_0)|^2$. All of these values may be combined as in Equation 4 to obtain the noise estimate for frequency ω_k at time t_0:

$$\begin{aligned} |\hat{N}(\omega_k, t_0)|^2 &= \gamma_1 |\hat{N}_q(\omega_k, t_0)|^2 \\ &+ \gamma_2 |\hat{N}_q(\omega_H, t_0)|^2 \\ &+ \gamma_3 |\hat{N}_q(\omega_L, t_0)|^2 \\ &+ \gamma_4 |D(\omega_H, t_0)|^2 \\ &+ \gamma_5 |D(\omega_L, t_0)|^2 \end{aligned} \quad (4)$$

where ω_H and ω_L denote the higher and lower frequency troughs in the spectra either side of ω_k. γ denotes a simple scaling factor for each component of the noise estimate. In Equation 4 $\hat{N}_q$ indicates original QBNE or temporal estimates while D indicates lateral instantaneous values obtained from the degraded spectrum. This approach to noise estimation is illustrated in Figure 4. Note that when the period, T, over which the quantile is constructed is reduced to a single sample, $|\hat{N}_q(\omega_k, t_0)|^2$ becomes equal to $|D(\omega_k, t_0)|^2$ by Equation 3.

In [11] the quantile statistics are used to explicitly determine when and where speech is present. When the absence of speech is detected, the quantile estimate is used as the noise estimate. When the presence of speech is detected, a DFT-derived spectrum is used to determine the lateral frequencies. The noise estimate is calculated from quantile estimates at the instantaneous and two lateral frequencies. The emphasis in this paper is different. In the LPC-based temporal-lateral approach the goal is simply to take the *instantaneous* measures when and wherever possible rather than the quantile-based estimates. For frequencies where speech is deemed to be present and the instantaneous values may not be used, the instantaneous lateral noise values are combined with the temporal estimates obtained through QBNE to provide a composite noise estimate, analogous to the mixed-decision based approach in [4].

In summary, when and wherever the noise is deemed to dominate, the instantaneous values are used for noise subtraction. At resonant peaks in the spectrum, the temporal quantile-based estimate is combined with the instantaneous lateral values to form the noise estimate.

4 ASR Experiments

The evaluation of LPC-based temporal-lateral noise estimation was conducted on the AURORA 2 Distributed Speech Recognition Database [13] which is a recent standard database on which there are many published results. See for example [10, 14, 15].

Training was performed on the untreated clean speech half of the database. The training set was not modified in any way for any of the experiments performed. The multi-condition training set was not included. Testing was performed on clean speech, artificially degraded with eight different noises (subway, babble, car, exhibition hall, restaurant, street, airport and train station) added across a broad range of SNRs (clean to -5dB) with two types of convolutional distortion. Recognition experiments were conducted on the untreated utterances as a baseline and repeated after being processed with spectral subtraction where the noise estimate was obtained through the LPC-based, temporal-lateral approach described in this paper. Comparisons were also made with the original QBNE approach [5], and the TF-QBNE approach [11]. Except for the front-end speech enhancement of the test data, the training and testing procedures were not altered. Training and testing were performed with the ETSI provided scripts. The full recogniser specification is in [13].

The degraded signal was analysed on a frame-by-frame basis, where frames were 32ms in duration and the frame rate was 8ms. The DFT of each frame was computed from which the temporal quantile was constructed for all ω_k. The period, T, over which the quantile was formed was fixed at 0.5 seconds, resulting in a 63 point quantile. In an additional step, the LPC frequency response was calculated to determine the resonant peaks and the troughs between them where noise was deemed to dominate and where noise estimates were obtained. The final noise estimate was then calculated for all ω_k and subtracted as in the implementation of spectral subtraction [2] with SNR-dependent noise over-estimation and noise floors:

$$|Y(\omega_k,t)|^2 = |D(\omega_k,t)|^2 - \alpha|\hat{N}(\omega_k,t)|^2 \quad (5)$$

$$|\hat{S}(\omega_k,t)|^2 = \begin{cases} |Y(\omega_k,t)|^2, \text{ if } |Y(\omega_k,t)|^2 > \beta|D(\omega_k,t)|^2 \\ \beta|D(\omega_k,t)|^2, \text{ otherwise} \end{cases}$$

where $|D(\omega,t)|^2$, $|\hat{N}(\omega,t)|^2$, and $|\hat{S}(\omega,t)|^2$ are the power spectra of the degraded speech, noise estimate and clean speech estimate respectively.

The location of ω_H and ω_L in Equation 4 were determined from the LPC-derived spectrum. In this work only the instantaneous values at ω_H and ω_L were combined with the quantile-based estimate at ω_k to form the noise estimate used in the noise estimation process at resonant peaks in the spectrum. γ_2 and γ_3 in Equation 4 were therefore set to zero. The three remaining values were combined with equal weighting so that γ_1, γ_4 and γ_5 in Equation 4 were all set to $\frac{1}{3}$.

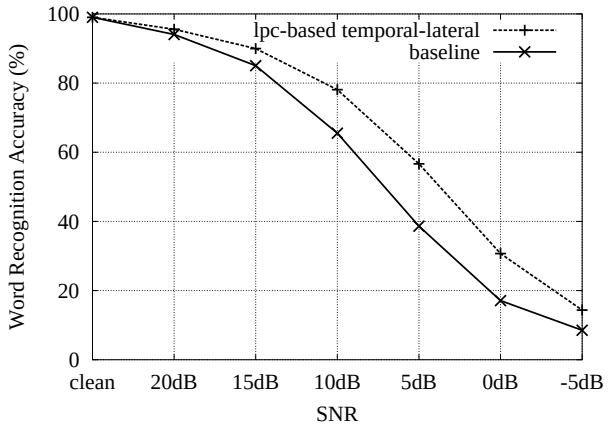

Figure 5. Average word accuracy against SNR for the ETSI front-end baseline and spectral subtraction using LPC-based temporal-lateral noise estimation for the clean training half of the AURORA corpus.

5 Experimental Results

Figure 5 illustrates the performance curves for the ETSI front-end baseline (lower solid line) and for the same data after processing with spectral subtraction using LPC-based temporal-lateral noise estimation (higher dashed line). Without added noise there is no improvement over the baseline. At all other SNRs there is an improvement in word recognition accuracy over the baseline, the best results coming from between 10dB and 0dB.

Table 1 illustrates a performance comparison in terms of the average word accuracy and average relative improvement for the baseline and for spectral subtraction where the noise estimate is obtained through the original QBNE approach [5], TF-QBNE approach [11], the proposed LPC-based temporal-lateral approach and finally for the harmonic tunneling approach as published in [10]. The LPC-based temporal-lateral based approach is shown to give superior results over the baseline, QBNE and TF-QBNE approaches but not over the root-normalised harmonic tunnelling approach. The QBNE, T-F QBNE and LPC-based temporal-lateral experiments were performed with no alteration to the feature extraction or recognition phases of the AURORA framework.

6 Conclusions

All speech enhancement techniques either explicitly or implicitly separate noise from speech. To accomplish this, knowledge of the speech must be beneficial. In this paper we use LPC analysis to identify areas of the spectrum which are likely to be dominated by speech, through regions of high energy. Low energy regions are used to provide noise estimates. This approach is combined with the recently published quantile-based approach to noise estimation where estimates are derived along the time course

Performance		
Approach	**Accuracy**	**Improvement**
Baseline	60%	-
QBNE	64%	10%
T-F QBNE	66%	15%
LPC-based temporal-lateral	70%	26%
Harmonic tunnelling [10]	83%	58%

Table 1. Performance in terms of average word accuracy and average relative improvement for the ETSI baseline, QBNE, T-F QBNE, LPC-based temporal-lateral and harmonic tunnelling.

rather than along the frequency axis. The combination of instantaneous, lateral estimation and temporal, quantile-based estimation is shown to give improved results over the single quantile-based and time-frequency quantile-based approaches. An average relative performance improvement of 26% over the ETSI baseline is achieved with the proposed approach.

References

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. on ASSP*, vol. 27(2), pp. 113–120, 1979.

[2] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars," in *Proc. Eurospeech*, 1991, vol. 1, pp. 79–82.

[3] H. Attias, L. Deng, A. Acero, and J. C. Platt, "A New Method for Speech Denoising and Robust Speech Recognition using Probabalistic Models for Clean Speech and for Noise," in *Proc. Eurospeech*, 2001, vol. 3, pp. 1903–1906.

[4] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Mixed Decision-based Noise Adaptation for Speech Enhancement," *Electronic Letters*, vol. 37, no. 8, 2001.

[5] V. Stahl, A. Fischer, and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," in *Proc. ICASSP*, 2000, vol. 3, pp. 1875–1878.

[6] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. EUSIPCO*, 1994, pp. 1182–1185.

[7] L. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptice Noise Suppression," in *Proc. ICASSP*, 1995, vol. 1, pp. 812–815.

[8] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands," in *Proc. Eurospeech*, 1995, vol. 2, pp. 1513–1516.

[9] H. G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in *Proc. ICASSP*, 1995, vol. 1, pp. 153–156.

[10] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic Tunnelling: Tracking Non-stationary Noises During Speech," in *Proc. Eurospeech*, 2001, vol. 1, pp. 437–450.

[11] N. W. D. Evans and J. S. Mason, "Time-Frequency Quantile-Based Noise Estimation," *to appear Proc. EUSIPCO*, 2002.

[12] N. W. D. Evans and J. S. Mason, "Noise Estimation Without Explicit Speech, Non-speech Detection: a Comparison of Mean, Median and Modal Based Approaches," in *Proc. Eurospeech*, 2001, vol. 2, pp. 893–896.

[13] H. G. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," *ISCA ITRW ASR2000 'Automatic Speech Recognition: Challenges for the next Millenium'*, 2000.

[14] U. Yapanel, J. H. L. Hansen, R. Sarikaya, and B. Pellom, "Robust Digit Recognition in Noise: An Evalutaion using the AURORA Corpus," in *Proc. Eurospeech*, 2001, vol. 1, pp. 209–212.

[15] J. P. Barker, M. Cooke, and P. Green, "Robust ASR Based On Clean Speech Models: An Evaluation of Missing Data Techniques For Connected Digit Recognition in Noise," in *Proc. Eurospeech*, 2001, vol. 1, pp. 213–216.