TIME-FREQUENCY QUANTILE-BASED NOISE ESTIMATION

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

ABSTRACT

This paper addresses the problem of noise estimation in the context of speech processing. A recently proposed quantile-based approach to noise estimation has the merit of not relying on the explicit detection of speech, non-speech boundaries. Here this approach is extended to both time and frequency. The resultant time-frequency quantile-based noise estimation is shown to give superior ASR performance. Results on the Aurora 2 Distributed Speech Recognition Database show an average relative performance improvement over the ETSI front-end baseline of 35%. The merits of the new system include: the relatively few parameters to optimise, the independence of absolute signal levels and minimal latency, all of which assist in real-time implementations.

1 INTRODUCTION

Ambient noise remains a challenging problem in automatic speech recognition (ASR) and with the broadening base of ASR applications the problem has grown in importance. Perhaps the best example of this is applications which involve mobile telephony since wide variations in environmental background noise conditions are often encountered. In this context the consequences of ambient noise are:

- direct contamination of the short-term spectral estimates upon which ASR systems are based
- induced changes in the speaking style of the persons subjected to the noise, known as the Lombard reflex [1]

Both of these consequences tend to have adverse effects on ASR performance. The two effects are fundamentally different and call for very different treatments.

The early work of Boll on spectral subtraction [2] is often regarded as the root of a large amount of research on speech enhancement and noise compensation. Many of these subsequent approaches are frame-based and aim to decouple the noise component from the speech component given the observed degraded speech. The original idea of Boll is based on estimates of the noise spectrum obtained in non-speech intervals. Then during speech intervals the estimate is subtracted from the degraded speech spectrum to give a new spectrum, which is then reverted to the time domain to give an enhanced speech waveform. The noise estimate is obtained from a series of short-term spectra outside of the speech interval; thus it is dynamic in that it models the immediate surrounding noise and it can be thought of as the simplest of noise models, namely a simple mean.

More sophisticated models of the noise are utilised by approaches such as parallel model combination (PMC) [3]. This more sophisticated model is of potential benefit since in practise the short-term spectra tend to exhibit significant short-term variations. This is illustrated in Figure 1 for a single frequency bin in the region of 500Hz for 32ms windows. The profiles are of the instantaneous rapidly varying spectral component (solid line) and the associated mean (dashed line). It is the instantaneous value that is sought for spectral compensation since these represent the noise at that time; the mean value is the best estimate with the simple mean approach. This raises the question of how to obtain a model of the noise spectrum that is an improvement over the simple short-term mean.

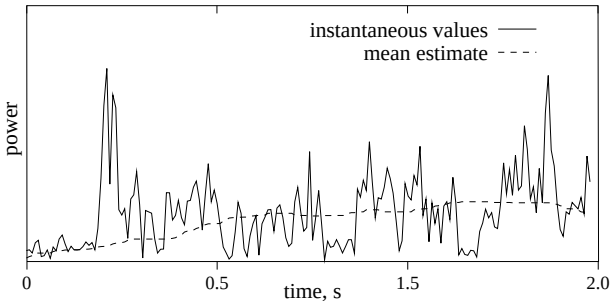

Figure 1: An illustration of the differences between the instantaneous noise values and mean noise estimate at 500Hz (window period = 32ms) for car noise from the Aurora 2 database.

Following the work of Berouti et al [4], Lockwood and Boudy address this question in the context of ASR [5] and put forward parameters extending the spectral subtraction approach, using ASR performance as the final cost function. The key parameters control noise over-estimation and noise floors. A review of related work is provided by [6]. It is interesting to note that this early work on spectral subtraction remains the basis for many approaches and provides a popular benchmark against which to judge new developments. See for example [7].

Of particular interest here is the appealingly named harmonic tunnelling approach [8]. The basis of this approach is to obtain estimates of the instantaneous noise from neighbouring non-speech regions, not simply in time (as is the case for the original Boll approach) but also in frequency. In other words when considering a given frame containing degraded speech, the idea is to consider lateral frequency bins where speech is deemed to be absent giving a time and

frequency approach to noise estimation. The harmonic tunnelling principle uses pitch harmonics to aid in the speech, non-speech decision. In this paper the idea of time and frequency estimates is investigated with an emphasis on how the values at adjacent frequencies and times are derived and combined to form the noise estimate using the quantile-based noise estimation (QBNE) approach [9].

The remainder of this paper is organised as follows. In Section 2 a brief description of the original quantile-based approach is given. In Section 3 the proposals for T-F QBNE are presented. A more detailed description at the implementation level of both original QBNE and T-F QBNE with details of the noise subtraction framework and experimental database is described in Section 4. Experimental results are reported in Section 5 with our conclusions in Section 6.

2 QUANTILE-BASED NOISE ESTIMATION

The main advantage of the quantile-based approach is that an explicit speech, non-speech detector is not required. Noise statistics are updated during non-speech *and* speech intervals. In [10] the authors show that such an approach is comparable in performance to conventional noise estimation in speech gaps even when the gaps are hand-labelled. There are relatively few parameters to implement and all parameters specific to the quantile are independent of absolute signal levels.

Approaches to noise estimation that do not require explicit speech, non-speech detection include those of Stahl et al [9], Martin [11], Arslan et al [12], Doblinger [13] and Hirsch and Ehrlicher [14]. In all cases noise statistics are continually updated during non-speech *and* speech periods. The QBNE approach, originally proposed in [9] is simple to implement, has relatively few parameters to optimise, is intrinsically independent of absolute signal levels and has minimal latency.

QBNE is an extention to the histogram approach, an idea originally put forward in [14]. The quantile-based and histogram approaches to noise estimation are based on two different statistical measures, the median and the mode. QBNE is based on the assumption that for speech periods, frequency bins tend not to be permanently occupied by speech. The non-speech, speech boundaries are implicitly detected on a per-frequency bin basis and the noise estimate is updated throughout non-speech *and* speech periods.

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. Inevitably the noise estimate is affected to some degree by the presence of speech. The QBNE noise estimate, $|\hat{N}_q(\omega_k, t_0)|^2$ at frequency ω_k and time t_0 is defined as:

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

where $|N(\omega_k)|^2$ is a numerically sorted buffer of length n containing values of $|N(\omega_k, t)|^2$ where $t - \frac{T}{2} < t < t + \frac{T}{2}$. 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 proved to provide a reasonable estimate of the noise and is as good as the mean used in the conventional approach [10]. However, there remain significant differences between the noise estimate and the actual instantaneous value. It is desirable therefore to utilise somehow the information provided by the quantile statistics to improve the noise estimate, so that they more accurately reflect the instantaneous statistics. To address this problem, a new time-frequency approach is proposed.

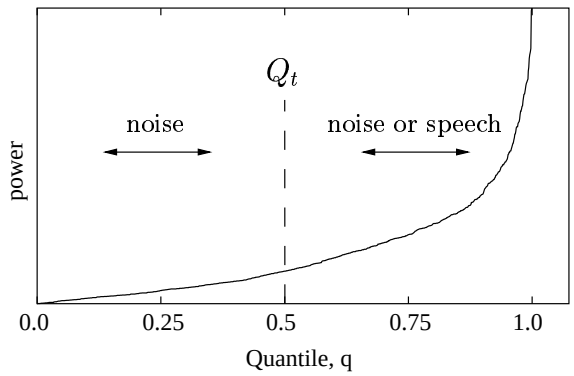

Figure 2: An illustration of Q_t. New values entering the quantile to the left of Q_t are assumed to provide reliable estimates of the instantaneous noise.

3 TIME-FREQUENCY QUANTILE-BASED NOISE ESTIMATION

An advantage of QBNE is that the quantile is automatically normalised for each frequency bin. In fact, the position in the quantile that the instantaneous noise value is placed can provide an indication of the presence of speech for each frequency. Assuming that the signal power during speech periods is higher than during speech gaps such that the new sample will enter the quantile above some threshold, Q_t as illustrated in Figure 2, it is possible to estimate the likelihood that a given bin contains a meaningful speech component or is dominated by noise. If new values enter the quantile to the left of Q_t, speech is deemed to be absent in that bin whereas if values enter the quantile to the right of Q_t, speech is deemed to be present. This the principle of the original QBNE along the time course.

In this paper noise estimation is improved by using Q_t to explicitely determine whether or not the the current sample represents speech on a per-frequency bin basis. Should the current sample be deemed not to contain speech, the noise estimate can be set to any combination of the quantile-based estimate and the instantaneous signal power. In this work, the original quantile-based estimate is used. When speech is deemed to be present, an improved estimate of the noise is sought from lateral estimates, troughs in the spectra either side of the current frequency. Note the quantile-based estimate at ω_k may be degraded by the presence of the speech at that frequency along the time course. 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, $|N(\omega_H, t_0)|^2$ and $|N(\omega_L, t_0)|^2$ may be combined as in Equation 2 to obtain the noise estimate for frequency ω_k at time t_0:

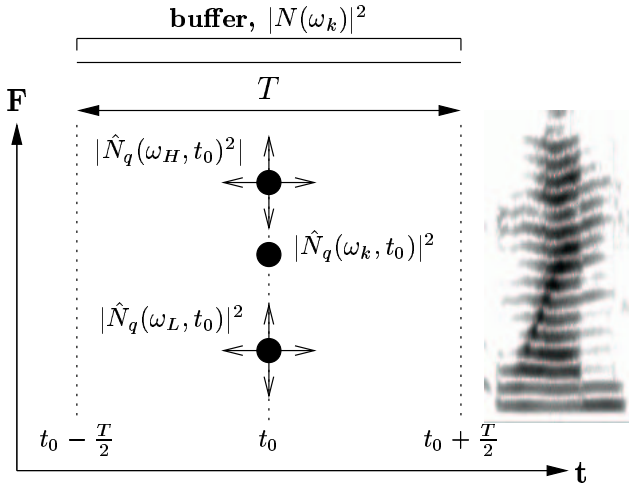

Figure 3: Time-frequency quantile-based noise estimation overview.

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

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 2 the $\hat{N}_q$ indicates original QBNE while the absence of the $\hat{}$ indicates lateral instantaneous values. 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 $|N(\omega_k, t_0)|^2$ by Equation 1. In the preliminary work presented here, both γ_4 and γ_5 are set to zero meaning that the noise estimate for each frequency ω_k is taken solely from the remaining three quantile-based estimates at frequencies ω_k, ω_H and ω_L. This is illustrated in Figure 3. The next stage is to use harmonic tunnelling or equivalent techniques to provide improved instantaneous lateral estimates.

In summary the basis of the idea presented in this paper is: take the quantile-based estimate for ω_k whenever and wherever possible in both time and frequency domains. In speech sections where the quantile-based estimate may be degraded by the presence of speech, take the noise estimate from a combination of the quantile-based estimate at ω_k and the quantile-based estimates at the low energy regions either side of the spectra, at ω_H and ω_L where noise is deemed to dominate.

4 ASR EXPERIMENTS

The evaluation of QBNE and the comparison with T-F QBNE on the Aurora 2 Distributed Speech Recognition Database [15] are reported here. The ETSI front-end uses 13 Mel frequency cepstral coefficients including the zeroth coefficient and the log energy resulting in a 14 coefficient feature vector. The full recogniser specification is in [15]. For all experiments, the models are trained on clean, unprocessed speech. Testing is on artificially degraded speech with real car noise added across a broad range of SNRs (clean to -5dB). The training set was not modified for any of the experiments performed.

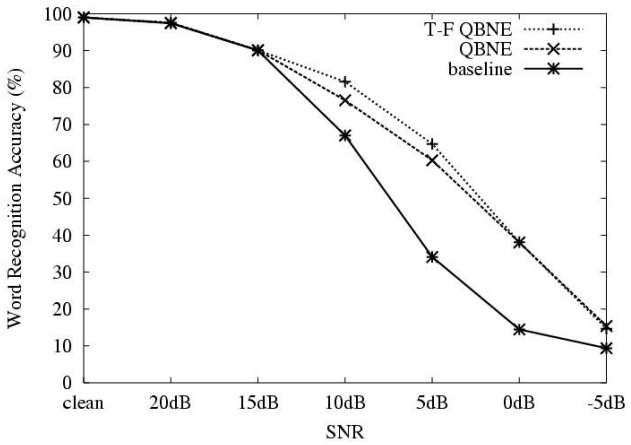

Figure 4: Word recognition accuracy against SNR for the ETSI front-end baseline, standard QBNE and T-F QBNE.

The degraded signal is analysed on a frame-by-frame basis, where frames are 32ms in duration and the frame rate is 8ms. The FFT of each frame is computed from which the quantile is constructed for all ω_k. Noise subtraction is then performed in two separate experiments using QBNE and T-F QBNE. In both sets of experiments the period, T, over which the quantile is formed was fixed at 0.5 seconds, resulting in a 63 point quantile.

The spectral subtraction algorithm is constant throughout. It is only the noise estimation algorithm that differs between the experimental sets. A standard SNR-dependent spectral subtraction framework as in [5] implemented as in Equation 3:

$$
\begin{aligned}
|Y(\omega_k, t)|^2 &= |D(\omega_k, t)|^2 - \alpha |\hat{N}(\omega_k, t)|^2 \\
|\hat{S}(\omega_k, t)|^2 &= \begin{cases} |Y(\omega_k, t)|^2, \text{ if } |Y(\omega_k, t)|^2 > \beta |D(\omega_k, t)|^2 \\ \beta |D(\omega_k, t)|^2, \text{ otherwise} \end{cases}
\end{aligned} \tag{3}
$$

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, applies.

The quantile-based estimates at ω_H and ω_L are used when the quantile-based estimate at ω_k may have been degraded by the presence of speech. The location of ω_H and ω_L are determined from a smoothed version of the instantaneous spectrum. Only values of 0.1 were considered for Q_t corresponding to 10% of the noise estimate being taken solely from the quantile-based estimate at ω_k. γ_1, γ_2 and γ_3 were all set at $\frac{1}{3}$.

5 EXPERIMENTAL RESULTS

Figure 4 illustrates the performance curves for the ETSI front-end baseline (lower solid line) and with non-linear spectral subtraction using both QBNE (middle dashed line) and T-F QBNE (higher dotted line). For the very highest SNRs there is little improvement over the baseline when using spectral subtraction with either noise estimation technique. At all SNRs below 15dB there is a noticeable improvement in

Performance		
Approach	**Accuracy**	**Improvement**
Baseline	67%	-
QBNE	72%	30%
T-F QBNE	75%	35%

Table 1: Performance in terms of average word accuracy for each approach and the average relative improvement for QBNE and T-F QBNE over the ETSI front-end baseline.

word recognition accuracy over the baseline, the best results being achieved with the time-frequency approach. For the lowest SNRs, whilst both quantile-based approaches give better results than the baseline, there is no noticeable difference between their performance.

Table 5 illustrates the improvements from each approach of noise estimation in terms of average word accuracy and relative improvement over the baseline across the range of noise levels. QBNE gives an average relative performance improvement of 30% over the baseline and T-F QBNE an average relative performance improvement of 35% over the baseline.

6 CONCLUSIONS

This paper presents an extension to the quantile-based noise estimation approach to encompass both time and frequency. For a given frequency bin, noise estimates are obtained from the said QBNE buffer and also from lateral buffers chosen because they are in local troughs and therefore with improved estimates of noise. Important advantages are at the implementation level and the inherent signal level independence. Results indicate that the new time-frequency noise estimation gives a performance advantage over the original QBNE with only a small increase in implementation cost. In harmonic tunnelling [8], the pitch of the speech is utilised to assist in determining the equivalent higher and lower lateral estimates. The next stage in this work is to incorporate harmonic tunnelling type techniques to assist in lateral frequency bin estimations.

REFERENCES

[1] Jean-Claude Junqua, "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

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

[3] M.J.F. Gales and S.J. Young, "HMM Recognition in Noise using Parallel Model Combination," in *Proc. Eurospeech*, 1993, vol. 2, pp. 837–840.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," in *Proc. ICASSP*, 1979, pp. 208–211.

[5] 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.

[6] Y. Gong, "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[7] Hagai Attias, Li Deng, Alex Acero, and John 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.

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

[9] 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.

[10] Nicholas W. D. Evans and John 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.

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

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

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

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

[15] Hans-Günter Hirsch and David 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.