EFFICIENT REAL-TIME NOISE ESTIMATION WITHOUT EXPLICIT SPEECH, NON-SPEECH DETECTION: AN ASSESSMENT ON THE AURORA CORPUS

Nicholas W. D. Evans[1], *John S. Mason*[1], *Benoît Fauve*[2]

[1]Department of Electronic and Electrical Engineering, University of Wales Swansea, UK
{eeevansn, j.s.d.mason}@swansea.ac.uk, http://eegalilee.swan.ac.uk
[2]E.N.S.P.M., F-13397 Marseille Cedex 20, France
bfauve@yahoo.fr

Abstract: This paper addresses the problem of noise estimation for speech enhancement and automatic speech recognition. In the context of mobile telephony, there is a requirement for low resource algorithms which must run at real-time. This paper describes the implementation of a recently published approach, termed quantile-based noise estimation, integrated within a conventional spectral subtraction framework. The novelty lies in the efficiency of the noise estimation process. Assessment is carried out on the AURORA corpus and demonstrates significant improvements in efficiency. Automatic speech recognition results show an average relative improvement of 26% over the baseline.

1. INTRODUCTION

Reliable noise estimation remains a particularly challenging problem in many speech enhancement and noise compensation tasks. There exists a plethora of research published in the field of noise robust automatic speech recognition (ASR). The effects of additive noise are well understood. With the recent broadening base of ASR applications, particularly in hand-held mobile telephony where wide variations in background noise are typically encountered, the problem has rapidly grown in importance. Reliable recognition performance is critical in the successful deployment of emerging speech technology applications in telephony and in other domestic situations.

In the mobile telephony area in particular, there is a requirement for low resource, efficient algorithms which typically must perform in real-time. Even where any such speech enhancement algorithm is performed remotely, say on a networked server, with the associated large volume of traffic efficiency remains highly important.

There are many published approaches to noise estimation and speech enhancement: spectral subtraction and Wiener filtering [1, 2, 3, 4, 5], parallel model combination [6, 7, 8] and a wide variety of probabilistic approaches including those of [9, 10, 11]. [3, 4, 5] all address the in-car mobile telephony scenario. Of particular interest in this paper is quantile-based noise estimation (QBNE) originally published by Stahl et al [5]. QBNE requires no voice activity detection and updates noise estimates in both speech gaps *and* speech intervals meaning it can react to changes in noise that occur during speech. Furthermore, there are relatively few parameters to optimise and parameters specific to the quantile are signal level independent. The primary theme here is that QBNE is a practical approach well suited to effective noise estimation. In [5] the noise estimate is used in the context of Wiener filtering and spectral subtraction. The approach is compared with conventional noise estimation in speech gaps and good results are reported.

In this paper, the original QBNE approach is modified to improve the reliability and efficiency of the noise estimation process. An indexing and ranking sort procedure is utilised. The proposed approach has been implemented on a 500MHz processor and is well within its capabilities. The system has been evaluated on the AURORA 2 [12] Distributed Speech Recognition Database. A comparison with the conventional QBNE approach is made in terms of efficiency and ASR word recognition accuracy. Results indicate that comparable performance is achieved through the new efficient algorithm with a great reduction in processing requirement.

The remainder of this paper is organised as follows. In Section 2 the conventional QBNE approach is described. In Section 3 the proposal for efficient QBNE is described and a comparison of the original QBNE and new proposal in terms of processing requirements is given. In Section 4 experimental work is described with the results. Conclusions are presented in Section 5.

2. QUANTILE-BASED NOISE ESTIMATION

Approaches to noise estimation that do not require explicit speech, non-speech detection include those of Stahl et al [5], Martin [13], Arslan et al [4], Doblinger [14] and Hirsch and Ehrlicher [15]. In all cases noise statistics are continually updated during non-speech *and* speech periods.

Quantile-based noise estimation (QBNE) originally proposed in [5] is an extension to the histogram approach, an idea originally put forward in [15]. The quantile-based and histogram-based approaches to noise estimation are based on two different statistical measures, the median and the mode. QBNE makes use of the fact that even during speech periods, frequency bins tend not to be *per-*

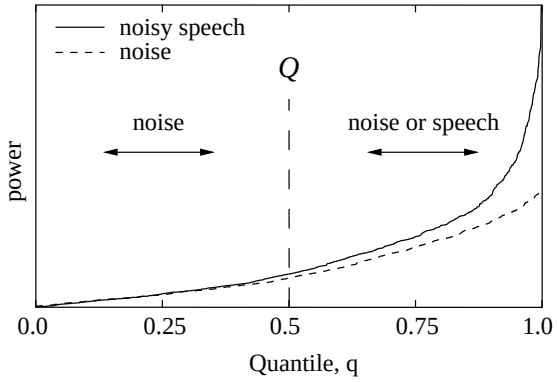

Fig. 1. Two quantile buffers for noise (dashed line) and noisy speech (solid line). Values at Q are assumed to provide reliable estimates of the instantaneous noise.

manently occupied by speech, i.e. exhibit high energy levels. This leads to non-speech, speech boundaries being detected implicitly on a per-frequency bin basis, with noise estimates updated throughout non-speech *and* speech periods. QBNE is simple to implement, with relatively few parameters to optimise and is intrinsically independent of absolute signal levels.

The degraded signal is analysed on a frame-by-frame basis. The discrete Fourier transform (DFT) is computed for each frequency ω_k over some period, T, and the power at that frequency in each frame (32ms) is placed in a first-in-first-out buffer and the buffer numerically sorted. The noise estimate for ω_k is then taken as the middle or median value of the corresponding buffer. The QBNE noise estimate, $|\hat{N}_q(\omega_k, t_0)|^2$ at frequency ω_k and time t_0 is defined as:

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

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 $D(\omega_k, t_0)$ is the same as $N(\omega_k, t_0)$. 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 proved to provide a reasonable estimate of the noise and is as good as the mean used in the conventional implementation of spectral subtraction, even when the speech intervals are hand-labelled [16].

Figure 1 shows a typical power distribution from a sorted buffer of ω_k. When speech is present, the assumption is that entries in the quantile to the right of Q may be attributed to noisy speech or high energy noise. Entries around Q are assumed to have come from speech gaps and to provide an estimate of the noise.

3. EFFICIENT QBNE

The degraded signal is sampled at 8kHz with 50% overlap, 32ms frames. The period T, over which the quantile is formed is fixed at 0.75 seconds, resulting in a 48 point quantile. This is illustrated in Figure 2.

As stated in [5] the rate that quantile-based noise estimation reacts to changes in the noise is proportional to the size of the buffer. Too small and the estimation is not accurate. Too large and the reaction time is slow. To enable the quantile to react to each replacement of old data in the quantile the quantile must be reconstructed and resorted each time the data is changed. However, the repetitive construction and sorting of each quantile for every frequency, ω_k, as the process progresses is highly inefficient. In the worst case scenario if there are n entries in k quantile buffers, there are $k(n^2)$ operations required to maintain a sorted quantile matrix with a basic sorting routine. Even with more efficient sorting algorithms such as those with $n \log_2 n$ operations, performance is still costly and there is significant redundancy in the algorithm.

In [5] the quantile is constructed and to gain efficiency when full, the smallest and largest values are discarded and replaced by the newer entries. In the worst case scenario now, this corresponds to $k(n-1)$ sorting operations to maintain the quantile matrix. However in such a scheme, should the noise change to (and remain at) a significantly higher or lower level then the noise estimation will not react to the change.

In this paper, a highly efficient indexing and ranking sort algorithm as in [17] is proposed that reacts to changes in instantaneous noise and overcomes the above limitations. As illustrated in Figure 2, an index and rank buffer, $I(\omega_k)$ and $R(\omega_k)$, to the quantile buffer, $D(\omega_k)$, are created. The purpose of the index and rank buffers are to minimise the processing required to sort each quantile buffer every time new data arrives.

Each index buffer, $I(\omega_k)$, is initialised with integers $1...n$. In the next step $D(\omega_k)$ is sorted numerically, exactly as in the conventional approach, and the same order is applied to the index. In other words, $I(\omega_k)$ is numerically sorted as if it contained the data in the quantile buffer, $D(\omega_k)$. The rank buffer, $R(\omega_k)$ then *ranks* the contents of $I(\omega_k)$. This concept is illustrated in [17], and by Figure 2.

Now when new data arrives, the processing requirement to maintain the quantile matrix is dramatically reduced. To add new data to each buffer, one need only look to the rank buffer, $R(\omega_k)$, to determine which location in $D(\omega_k)$ should be replaced with the new data. $R(\omega_k)$ effectively labels $D(\omega_k)$ by age so that the oldest value in $D(\omega_k)$ is pointed to by the first element of $R(\omega_k)$. As the processing proceeds in time, the oldest values are replaced by the newest. Now when each $D(\omega_k)$ is sorted there is only ever one value that is not in numerical order. When $D(\omega_k)$ is sorted the single out-of-order element simply cascades through the buffer to its correct location and the noise estimate is kept up-to-date upon every replacement of old data in the quantile buffer. This now corresponds to the original $k(n-1)$ operations to maintain the quantile matrix in the worst case scenario. In practise the new approach is more efficient since the new data is not forced to

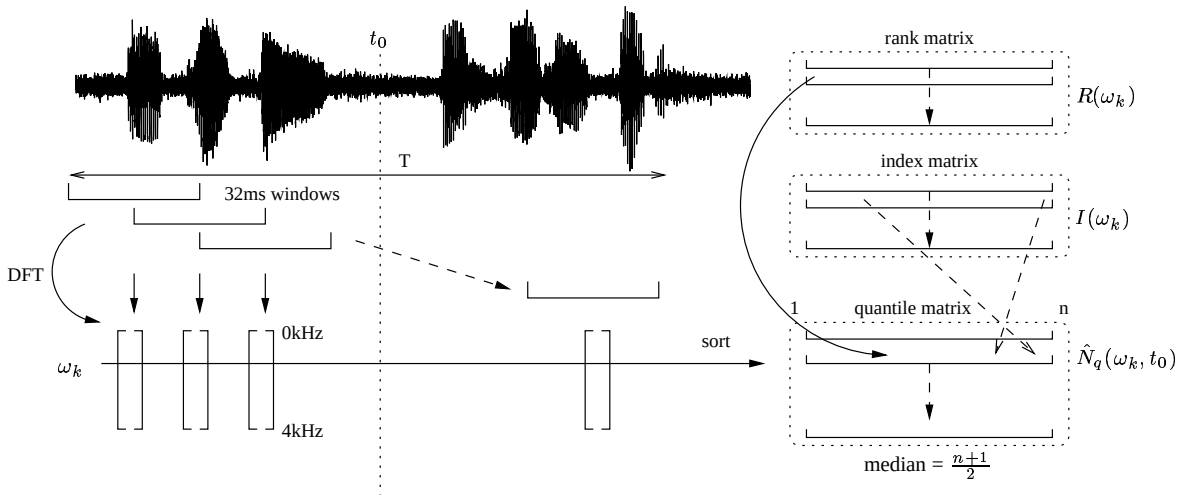

Fig. 2. Quantile-based noise estimation. A typical example of the contents of $\hat{N}_q(\omega_k, t_0)$ is illustrated in Figure 1.

enter the quantile at the extremities with far fewer sorting operations required.

4. EXPERIMENTAL WORK

The experimental work presented here is two fold. First, an evaluation of the computational efficiency and second, an assessment in terms of automatic speech recognition (ASR). Both are performed on the AURORA Distributed Speech Recognition Database [12] which is a recent standard database on which there are many published results. See for example [18, 19, 20].

To evaluate the computational efficiency of the approach, two sets of experiments were performed first, with the conventional QBNE noise estimation approach and second with the efficient QBNE approach proposed here. A subset of 1000 utterances from the AURORA database was chosen where the average utterance length was approximately 1.7s. The subset was treated with spectral subtraction with both approaches. The experiments were performed with differing buffer lengths to observe the dependence of both algorithms. The results are illustrated in Figure 3. Though the performance is dependent on the processor, the comparison between the conventional and efficient approach is valid. For the given processor, the conventional approach does not deliver real-time performance for any of the tested buffer lengths. In contrast, the efficient version performs in real-time for all of the tested quantile buffer lengths.

The second set of off-line experiments are concerned with ASR performance. An ASR system was trained on the untreated clean half of the AURORA 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 approach proposed in this paper with a buffer length of 0.75 seconds which was found to give the best performance for the tested quantile buffer lengths. 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 [12].

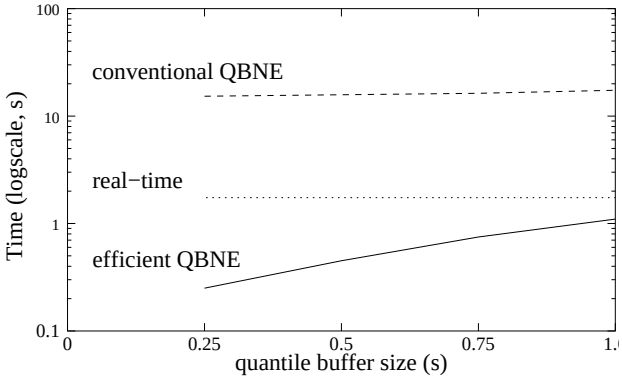

Fig. 3. Log (average) processing time against quantile buffer length in seconds. Performance for conventional QBNE and efficient QBNE.

The noise estimate was calculated for all ω_k and subtracted as in the implementation of spectral subtraction [3] with SNR-dependent noise over-estimation and noise floors:

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

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

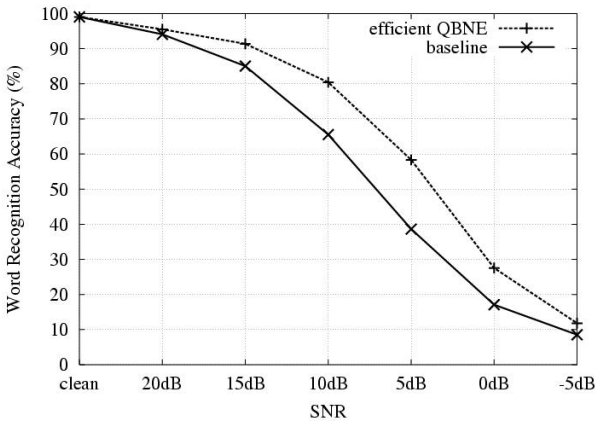

Fig. 4. Word recognition accuracy against SNR for the ETSI front-end baseline and spectral subtraction with efficient QBNE.

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.

Figure 4 illustrates the performance curves for the ETSI front-end baseline (lower solid line) and with non-linear spectral subtraction using the proposed approach (higher dashed line). For the very highest SNRs there is little improvement over the baseline. At all SNRs below 20dB there is a noticeable improvement in word recognition accuracy over the baseline, the best results being achieved between 10dB and 0dB. The average performance is 71% which corresponds to a 26% average relative improvement over the baseline.

5. CONCLUSIONS

This paper presents a highly efficient algorithm for noise estimation that does not require speech, non-speech detection. The proposed algorithm delivers a significant performance improvement in terms of processing requirement with some cost of additional memory. The algorithm performs in real-time for all conducted experiments on a personal computer. Automatic speech recognition experiments show the proposed algorithm gives an average relative performance improvement of 26% over the ETSI baseline on the AURORA 2 distributed speech recognition database.

6. REFERENCES

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

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

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

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

[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] D. Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," *Computer Speech and Language*, vol. 3, pp. 151–167, 1989.

[7] A. P. Varga and R. K. Moore, "Simultaneous Recognition of Concurrent Speech Signals using Hidden Markov Model Decomposition," in *Proc. Eurospeech*, 1991, vol. 3, pp. 1175–1178.

[8] 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.

[9] 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.

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

[11] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE Algorithm on the Aurora2 Database," in *Proc. Eurospeech*, 2001, vol. 1, pp. 217–220.

[12] 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.

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

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

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

[16] 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.

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C*, Cambridge University Press, 1999.

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

[19] 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.

[20] 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.