ABSTRACT

In the context of mobile telephony there is a need for low resource, computationally efficient noise compensation and speech enhancement approaches. This paper assesses the performance of efficient quantile-based noise estimation integrated into a non-linear spectral subtraction framework. The approach has been implemented in real-time with minimal latency on a 500Mhz processor and is well within the processing capabilities. Experiments are reported on the AURORA 2 and AURORA 3 corpora. Results show an average relative improvement of 15% on the clean and multi-condition training sets of the AURORA 2 database and an overall average relative improvement of 20% across the four AURORA 3 databases. It is acknowledged that these are not state-of-the-art results and further optimisation is anticipated.

1. INTRODUCTION

Recent years have seen an increased commercial deployment of speech recognition technology. The use of automatic speech recognition (ASR) in the mobile telephony scenario is particularly appealing. However, coupled with the convenience of mobility is the susceptibility to wide variations of degradation. The effects of background noise on ASR and the many diverse approaches to compensate for its effects have long been the interest of many speech researchers and there is a wealth of literature in the field. In this context the consequences of background noise are:

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

Both of these consequences may have an adverse effect on ASR performance, particularly when they lead to a mis-match between the training and testing stages of ASR. The Lombard reflex is often considered to be detrimental to ASR, even though the effect arises from the speakers efforts to raise the intelligibility of their speech. However, unless the ASR is trained on matched speech, the Lombard reflex is just another unwanted variation and will inevitably degrade ASR performance.

To cope with additive noise, modern ASR systems include noise compensation in one form or another. Though performance can be improved by matched training, there often remains a significant gap between performance of ASR in laboratory-clean conditions and appropriate noise compensation invariably proves beneficial. 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 noise compensation algorithm is performed remotely, say on a distributed server, efficiency remains highly important with the associated large volume of traffic. This paper assesses a recently published approach to real-time noise compensation. An improved and modified version of quantile-based noise estimation, an idea originally proposed by Stahl et al [2], is utilised in a fully automatic, frequency and SNR dependent, non-linear spectral subtraction framework [3, 4, 5] to attenuate the noise component in degraded speech. Results show an average relative improvement of 15% on the clean and multi-condition training sets of the AURORA 2 database and an average relative improvement of 20% across the four AURORA 3 databases. The algorithm has been implemented in real-time on a 500Mhz processor with minimal latency.

The remainder of this paper is organised as follows. In Section 2 a description of the modified quantile-based noise estimation algorithm used in the experimentation is given. In Section 3 the experimental setup is described. Results on both AURORA 2 and AURORA 3 databases are given in Section 4. Conclusions are presented in Section 5.

2. MODIFIED QUANTILE-BASED NOISE ESTIMATION

Quantile-based noise estimation (QBNE) proposed by Stahl et al [2] is an extension of the histogram approach of Hirsch and Ehrlicher [6]. 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 relative, that is they are independent of absolute signal levels.

For each frequency \( \omega_k \) 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, \( \hat{N}(\omega_k, t_0) \), at frequency \( \omega_k \) and time \( t_0 \) is defined as:

\[
\hat{N}(\omega_k, t_0) = \left| D_{\omega_k}(t_0) \right| \quad (1)
\]

where \( D(\omega_k) \) is the short-term power spectral estimate at \( \omega_k \), stored in a numerically sorted buffer of length \( n \) (n odd) containing values of \( D(\omega_k, t) \) where \( t_0 - \frac{2}{T} < t < t_0 + \frac{2}{T} \).
Note that when speech is absent and \( n = 1 \), \( D(\omega_k, t_0) \) becomes \( N(\omega_k, t_0) \) and ideal noise estimation for that instant is achieved. The process is continuous and new 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-labeled [7]. The assumption is that entries in the quantile greater that the median may be attributed to noisy speech or high energy noise. Entries at the median are assumed to have come from speech gaps and to provide a reliable estimate of the noise. This is illustrated in Figure 1. Figure 1(a) illustrates the time waveform of a clean utterance (top plot) from the AURORA 2 database degraded (bottom plot) by car noise with a superimposed amplitude envelope (middle plot). The profile is chosen to illustrate the behavior of QBNE. Figure 1(b) illustrates QBNE values at 1000Hz over the period of the utterance where the noise estimate is taken from \( Q=0.5 \) as illustrated. Finally, Figure 1(c) illustrates the quantile-based noise estimate (solid line) and the true noise energy (dashed line) at 1000Hz. The graph illustrates how QBNE reacts to changes in the noise during both non-speech and speech periods. It is the rapidly varying, true instantaneous values for which estimates are sought.

In this paper, QBNE is integrated into a fully automatic, frequency and SNR dependent non-linear spectral subtraction framework. The degraded signal is sampled at 8kHz with 32ms frames, 50\% overlapped. The period \( T \), over which the quantile is formed is fixed at 1 second, resulting in a 64 point quantile.

As stated in [2] 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, \( \omega_k \), as the process progresses is highly inefficient.

In [2] to gain efficiency when full, the smallest and largest values in the quantile buffer are discarded and replaced by the newer entries. 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 the experimental work presented here, a highly efficient indexing and ranking sort algorithm was implemented to perform efficient QBNE in real-time. The index and ranking sort algorithm minimises the time required to sort the quantile buffers and reacts to changes in noise conditions each time new data arrives. A full description of the index and ranking sort algorithm is presented in [8].

3. EXPERIMENTAL SETUP

An evaluation of QBNE is reported here on the AURORA 2 and AURORA 3 corpora [9]. For all reported experiments the standard ETSI front-end was used. There are 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 [9].

Efficient QBNE was integrated into a frequency and SNR dependent, non-linear spectral subtraction (NLSS) algorithm. The two major parameters of NLSS are the noise over-estimate (\( \alpha \)) and noise floor (\( \beta \)). In [5, 10] the benefits of an SNR dependent noise over-estimation factor are discussed. The idea is to subtract

![Fig. 1](image-url) An illustration of quantile-based noise estimation. (a) an utterance from the AURORA 2 database is degraded by car noise with an envelope amplitude superimposed. (b) the quantile values at 1000Hz over the period of the utterance. (c) the true instantaneous noise magnitude and noise estimate taken from \( Q=0.5 \) on the quantile, at 1000Hz.
Fig. 2. Absolute word accuracy scores for the AURORA 2 corpus after QBNE and NLSS speech enhancement where the algorithm operates in real-time and is fully automatic with respect to noise level.

Fig. 3. Word accuracy scores for the AURORA 3 corpus after QBNE and NLSS speech enhancement. Parameters of QBNE and NLSS taken from AURORA 2 results and not optimised. The same QBNE and NLSS parameters are used for training and test sets.

a greater over-estimate of the noise in more noisy periods and a smaller estimate in less noisy periods. In this work, the NLSS algorithm is fully automatic. The SNR of the degraded signal is approximated using the ratio of the degraded signal and quantile-based noise estimate energies. The noise estimate is then scaled as in [11] and subtracted as in Equation 2:

\[ S(\omega_k, t)^2 = \begin{cases} Y(\omega_k, t)^2 - \alpha N(\omega_k, t)^2, & |Y(\omega_k, t)|^2 > \beta |D(\omega_k, t)|^2 \\ \beta |D(\omega_k, t)|^2, & \text{otherwise} \end{cases} \]

where \( |D(\omega, t)|^2, |N(\omega, t)|^2, \) and \( |S(\omega, t)|^2 \) are the power spectra of the degraded speech, noise estimate and clean speech estimate respectively. An SNR dependent noise floor, \( \beta \), was observed to be beneficial. A greater noise floor was used for lower noise conditions than for higher noise conditions.

For the clean training sets of the AURORA 2 database, speech compensation through QBNE and NLSS was performed on the test sets only. Training was performed on the original speech data. For the multi-condition training half of the AURORA 2 database, identical speech enhancement was performed on both the training and testing sets as was the case for AURORA 3 experiments. In all cases the algorithm was fully automatic and no additional measure of SNR was performed. No end-pointing was performed so any observed performance improvement can be solely attributed to the automatic QBNE, NLSS combination.

4. EXPERIMENTAL RESULTS

Figure 2 illustrates the performance of the proposed approach on both the multi-condition and clean training sets. Quantile-based noise estimation and non-linear spectral subtraction gives a performance improvement across the range of noise levels for the clean training set where the goal is to minimise the mismatch between laboratory-clean conditions of the training data and the noisy conditions of the test data.

The goal in the multi-condition training set is subtly different. Whereas both training and testing sets contain added background noise, as might be expected performance degrades as the noise level increases. However, the performance is superior to that of the clean training set. The aim of QBNE and NLSS now becomes to enhance the speech signal and minimise the effect of the inherent variation between the training and testing noise conditions. QBNE and NLSS do not perform as well at this task as they do for the clean training task and smaller improvements are obtained on the multi-condition training set. However, significant performance improvements are obtained on the noisiest sets whose noise levels are not included in the training sets.

Figure 3 illustrates preliminary results of experiments on the four AURORA 3 databases. The results presented here are not optimised and are greatly dependent on the exact noise levels in each set.

Figures 4 and 5 illustrate the performance of the QBNE and NLSS combination in terms of absolute word error rates and average relative performance improvement. QBNE and NLSS are...
more suited to the clean training task and better results were observed on the clean training half of the AURORA 2 database and on the medium and high mis-match sets of the AURORA 3 databases.

5. CONCLUSIONS

This paper has presented a computationally efficient approach to noise estimation that does not require explicit speech, non-speech detection or explicit SNR estimation. These two tasks become absorbed into the QBNE and NLSS functions. Experimental results are reported which show an average relative improvement of 15% on the clean and multi-condition training sets of the AURORA 2 database and an overall average realtive improvement of 20% across the four AURORA 3 databases. Whilst the performance is not state-of-the-art, the algorithm operates in real-time and has been implemented on a 500Mhz processor, well within its capabilities illustrating the potential for low resource implementation. The system is not fully optimised on either database and experimental work is ongoing.

6. REFERENCES