Speech Enhancement for Forensic and Telecommunication Applications

A. J. Fisher and S. Sridharan

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering,
Queensland University of Technology

ABSTRACT - This paper describes speech enhancement applied to covert recordings to improve both quality and intelligibility of noise corrupted speech. In this situation, intelligibility is the key issue since the recording is likely to used as audio evidence. It is shown that using improved noise estimation and post-processing applied to a noise subtraction technique, both the above features can be substantially improved. It is also shown that these results are also significant to applications in the telecommunications industry where enhancement of speech is used as a pre-processing stage in speech recognition, coding and speaker verification.

INTRODUCTION

Law enforcement agencies often engage in surveillance operations which involve the recording of spoken conversations. As is often the case, these recordings are made with a single microphone under covert conditions and, as a result, are corrupted by various forms of additive noise. This degrades both the quality and intelligibility of the speech. Forensic analysis of the recording may involve either simple transcription, speaker identification or speaker verification. All of which would benefit from some form of enhancement. In addition, the recording itself may be submitted as an item of evidence in a court of law hence good speech intelligibility holds equal, if not greater, importance than good quality.

Of the many forms of noise which may corrupt a given recording, broadband noise is certainly the most common and possibly the most difficult to completely remove. Its initial estimation thus plays a significant role in any noise suppression process. One such process is the Spectral Subtraction technique proposed by Boll (1979). Its principle of implementation has formed the basis of many broadband noise suppression methods (Berouti, *et al*, 1979), (Preuss, 1979), (McAuley & Malpass, 1980), (Victorin, 1985), (Ahmed, 1989), (Kang & Fransen, 1989), (Lockwood & Boudy, 1992) which have achieved a relative degree of success. As its name suggests, spectral subtraction consists of subtracting an estimate of the power spectrum of the noise from that of the noisy speech.

Spectral subtraction, in both its original and modified forms, is effective in achieving good quality (Signal to Noise Ratio - SNR) improvement, however, it is only capable of improving the intelligibility to a moderate extent. This is largely due to the subtraction of a uniform noise estimate from the whole speech signal. As the unvoiced segments of speech are usually of a low amplitude, the subtraction tends to suppress these along with the noise, resulting in reduction in intelligibility despite improvement in quality.

In addition, using only an estimation of the noise in the subtraction never results in its complete elimination. Instead, a new residual noise resulting from the errors of subtraction in the frequency domain is formed. Essentially comprising of low amplitude frequency tones, this new noise has an audible musical quality, and so, is often referred to as *musical noise*. This also contributes to a reduction in intelligibility of the speech. In an effort to combat this, attempts to remove the musical noise through spectral whitening via noise flooring, have only served to improve the quality at the cost of reducing intelligibility even further (Berouti, 1979), (Lockwood & Boudy, 1992).

From this, it can be seen that improvement of the situation may not necessarily lie in further refinement of the suppression algorithm, but rather in improving the initial noise estimation. The effectiveness of spectral subtraction's implementation is largely constrained by its noise estimate and hence it is only logical that an improvement in the noise estimate will improve the performance of the algorithm.

To date, methods for noise estimation requiring no *a priori* knowledge of the signal have been developed and operate on an energy thresholding basis. Kang and Fransen (1989) implemented a binary decision based on a nominated threshold. If the energy of the first 1 kHz was found to exceed the given threshold, then the frame was classified as voiced otherwise it was unvoiced. Although it is clear that unvoiced speech information will corrupt the noise estimate, it has been argued that the total duration of unvoiced speech to silence is such that the corruption is not greatly significant. This may well be the case where the main concern is quality improvement, but when intelligibility is the issue, the significance becomes more apparent. This will be shown later.

A new noise estimation method proposed by Krubsack & Niederjohn (1994), based on voicing confidence rather than thresholding is implemented here to provide an estimate which can be applied to facilitate both quality and intelligibility improvement requirements. In addition, a post-processing subtraction technique is applied in the cepstral domain to remove residual noise without compromising intelligibility.

NOISE ESTIMATION

The method first begins with determining the pitch and voicing confidence of given speech segments (Krubsack & Niederjohn, 1991) sampled at 20 kHz. It is first Low pass filtered at 600 Hz and then decimated to 10 kHz. Following this is band pass filtering between 20 - 600 Hz and then segmenting at 51.2 ms with an overlap of 75%. The pitch is determined using an autocorrelation method correcting for doubling and tripling errors. The voicing confidence is then determined for each segment based on three criteria: 1) The RMS energy of the segment (e'), 2) the maximum value of the autocorrelation function over the pitch range normalized by the value at zero lag (p'), and 3) the RMS of the normalized autocorrelation function over the pitch range (r'). Based on these, a piece-wise linear confidence space is determined and each segment is attributed a confidence (c_v) given its position within this space. The confidence ranges between -1 to 1 with 0 being the boundary between voiced and unvoiced.

Given the pitch and voicing confidence for each segment, noise estimation can be determined for both voiced and unvoiced segments. However, given that covert recordings are often bandlimited owing to transmission over a telephone line (O'Shaughnessy, *et al*, 1990), only, the unvoiced segments were used to generate the noise estimate (Krubsack & Niederjohn, 1994).

For unvoiced segments ($c_v < 0$) a minimum three term window (Nuttall, 1981) is applied and the modified periodogram calculated. The modified periodograms $P(k)$ of the present and all past unvoiced segments are averaged using the Welch method (Welch, 1967) applied to Logarithmic modified periodograms (Krubsack & Niederjohn, 1994)

$$N_u(k)_{dB} = \left(\frac{1}{Mu} \sum_i P_i(k)_{dB} \right) + \zeta \qquad (1)$$

where $N_u(k)$dB is the power spectrum noise estimate in dB, Mu is the number of unvoiced segments and ζ is the log-modified Euler's constant (ζ = 2.506815781... dB). Finally, the noise estimate, $\tilde{N}_u(k)$ is obtained by smoothing $N_u(k)$ (Krubsack & Niederjohn, 1994)

$$\tilde{N}_u(k)\ dB = F[\ wc\ F^{-1}\ [\ N_u(k)\ dB\ ]\ ] \qquad (2)$$

where F and F^{-1} are the standard DFT and IDFT respectively and *wc* a triangular window applied in the cepstral domain for smoothing to reduce the effects of any unvoiced speech in the noise estimate (Krubsack & Niederjohn, 1994).

SPECTRAL SUBTRACTION

This simply involves the subtraction of the noise power spectral estimate segment from each corresponding noisy speech power segment. For each voiced segment, the closest preceding unvoiced noise estimate is used. All negative values are zeroed.

$$P_i(k) = Y_i(k) - N_{ui}(k) \qquad \text{for all k}$$
$$P_i(k) = 0 \qquad \text{for } P_i(k) < 0 \qquad (3)$$

where $P_i(k)$ is the power spectrum of the enhanced speech, $Y_i(k)$ is the power spectrum of the original noisy speech and $Nu_i(k)$ is power spectrum of the estimated noise.

CEPSTRAL SUBTRACTION

In the root cepstral domain (Alexandre & Lockwood, 1993), broadband noise is more compressed about the origin, and as a result, greater separation between speech and noise occurs. Hence, this property can be exploited by applying subtraction in this domain (Wu, *et al*, 1991). As Figure 1 illustrates, first an estimate of the noise (residual noise in this case) is taken and an initial estimate of the clean speech is made by subtracting in the spectral domain. Then in the root cepstral domain, subtraction of the noise estimate again occurs based upon a scaling factor determined from the relative SNR provided by the spectral speech estimate. As will be seen later, cepstral subtraction is superior to spectral subtraction at moderate to high SNRs in terms of maintaining intelligibility levels of the speech. However, at lower SNRs, its performance approaches that of spectral subtraction. Now, in observing that spectral subtraction is much better than cepstral subtraction in relation to quality improvement at low SNRs, it is apparent that these two methods can be combined to provide optimum conditions for intelligibility enhancement. Implementation would consist of applying spectral subtraction with the above mentioned noise estimation technique to provide quality improvement while preserving intelligibility and then using cepstral subtraction, in the better SNR conditions, to remove residual noise and enhance intelligibility.

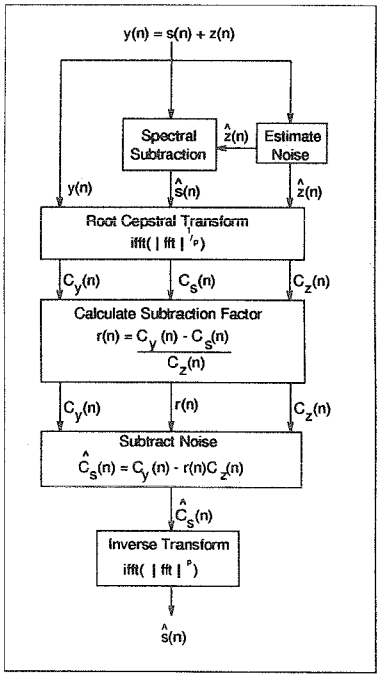

Figure 1: Cepstral subtraction

RESULTS

A given utterance, "*Every salt breeze comes from the sea*", was corrupted at signal to noise ratios from +15dB down to 0 dB. Processing was performed using spectral subtraction alone with the noise estimation method proposed by Kang & Fransen (K1) method and also with the noise estimation method proposed by Krubsack and Niederjohn (K2). In addition cepstral subtraction was performed alone and finally spectral subtraction (K2) was performed combined with cepstral subtraction. Figure 4 shows the time domain representation for each processing method at a SNR of 5dB.

As can be seen in figure 4(b), at the 1.5 second mark the /z/ ("*breeze*") has been incorrectly removed by the spectral subtraction(K1) method. Also the /k/ and the /z/ ("*comes*") at the 1.75 and 1.9 second marks as well as the /s/ ("*sea*") at the 2.45 second mark have also been incorrectly suppressed. However, in the other methods using K2, these unvoiced speech segments have been preserved.

Note also that the spectral subtraction(K1) method provides better SNR improvement over cepstral subtraction. The combination of both methods using K2 serves to even better that.

Figures 2 and 3 show quality and intelligibility improvement respectively, for the SNRs outlined above. Quality assessment is based on average segmental SNR over the whole utterance. Intelligibility assessment is based on the objective measurement of the 20-band Articulation Index (Kryter, 1967) for the given utterance.

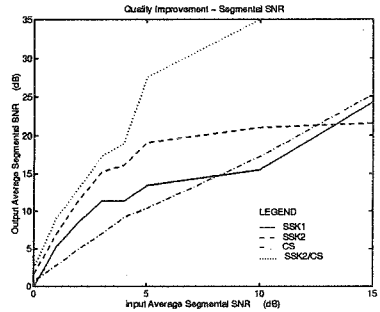

Figure 2: Quality Enhancement

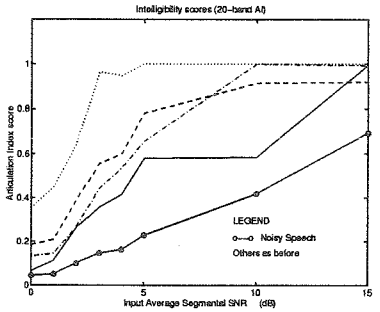

Figure 3: Intelligibility Enhancement

CONCLUSIONS

Spectral subtraction is an efficient technique for substantially improving the quality and moderately improving intelligibility of broadband noise corrupted speech. Using previous noise estimation techniques, however, its performance is constrained by the level of corruption of the noise estimate by unvoiced speech segments. On the other hand, cepstral subtraction is able to provide significant increase in intelligibility at moderate to high SNRs, however, its performance degrades at lower SNRs.

It has been proposed to implement spectral subtraction, with a better noise estimate, to substantially increase the SNR of the noisy speech while maintaining satisfactory intelligibility levels, at lower SNRs so that cepstral subtraction can be more effectively applied to improve intelligibility. The results demonstrate that the proposed method is indeed superior to either of the other implementations tested for both quality and intelligibility enhancement.

Hence it can be said that this form of enhancement satisfies the requirements for speech enhancement applied to covert recordings in both quality and intelligibility improvement. In addition the substantial improvement in quality facilitates its candidacy for application as a pre-processor for telecommunication applications such as speech recognition, speech coding and speaker verification.

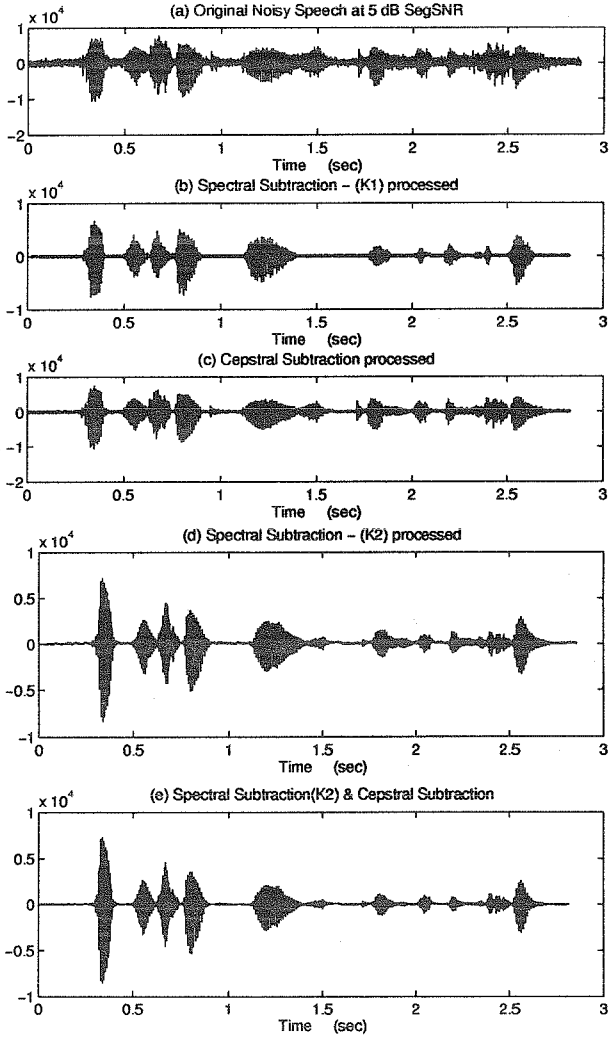

Figure 4: Time domain noisy and processed signals

ACKNOWLEDGMENT

This work was supported in part by grants from the National Institute of Forensic Science, the Queensland Police Services and a QUT Research Encouragement Award

REFERENCES

Ahmed, M.S. (1989) *Comparison of Noisy Speech Enhancement Algorithms in Terms of LPC Perturbation*, IEEE Trans. Acoust. Speech and signal Processing, ASSP-37, 121-125.

Alexandre, P. & Lockwood, P. (1993) *Root Cepstral Analysis: A unified view. Application to speech processing in car noise environments*, Speech Communication, v12, 277-288.

Berouti, M., Schwartz, B. & Makhoul, J. (1979) *Enhancement of speech corrupted by acoustic noise*, Proc. ICASSP-79, 208 - 211.

Boll, S.F. (1979) *Suppression of acoustic noise in speech using spectral subtraction*, IEEE Trans. Acoust. Speech and signal Processing, ASSP-27, 113 - 120.

Kang, G.S. & Fransen, L.J. (1989) *Quality improvement of LPC-processed noisy speech by spectral subtraction*, IEEE Trans. Acoust. Speech and signal Processing, ASSP-37, 939 - 942.

Krubsack, D.A. & Niederjohn, R.J. (1991) *An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech*, IEEE Trans. Acoust. Speech and signal Processing, ASSP-39, 319 - 329.

Krubsack, D.A. & Niederjohn, R.J. (1994) *Estimation of noise corrupting speech using extracted speech parameters and averaging of logarithmic modified periodograms*, Digital Signal Processing: A Review Journal, v4,154 -172.

Kryter, K.D. (1967) *Methods for the calculation and use of the Articulation Index*, J. Acoust. Soc. Am., v24, 175 - 184.

Lockwood, P. & Boudy, J. (1992) *Experiments with a Nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars*, Speech Communication, v11, 215 - 228

McAulay, R.J. & Malpass, M.L. (1980) *Speech enhancement using a soft-decision noise suppression filter*, IEEE Trans. Acoust. Speech and signal Processing, ASSP-28, 137 - 145.

Nuttall, A.H. (1981) *Some windows with very good sidelobe behaviour*, IEEE Trans. Acoust. Speech and signal Processing, ASSP-29, 1352 - 1354.

O'Shaughnessy, D. *et al.* (1990) *Applying speech enhancement to audio surveillance*, J. Forensic Sc., v35, 1163 - 1172.

Preuss, R.D. (1979) *A frequency domain noise canceling preprocessor for narrowband speech communications systems*, ICASSP-79, 212 - 215

Victorin, J. (1985) *Enhancement of noisy speech enhancement using spectral subtraction*, Proc. European Conf. Circuit Theory & Design, 569 - 572.

Welch, P.D. (1967) *The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms*, IEEE Trans. Audio and Electroacoust. AU-15, 70 - 73.

Wu, C.S., et al (1991) *Fast self-adapting broadband noise removal in the cepstral domain*, ICASSP-91, 957 - 960.