Noise Compensation using Spectrogram Morphological Filtering

Nicholas W. D. Evans, John S. Mason and Matt J. Roach
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 describes the application of morphological filtering to speech spectrograms for noise robust automatic speech recognition. Speech regions of the spectrogram are identified based on the proximity of high energy regions to neighbouring high energy regions in the three-dimensional space. The process of erosion can remove noise while dilation can then restore any erroneously removed speech regions. The combination of the two techniques results in a non-linear, time-frequency filter. Automatic speech recognition experiments are performed on the AURORA database and results show an average relative improvement of 10% is delivered with the morphological filtering approach. When combined with quantile-based noise estimation and non-linear spectral subtraction, the average relative performance improvement is also 10% but with a different performance profile in terms of SNR.

KEY WORDS

Noise Compensation, Morphological Image Filtering, Automatic Speech Recognition

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 background noise. 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 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 approaches to compensation. This paper addresses the first issue; here we are concerned with noise estimation in the context of noise compensation and speech enhancement for ASR.

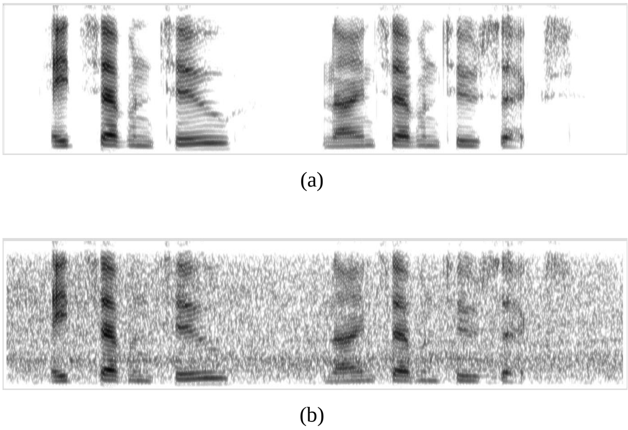


Figure 1. Two spectrograms from (a) a ‘clean’ utterance and (b) with added car noise from the AURORA database at 10dB SNR (vertical frequency axis 0-4kHz, horizontal time axis 0-3s).

Approaches to noise compensation and speech enhancement fall into two very broad categories:

- front-end processing of test data (time waveform, spectral processing, robust feature extraction)
- back-end compensation of pre-trained models

In front-end processing for example, a spectral estimate of the noise signal can be obtained and then *subtracted* from the degraded signal. In back-end processing a similar noise estimate may be used to compensate speech models for the presence of ambient noise at the test source. Whatever the approach taken in either category, the goal is to separate the noise and speech signals, whether it be implicitly or explicitly, and then to minimise the inevitable difference between the conditions at the training and testing stages. This is illustrated in Figure 1 by the spectrograms for ‘clean’ and noisy speech.

A particular difficulty of estimating background noise is caused by frame-to-frame variations of the energy in each frequency bin. This is illustrated in Figure 2 by the energy in a single frequency bin in the region of 500Hz for 32ms DFT windows for a period of 2 seconds taken from car noise. The profiles are of the instantaneous rapidly varying spectral component (solid line) and the associated

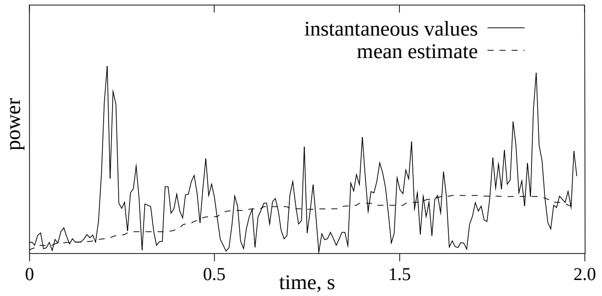


Figure 2. An illustration of the differences between the instantaneous noise values and mean noise estimate at 500Hz (window period = 32ms) for car noise.

mean noise estimate (dashed line). When only noise is present, the instantaneous values may be used as the (perfect) noise estimate but when speech is present the task becomes much more difficult and often the mean computed over a suitable interval is taken as an estimate. This raises the question of how to obtain a model of the noise spectrum that is an improvement over the simple short-term mean.

Reliable noise estimation remains a challenging problem in many noise compensation and speech enhancement tasks. Spectral subtraction [2, 3, 4] is a simple example of spectral noise estimation and subtraction, where noise estimates are updated during speech gaps. Here the noise and speech intervals are explicitly separated. In the histogram approach of [5] and quantile-based approach of [6, 7] the noise and speech signals are implicitly separated. Noise estimation is performed in both non-speech *and* speech intervals. In [8] instantaneous noise estimation is performed based on the harmonic analysis of degraded speech. In [9] an approach to noise estimation is proposed that combines implicit and explicit estimates. In [10] and [11] two different extensions to the quantile-based approach [6] and harmonic tunnelling approach [8], are proposed where the noise estimate is improved based on the profile of a spectrum derived from the discrete Fourier transform (DFT) and linear predictive coding (LPC). In [12] time-frequency filtering is employed to attenuate low residual noise (termed musical noise in the literature) after spectral subtraction. Non-speech-like peaks in the processed spectrum which remain after spectral subtraction are identified based of their duration, bandwidth and proximity to other peaks in the three-dimensional, time-frequency approach. Peaks deemed to be residual noise are removed whereas peaks deemed to come from speech are left untreated.

The common element to these latter approaches is that noise estimation is a three-dimensional operation based of the spectrogram of the degraded signal. It is reasonable to hypothesise that the more successful techniques not only implement the idea of noise estimation but also speech estimation. In this paper we introduce the idea of speech region identification using techniques usually applied to image processing. In essence the goal is to identify areas of the spectrogram where the presence of speech is dominated by the presence of noise. The combined morphological operations of first erosion and then dilation known as opening are applied to a binary thresholded version of the degraded spectrogram to highlight probable speech regions. The idea is then that noise statistics in the remaining area of the spectrogram may be used to calculate an estimate of the underlying noise signal in both non-speech *and* speech regions.

The remainder of this paper is organised as follows. In Section 2 the morphological filtering operations of erosion, dilation and opening are described. In Section 3 we discuss experimental work with results; conclusions are presented in Section 4.

2 Morphological Filtering and Noise Compensation

Upon visual inspection of typical spectrograms at a reasonable SNR such as those in Figure 1, the speech regions and structures are intuitively identified. Morphological filtering is a technique applied to image processing and is used for a number of purposes. These encompass structure enhancement, object marking, shape simplification, fore-ground/back-ground segmentation and noise filtering. A familiar example of morphological processing is optical character recognition. A scanned binary image is preprocessed to remove noise and the objects (the letters) are marked and enhanced as a precursor to recognition. Morphological filtering is proposed here as a non-linear tool for the pre-processing of speech spectrograms for noise estimation and noise compensation.

Figure 3(a) illustrates a time waveform of an utterance from the AURORA database at 10dB SNR. Figure 3(b) illustrates the corresponding spectrogram. To the human eye the speech structures are clearly visible. Noise estimation and speech enhancement are typically employed to attenuate the noise component to improve ASR performance. With approaches such as spectral subtraction, the underlying noise signal is estimated in some fashion, conventionally during non-speech intervals in the frequency domain, and then subtracted from the degraded spectrum. The assumption is made that the information is contained in the short-term spectral *magnitude* and that the *phase* is of negligible importance.

Successful noise estimation approaches consider not only noise statistics from the same frequency along the time course, but also those statistics gained from neighbouring frequencies along the time course. To accomplish this, the detection of speech becomes a three-dimensional problem. Attributes that may be used to identify speech regions include:

- pitch harmonic analysis [8]
- relative energy levels
- peak proximity-based analysis

An understanding of morphological filtering may be gained from [13, 14, 15]. Morphological analysis may be applied to grey-scale images but in the preliminary work presented in this paper, erosion and dilation with isotropic structuring elements are combined in *opening* pre-emphasised and thresholded, binary spectrograms as illustrated in Figure 3.

2.1 Erosion

Erosion is used to shrink objects. A structuring element or mask, B, is moved over the image, X, pixel-by-pixel. Using the notation from [14] the result of eroding image, X, by the structuring element, B, written:

$$X \ominus B = \{p \in \varepsilon^2 : p + b \in X \text{ for every } b \in B\} \qquad (1)$$

is defined by all those points p for which all possible $p + b$ are in X where ε^2 represents the two-dimensional space. The result is a smaller or *eroded* version of the original image.

2.2 Dilation

Dilation is the dual of erosion. Dilation results in the expansion of a shape and is used to fill small gaps or valleys between shapes. Again, using the notation from [14] the result of dilating image X by structuring element, B written:

$$X \oplus B = \{p \in \varepsilon^2 : p = x + b, x \in X \text{ and } b \in B\} \qquad (2)$$

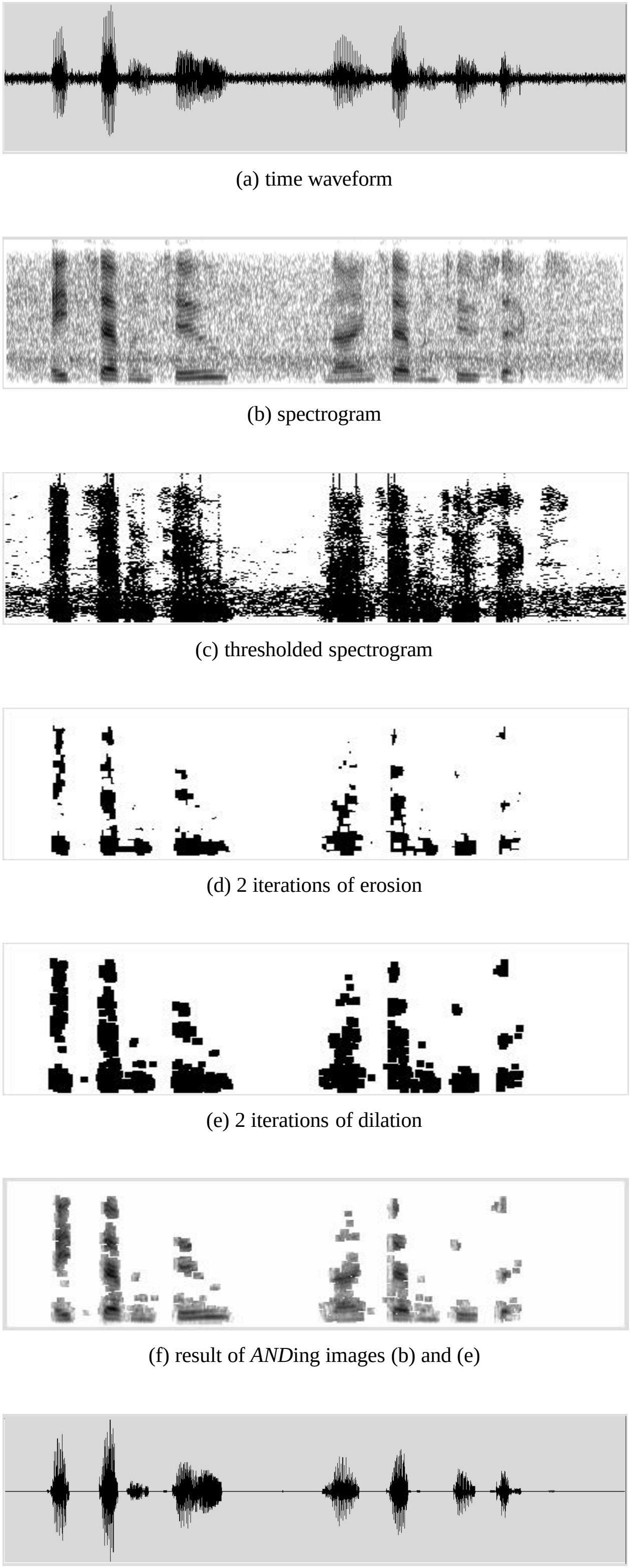
is defined as the point set of all possible vector additions of pairs of elements, one from each of the sets X and B.

2.3 Opening

Each process of erosion and dilation may be applied repeatedly. However, neither process is generally reversible. If an image is eroded, dilation will not restore the image to its original since once an object or element has disappeared, it can not generally be reformed. This property is utilised in the process of opening written as:

$$X \circ B = (A \ominus B) \oplus B \qquad (3)$$

This implies that the opening of image, X, is the erosion of X by B followed by the dilation of the results, also by B. Usually erosion and dilation operations may be performed n times. Opening is commonly used to remove noise. Unwanted small regions of noise in the image are eroded until they disappear. The image is then dilated to restore a smoothed version of the original image, without the noise.



(a) time waveform

(b) spectrogram

(c) thresholded spectrogram

(d) 2 iterations of erosion

(e) 2 iterations of dilation

(f) result of *AND*ing images (b) and (e)

(g) resulting time waveform

Figure 3. An overview of morphological filtering (spectrogram vertical frequency axis 0-4kHz, horizontal time axis 0-3s).

An example of opening is given in Figure 3. Figure 3(d) shows the results after two iterations ($n = 2$) of erosion with the thresholded spectrogram in Figure 3(c) and a simple 3*x*3 isotropic structuring element. Figure 3(e) illustrates the results of dilating Figure 3(d) with the same structuring element and so represents the opening of Figure 3(c). Figure 3(f) provides an example of how the detected speech regions are extracted from the original spectrogram using the binary map and restored to the time domain to perform speech enhancement as in Figure 3(g).

In this paper we examine the process of opening applied to speech spectrograms and draw the analogy between techniques such as harmonic tunnelling [8] and time-frequency quantile-based noise estimation [10].

3 Experimental Work

The evaluation of morphological image filtering techniques applied to speech spectrograms for robust speech recognition in noise was conducted on the AURORA 2 Distributed Speech Recognition Database [16] which is a recent standard database on which there are many published results. See for example [8, 17, 18].

An ASR system was trained on the untreated clean speech half of the AURORA 2 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, 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 morphological filtering and quantile-based, non-linear spectral subtraction similar to [6]. Except for the front-end speech enhancement of the test data, the training and testing procedure was not altered. Training and testing were performed with the ETSI provided scripts. 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 [16].

The degraded signal was analysed on a frame-by-frame basis, where frames were 32ms in duration and the frame rate was 16ms. Morphological filtering was applied to the spectrograms and in the first set of experiments, where speech was detected to be absent (white regions in spectrogram images) by the morphological filtering, the energy values where set to some small sensible noise floor. The processed spectrogram was then reverted to the time domain and the first set of recognition experiments performed. In the second set of experiments the non-speech regions were used to obtain noise estimates at all frequencies and combined with quantile-based noise estimation to obtain a composite noise estimate used in non-linear spectral subtraction. The period, T, over which the quantile was formed was fixed at 0.5 seconds, resulting in a 32 point quantile. The composite noise estimate was calculated for all ω_k and subtracted as in the implementation of spectral subtraction in [4] with SNR-dependent noise over-estimation and noise floors:

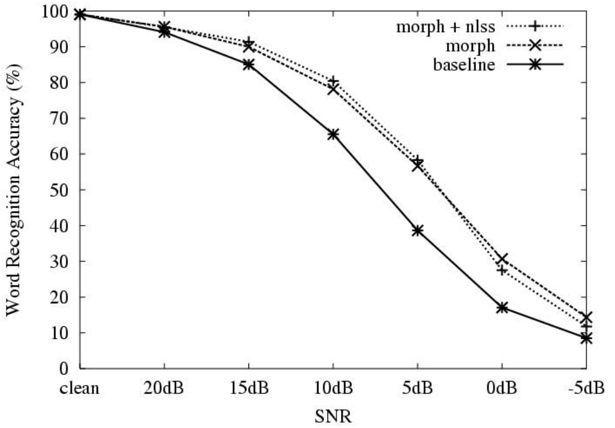


Figure 4. Word recognition accuracy against SNR for the ETSI front-end baseline (lower solid line), after binary morphological filtering (dashed line) and the combination of binary morphological filtering with quantile-based noise estimation and non-linear spectral subtraction (dotted line).

$$|Y(\omega_k,t)|^2 = |D(\omega_k,t)|^2 - \alpha|\hat{N}(\omega_k,t)|^2 \tag{4}$$

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

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.

Figure 4 illustrates the performance curves for the ETSI front-end baseline (lower solid line), after morphological filtering (dashed line) and after quantile-based, non-linear spectral subtraction using the composite noise estimate (dotted line). For the very highest SNRs there is little improvement over the baseline. At all SNRs below 15dB there is a noticeable improvement in word recognition accuracy over the baseline, the best results being achieved between 10dB and 0dB. The average performance for the baseline is 58% and after morphological filtering alone is 66% which corresponds to a 10% average relative improvement over the baseline. When the morphological filtering is combined with quantile-based noise estimation and non-linear spectral subtraction, the average performance is also 66% with the same average relative improvement. It is observed that whilst there is no difference in average performance between the two approaches, they exhibit different performance profiles in terms of SNR as illustrated in Figure 4.

4 Conclusions

This paper has presented the idea of morphological filtering for noise compensation and speech enhancement for noise robust automatic speech recognition. The morphological operations of erosion are shown to remove noise from the noisy spectrogram and dilation is shown to restore any erroneously removed speech regions. The combination of erosion and dilation techniques were applied to attenuate the noise component in noisy speech spectrograms and an average relative improvement of 10% over the ETSI-baseline was acheived with the two approaches.

This work has demonstrated the suitability of morphological filtering techniques applied to noise compensation and speech enhancement. The next stage is to implement grey-level morphological filtering techniques where the grey level represents the signal energy. This approach is expected to deliver improved results over the binary thresholded version.

References

[1] J. C. 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*, vol. 27(2), pp. 113–120, 1979.

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

[4] 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.

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

[6] 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.

[7] 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.

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

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

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

[11] N. W. D. Evans and J. S. Mason, "LPC-Based, Temporal-Lateral Noise Estimation Evaluated on the AURORA Corpus," *to appear Proc. IASTED SPPRA*, 2002.

[12] G. Whipple, "Low Residual Noise Speech Enhancement Utilising Time-Frequency Filtering," in *Proc. ICASSP*, 1994, vol. 1, pp. 5–8.

[13] J. R. Parker, *Algorithms for Image Processing and Computer Vision*, Wiley & Sons, 1997.

[14] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, PWS Publishing, 1999.

[15] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley Publishing Company, 1993.

[16] 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.

[17] 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.

[18] 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.