Morphological Filtering of Speech Spectrograms in the Context of Additive Noise

Francisco Romero Rodriguez[1], *Wei M. Liu*[2], *Nicholas W. D. Evans*[2], *John S. D. Mason*[2]

[1]Escuela Superior de Ingenieros, Seville, Spain
franciscororo@hotmail.com
[2]School of Engineering, University of Wales Swansea, UK
w_ming99@yahoo.com, {n.w.d.evans, j.s.d.mason}@swansea.ac.uk

Abstract

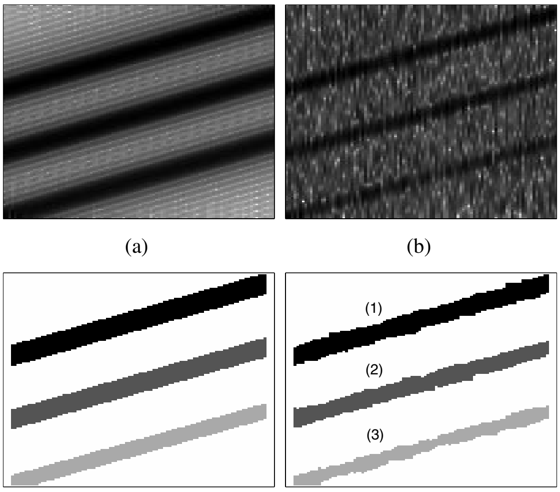
A recent approach to signal segmentation in additive noise [1, 2] uses features of small spectrogram sub-units accrued over the full spectrogram. The original work considered chirp signals in additive white Gaussian noise. This paper extends this work first by considering similar signals at different signal-to-noise ratios and then in the context of speech recognition. For the chirp case, a cost function based on spectrogram area is introduced and this indicates that the segmentation process is robust down to and below 0 dB SNR. For the speech experiments the objectives are again to assess the segmentation capabilities of the process. White Gaussian noise is added to clean speech and the segmentation process applied. The cost function now is automatic speech recognition (ASR) accuracy. After segmentation speech areas are set to one constant level and non-speech areas are set to a lower constant level, thereby assessing the segmentation process and the importance of spectral shape in ASR. For the ASR experiments the TIDigits database is used in a standard AURORA 2 configuration, under mis-matched test and training conditions. With 5 dB SNR for the test set only (clean training) a word accuracy of 56% is achieved. This compares with 16% when the same noisy test data is applied directly to the ASR system without segmentation. Thus the segmentation approach shows that spectral shapes alone (without normal spectral amplitude variations) leads to perhaps surprisingly good ASR results in noisy conditions. The next stage is to include amplitude information along with appropriate noise compensation.

1. Introduction

The task of separating speech from noise has proven to be a particularly challenging one over a number of years. Whether the speech is destined for an automatic recognition system or for a person the normal goal of the task is essentially the same and can be summarised as that of extracting a representation of the speech signal that leads to improved recognition. In this context the early work of Boll [3] is generally acknowledged as the forerunner of experimental investigations of many variants under the general heading of spectral subtraction. All involve deriving noise estimates which are then subtracted from the corrupted signal. Typically, these estimates relate to short-term discrete Fourier transform frequency bins and in the early procedures they were derived from non-speech intervals. However, more recently procedures have been examined which derive noise estimates continuously, during speech and non-speech intervals. These include the quantile-based approach of Stahl *et al* [4] and extensions which utilise both local time and local frequency bins [5]. Another recent approach which uses local time and frequency bins to derive noise estimates is the harmonic tunnelling of [6]. Clearly these latter approaches have the potential for deriving better noise estimates since they make use not only of the complete time course but also can provide estimates from within the same instantaneous window that is subjected to compensation, a feature particularly beneficial for distinctly non-stationary noise.

This paper considers another approach which possesses these benefits in that it attempts to separate a signal from noise. However here the process is one of segmentation, classifying regions of a short term spectrogram as either signal or non-signal (noise). The process is based on the statistical properties of the short term spectrogram and morphological filtering. Regions in the spectrogram are identified as either noise or signal and then the signal regions are grown via morphological processing. Very little work has been published on the morphological filtering of speech. The work of Hansen [7] considers morphologically based feature enhancement in the context of noisy speech and Lombard, but little else has followed. The motivation for the work presented here stems from the recent work of Hory *et al* in signal segmentation [1, 2].

The remainder of the paper is structured as follows. Section 2 contains a description of the morphological filtering approach as proposed by Hory *et al* [1, 2] including supporting evidence independently obtained for chirp-based signals extracted from additive white Gaussian (AWG) noise. The basic work is first extended by considering segmentation performance for different signal-to-



(a) (b)

(c) (d)

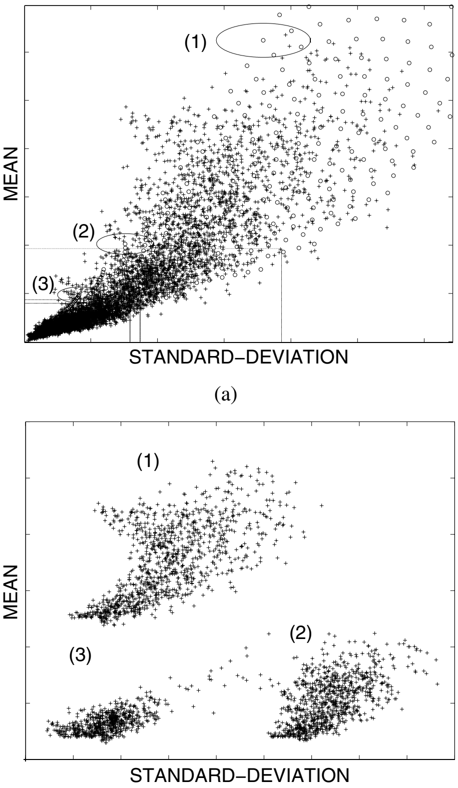
Figure 1: Spectrograms of chirp test signal: (a) without and (b) with AWG noise added at +5 dB (higher chirp), 0 dB (middle chirp), -5 dB (lower chirp), and corresponding segmentation results (c) and (d). Horizontal time axis (0-0.5 seconds), vertical frequency axis (0-4 kHz).

noise ratios, again for the same synthesised chirp signal (Section 3). Then, in Section 4, some initial speech recognition results are presented using the TIDigits database [8], followed by some observations and suggestions for further work.

2. Background to Signal Segmentation

The procedures considered here to segment signals from noise follow closely those of Hory *et al* [1, 2]. Consequently, since the details and background are covered extensively in [1, 2], only the concepts and outlines of the procedures are presented. In conceptual terms it is convenient to consider the spectrogram as an image. Then the segmentation process is based on the statistics of features derived from sub-images and accrued over the full image. The assumption is that the statistics associated with noise differ from those associated with the signal. The sub-image features recommended in [1, 2] are the mean and standard deviation of pixel values (ie power spectra) of each sub-image. Figure 1 (a) and (b) show spectrograms of clean and noisy chirp test signals respectively (as used in [1, 2]). The corresponding spectrograms after segmentation are shown in Figure 1 (c) and (d).

The segmentation procedure begins by computing the local features across the spectrogram and then locating seeds for morphological growth of signal regions. A grid is obtained from estimations of the SNR of the sub-images. The grid is superimposed onto the feature space and determines the seed selection. Examples of seeds are illustrated in Figure 2 (a), which shows a plot of the two features, mean against standard deviation for the 3-chirp examples shown in Figure 1 (b) and (d). Seeds are taken, one by one, starting with the highest values of the grid (region 1 in Figure 2 (a)) until reaching a region deemed to be noise. Each seed is grown in the spectrogram to provide signal regions. Subsequently a new grid and a new estimated noise-region are computed from the un-segmented elements. The process is repeated until the normalised maximum likelihood calculated to estimate the noise converges [1, 2]. The segmented feature space is given in Figure 2 (b).



(a)

(b)

Figure 2: An illustration of seeds (a) for the test signals in Figure 1 (b) and (d), and the resulting segmented feature space (b) showing three distinct regions, one for each chirp.

3. Signal to Noise Ratio

The results presented in Figures 1 and 2 corroborate those reported in [1, 2]. In this section the procedures are applied for different levels of signal-to-noise ratio in an attempt to assess the robustness of the process. An objective cost function is defined in the form of segmentation areas. The procedure is applied first to a clean signal leading to the baseline segmentation area. Then the cost function comes from the integration of the image difference signal:

$$\text{Seg. Acc.}|_{\text{SNR}=x\text{dB}} = \frac{\sum_{n,k} \text{Seg}(n,k)|_{x\text{dB}}}{\sum_{n,k} \text{Seg}(n,k)|_{\text{clean}}},$$

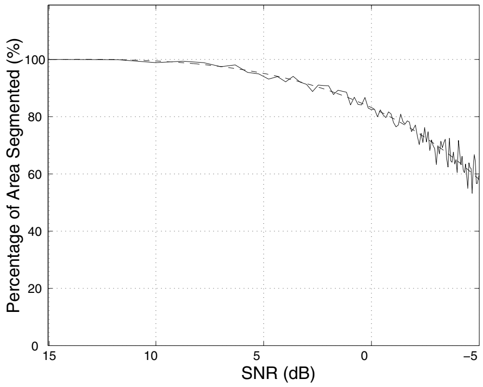


Figure 3: Dependence on SNR: area based segmentation error against SNR for chirp signal.

where the segmentation accuracy at x dB is given by the ratio of segmented areas at x dB and the original clean area, n and k are the dimensions of the spectrogram image. Figure 3 shows values of this cost function for chirp signals with different SNR, decreasing from +15 dB to -5 dB. The associated segmentation error shows a marked increase when the SNR reaches -5 dB, but is reassuringly level for higher SNRs. Example time waveforms, spectrograms and results of segmentation are given in Figure 4. At -5 dB (Figures (b), (d) and (f)) some of the signal area has been lost and the chirp is broken at one point early along the time course; also, there are some small areas wrongly classed as signal mainly at the lower end of the frequency range. The corresponding segmentation cost function in Figure 3 indicates a 40% signal area error for the same conditions.

4. Speech Experiments

The objective of the experiments reported here are twofold. The first is to assess the contribution of speech shapes or structures in the spectrogram to ASR performance and second, to extend the assessment of SNR dependence reported in Section 3, to the speech context.

The experiments are performed on a subset of the TIDigits database [8]. A set of 8440 utterances are used for training and a set of 1001 utterances for testing. The experimental setup is the same as that of the AURORA 2 database [9] except that here AWG noise is added to the test set only at six different noise levels (20, 15, 10, 5, 0, and -5 dB).

In order to investigate the contribution of speech shapes or structures, the segmentation is applied to speech spectrograms as described in Section 2. Following segmentation, the energy at all areas of the spectrogram deemed to be speech are set to one common ceiling value. Similarly, all areas of the spectrogram deemed to be noise are set to one common floor value. An illustration of the results is given in Figure 5 where the spectrogram of an original utterance from the TIDigits database is given in (a), the spectrogram of the same utterance corrupted by AWG noise at 5 dB is given in (b) and the result of segmentation given in (c).

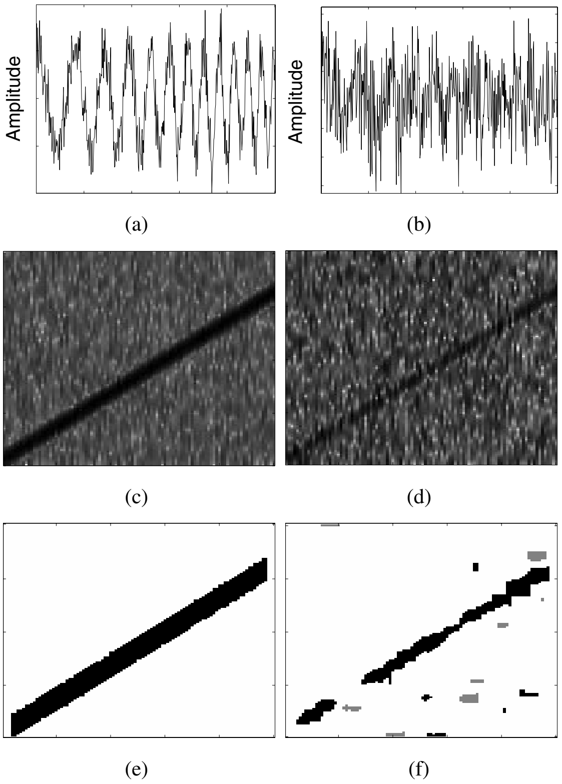
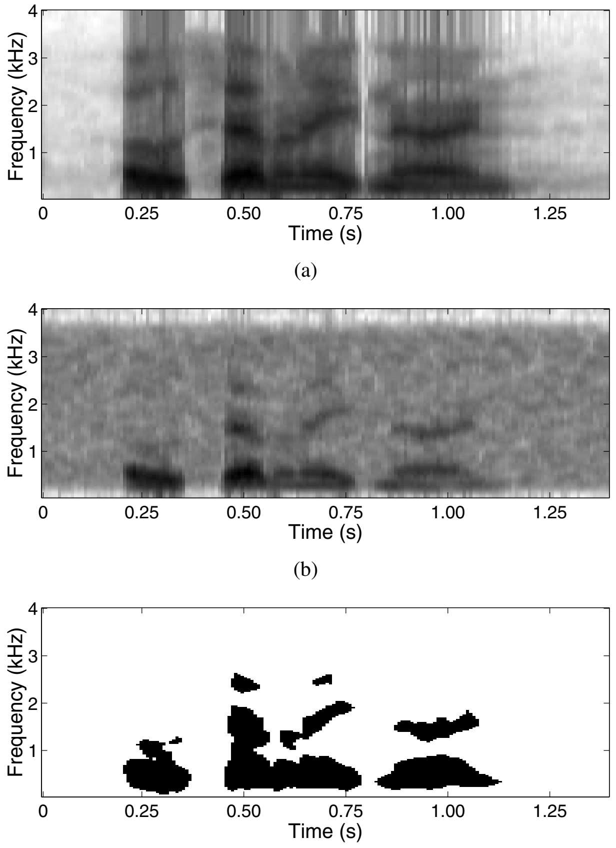


Figure 4: Dependence on SNR: time waveforms at (a) 10 dB and at (b) -5 dB, corresponding spectrograms (c) and (d), and results of segmentation (e) and (f). Horizontal time axis (0-0.5 s), vertical frequency axis (0-4 kHz).

Figure 6 shows ASR results in terms of word accuracy plotted against SNR levels of mismatched noise conditions for the untreated data set and the morphologically filtered set (morph set). In the clean case for both test and training the morph set gives perhaps surprisingly good results at 89%, given that all spectral amplitude information has been replaced by a single level (cf 98% for untreated set). Furthermore the robustness of the shape segmentation is illustrated by the more gradual degradation in the case of the morph set, falling to just 56% at 5 dB (cf 16% for untreated). The next stage is to replace some of the amplitude variation to the segmented speech signal areas.

5. Conclusions

In this paper experimental results relating to a recently proposed process for segmenting signals in the domain of spectrograms [1, 2] are described. The original work considers chirp test signals and experiments reported here use similar signals thereby independently corroborating the original findings. These first results are extended by then



(a)

(b)

(c)

Figure 5: Three spectrograms for (a) original clean speech utterance from the TIDigits database, (b) the same utterance degraded by AWG noise at 5 dB and (c) the result of signal segmentation.

examining segmentation performance at different SNRs. It is shown through a spectrogram area cost function that the segmentation process is robust down to 0 dB and below.

The experimental work is then extended to the speech domain. Areas deemed to be dominated by speech are segmented and set to a constant high amplitude. All other areas are set to a constant low floor level. Clearly this process preserves only the spectral shape of the speech. In so doing it looses local amplitude information of the speech but in doing so it also removes all noise (accept that which has corrupted the spectral shapes of the speech).

For the ASR experiments the TIDigits database is used in a standard AURORA 2 configuration. Under mismatched test and training conditions of 5 dB SNR for the test set only (clean training) a word accuracy of 56% is achieved. This compares with 16% when the same noisy test data is applied directly to the ASR system. In conclusion the approach to signal segmentation applied here to speech, shows that spectral shapes alone (without normal spectral amplitude variations) leads to perhaps surprisingly good ASR results in noisy conditions and provides good noise suppression. The next stage is to include spectral amplitude information along with appropriate noise compensation in the spectrogram regions occupied by speech.

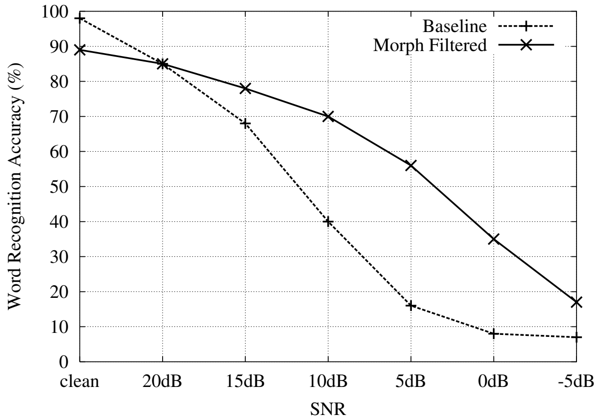


Figure 6: Word recognition accuracy against SNR for the baseline and morphologically filtered sets. The recognisers are trained on clean data and tested against data with noise added at the indicated level.

6. References

[1] Hory, C., Martin, N. and Chehikian, A., “Spectrogram Segmentation by means of Statistical Features for Non-stationary Signal Interpretation”, IEEE Trans. on Signal Processing, 50:2915–2925, 2002.

[2] Hory, C. and Martin, N., “Maximum Likelihood Noise Estimation for Spectrogram Segmentation Control”, Proc. ICASSP, Vol. 2, 2002, 1581–1584.

[3] Boll, S. F., “Suppression of Acoustic Noise in Speech using Spectral Subtraction”, IEEE Trans. on Acoustics Speech and Signal Processing, 27(2):113–120, 1979.

[4] Stahl, V., Fischer, A. and Bippus, R., “Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering”, Proc. ICASSP, Vol. 3, 2000, 1875–1878.

[5] Evans, N. W. D. and Mason, J. S., “Time-Frequency Quantile-Based Noise Estimation”, Proc. EUSIPCO, Vol. 1, 2002, 539–542.

[6] Ealey, D., Kelleher, H. and Pearce, D., “Harmonic Tunnelling: Tracking Non-stationary Noises During Speech”, Proc. Eurospeech, Vol. 1, 2001, 437–450.

[7] Hansen, J. H. L., “Morphological Constrained Feature Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect”, IEEE Trans. on Speech and Audio Processing, 2(4):598-614, 1994.

[8] Leonard, R. G., “A database for Speaker Independent Digit Recognition”, Proc. ICASSP, Vol. 3, 1984, 42.11–14.

[9] Hirsch, H. G. and Pearce, D., “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.