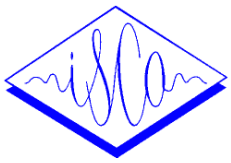

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A Dynamic-Window Weighted-RMS Averaging Filter Applied to Speaker Identification

Hong Tang, Xiaoyuan Zhu, Iain Macleod, Bruce Millar and Michael Wagner
TRUST* Project
Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, Australia
E-mail: hong@trust.anu.edu.au

Abstract

Based on concepts which are known to work well with 2-D images, this paper explores the application of a statistical noise reduction method to improving the robustness of speaker identification against background noise. A dynamic-window weighted-rms averaging criterion is used to detect noise and preserve the speech quality. Various forms of noise from the NOISEX-92 database were added to digitised speech to test the enhancement properties of our methods. Experimental results show that the proposed filter leads to improved speaker identification performance, even with certain of our supportedly "clean" speech data sets which were contaminated by high-level background noise.

Keywords

Noise Reduction, Speaker Identification.

1. Introduction

The performance of automatic speaker identification methods typically degrades markedly in the presence of additive noise [1]. As speaker identification or verification technologies move from the laboratory to practical application environments there is a clear need to develop systems which are more robust to likely forms of additive noise [2,3].

Many different methods have been studied for achieving noise robustness [1, 3-10]. These methods have included spectral subtraction [4] (in which a steady-state estimate of the noise spectrum is removed from the measured spectrum of the noisy speech), spectral mapping [5] (which directly exploits the approximate one-to-one relationship between spectral vectors in a clean speech environment and those in a noisy environment), and comb filtering [6] (in which the quasi-periodic nature of voiced-speech waveforms, which have narrow harmonically spaced bands of energy in the frequency domain, is exploited by tracking the harmonic fundamental and eliminating spectral components which do not conform). Each of these methods has its disadvantages. The first two rely on knowledge of a fixed noise environment and will generate their own "system" noise when the actual noise varies in spectrum or energy. Comb filters are highly dependent on accurate tracking of the fundamental frequency of voiced speech, which can be difficult at the onset of voicing and when this frequency varies rapidly. Furthermore a comb filter is applicable only to voiced speech.

This paper presents a dynamic-window weighted-rms averaging filter (DWWRAF) which was designed to enhance the performance of speaker identification when the speech signals contained additive noise. The DWWRAF filter is a multi-step adaptive filter which is designed (i) to respond to changes in the additive noise while not adding unpredictable "system" noise components and (ii) to be able to follow typical variations of fundamental frequency. It uses local statistical properties of the pressure-time waveform to suppress unexpectedly large variations while adaptively preserving the underlying speech signal information. The concepts on which the DWWRAF filter is based have been used quite successfully in the 2-D image processing domain [11], yielding improvements in effective SNR while at the same time preserving significant image edge details. The following sections describe the filter operation and its performance in detail.

2. Noise Model

The focus of this research is on reducing background noise that is acoustically or digitally added to speech. The additive noise model in the time domain is:

$$g_i = (1 - p) \times s_i + p \times n_i$$

where g_i denotes the corrupted speech, s_i denotes the clean speech, and n_i denotes the acoustic noise, all at position i. p is a noise weighting parameter which generates different signal-to-noise ratios (SNRs).

3. The DWWRAF Filter

Before describing the filter, we first define some symbols:

1. Let D be a sequence of M samples about centre i, and D' be a sequence of M' samples also about centre i, where $M' < M$.
2. For a noisy speech sample g_i, the local averages μ_i^{long} and μ_i^{short} in D and D' are defined as

*Technology for Robust User-conscious Secure Transactions

$$\mu_i^{long} = \frac{1}{M-1} \sum_{j \in D, j \neq i} g_j,$$

and

$$\mu_i^{short} = \frac{1}{M'} \sum_{j \in D'} g_j.$$

3. The difference d_i between g_i and the local average μ_i^{long} is defined as

$$d_i = (g_i - \mu_i^{long}).$$

4. The root mean square (RMS) deviation for sample points within D is defined as

$$\sigma_i = \left(\frac{1}{M-1} \sum_{j \in D} d_j^2 \right)^{1/2}$$

Two versions of the filter are described as follows:

$$g_i' = \begin{cases} g_i & \text{if } |d_i| \leq \alpha\sigma_i \\ \mu_i^{long} & otherwise \end{cases} \quad (1)$$

$$g_i' = \begin{cases} g_i & \text{if } |d_i| \leq \alpha\sigma_i \\ \mu_i^{short} & otherwise \end{cases} \quad (2)$$

where g_i' is the i^{th} output filtered speech sample, g_i is the i^{th} input noisy speech sample, μ_i^{long} and μ_i^{short} are the means of the sequences of speech samples M and M' (where $M' < M$ inside the processing window D), d_i is the difference between g_i and μ_i^{long}, σ_i is the standard deviation of g_k relative to μ_k^{long} (inside window D), and α is a parameter which defines the proportion of that standard deviation beyond which a deviant sample will be replaced. α was chosen to provide good performance.

Algorithm 1 was designed to smooth speech with high-level additive noise while Algorithm 2 was designed for speech which is less severely degraded. The rationale here is that the average error in replacing a suspected noise point by the local mean will be reduced in high-noise conditions by using a relatively wide window for computing this mean. The disadvantage is that fine detail in the underlying speech signal will tend to be attenuated, suggesting that a narrower window should be used where the amplitude of the added noise and hence the potential "replacement error" are smaller.

4. Experimental Evaluation

The impact of this filter was evaluated by examining the performance of a speaker identification system which had been trained on clean speech. Vector quantisation (VQ) codebooks were trained for 25 speakers (12 male/13 female) speaking 170 utterances which were designed to be representative of the phonemes of Australian English. The speech was sampled at 20,000 samples/second and at 16 bit resolution. The training utterances were recorded in a single session and an equivalent set of test utterances was recorded in separate session more than one week later. The identification algorithm was simply the selection of the speaker whose VQ codebook gives the minimum variance-weighted VQ distortion score when presented with the set of test utterances.

The DWWRAF approach has been applied to the set of test utterances both in their "clean" form and when subjected to additive noise from the NATO NOISEX-92 database, using the experimental procedure shown in Figure 1. Each noise source was added in turn to the test utterances in three proportions to create a wide range of recognition performances. The filter parameters were set to $M = 9, M' = 3$ and $\alpha = 0.04$ empirically.

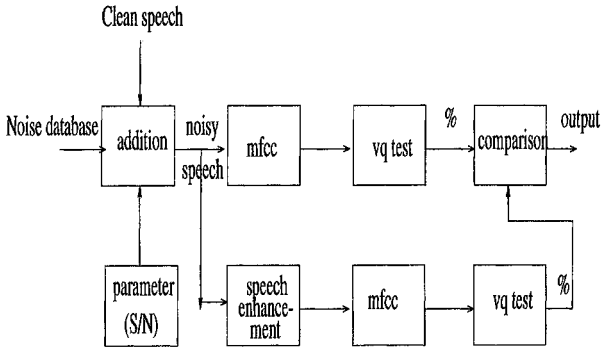

Fig. 1. The experimental procedure for speech enhancement

Seven types of noise have been added to the "clean" speech to test our algorithms: Speech babble noise, Office noise, Lynx helicopter noise, Factory noise, Speech-shaped white noise (STITEL), Car noise, and F16 cockpit noise. The dominant energy of these noise types is distributed in different frequency ranges. A description of the noise categories is given in Table I.

Tables II to VIII show the results of text-independent speaker identification using our two DWWRAF algorithms under the seven types of additive noise. Algorithm 3 was a simple nine point smooth of the data and was used as a benchmark against which to evaluate Algorithms 1 and 2.

The results indicate that of the two algorithms, Algorithm 1 is generally more effective in dealing with high-level noise, whereas Algorithm 2 is better at dealing with low-level noise. This is consistent with our filter design concepts.

For example, for Speech spectrum noise, Algorithm 1 is better when the SNR is less than 5dB whereas when the SNR is greater than 9dB, Algorithm 2 is more effective. For Office noise, when the SNR is less than 12dB, Algorithm 1 works better. For Factory noise, Algorithm 1 is better when the SNR is less than 15dB. For STITEL noise, Algorithm 2 is better when the SNR is greater than 9dB. For F16 cockpit noise, Algorithm 1 works better when the SNR is less than 12dB. Both of the algorithms work well in dealing with NATO noise except for Lynx helicopter noise when the SNR is less than 10dB and with Car noise when the SNR is greater than -6dB.

TABLE I

Description of the noise categories

No.	Source	Remarks
6	Multispeaker babble	similar to long-term speech spectrum; 100 people speaking in a canteen, 88dB.
12	Lynx helicopter noise	platform 97dB
14	office noise	opsroom of destroyer 70 dB
18	STITEL noise	Speech-shaped noise
20	F16 cockpit noise	co-pilot's seat, 300-600 feet, 500 knots, 103 dB
21	factory noise	car floor production, electrical welding, 83dB
23	car noise	Volvo 340, 120 km/h, 4th gear, wet asphalt road, 76.5dB

TABLE II

Speech spectrum noise

Recognition rate (before filtering)			Recognition rate (a fter filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.8	4.69	0.62	1	38.51
			2	16.65
			3	14.91
0.7	9.38	19.88	1	66.46
			2	71.43
			3	65.84
0.6	13.21	61.49	1	81.37
			2	91.93
			3	78.88

TABLE III

Car noise

Recognition rate (before filtering)			Recognition rate (after filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.8	-15.28	1.86	1	1.86
			2	5.59
			3	0
0.7	-10.60	34.16	1	9.32
			2	39.75
			3	2.48
0.6	-6.77	72.05	1	5.59
			2	69.57
			3	5.59

TABLE IV

Office noise

Recognition rate (before filtering)			Recognition rate (after filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.7	8.61	2.48	1	43.48
			2	18.01
			3	13.66
0.6	12.45	11.80	1	62.11
			2	55.63
			3	50.93
0.5	15.97	38.51	1	73.29
			2	79.50
			3	69.57

TABLE V

Factory noise

Recognition rate (before filtering)			Recognition rate (after filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.8	5.77	0	1	43.48
			2	41.72
			3	2.48
0.7	10.45	1.86	1	68.94
			2	48.54
			3	37.89
0.6	14.29	16.77	1	83.23
			2	80.12
			3	77.64

TABLE VI

STITEL noise

Recognition rate (before filtering)			Recognition rate (after filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.8	4.45	1.86	1	18.01
			2	8.07
			3	5.59
0.7	9.13	17.39	1	43.48
			2	57.14
			3	18.01
0.6	12.97	54.66	1	73.29
			2	84.47
			3	70.81

TABLE VII

Lynx helicopter noise

Recognition rate (before filtering)			Recognition rate (after filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.8	3.15	5.59	1	4.35
			2	4.35
			3	0
0.7	7.83	59.63	1	2.48
			2	52.80
			3	1.24
0.6	11.67	90.68	1	15.53
			2	85.71
			3	13.65

TABLE VIII

F16 cockpit noise

Recognition rate (before filtering)			Recognition rate (after filtering)	
p	SNR in dB	rate in %	Algorithm	rate in %
0.8	7.74	13.04	1	93.17
			2	54.66
			3	45.65
0.7	12.42	46.58	1	96.89
			2	54.66
			3	41.30
0.6	16.26	74.53	1	96.89
			2	**100**
			3	96.27

Note that the raw noise power was much higher for Car noise than other forms, so the chosen values of p give poorer SNRs. However, relatively good speaker identification performance is still obtained, suggesting that the dominant energy bands in the car noise lie outside those which are important for speaker identification.

Algorithm 3 performed uniformly worse than the other two algorithms. This indicates that while the DWWRAF approach does contain elements of smoothing, its adaptive and non-linear behaviour is a major contributor to its performance.

A further experiment with a single word has been carried out to illustrate the performance of the filter in more detail. The waveform of the word 'say' is presented in Figure 2a. Figure 2b shows the impact of additive F16 cockpit noise at an SNR of 10dB. The fricative is obliterated and the waveform envelope of the vocalic part is perturbed. The result of application of Algorithm 1 (Fig.2c) is to reduce the high-frequency components of the noise covering the fricative section and to make no obvious change in the vocalic part. The differential result of application of Algorithm 2 (Fig.2d) is to retain some of the high-frequency components removed when Algorithm 1 is used.

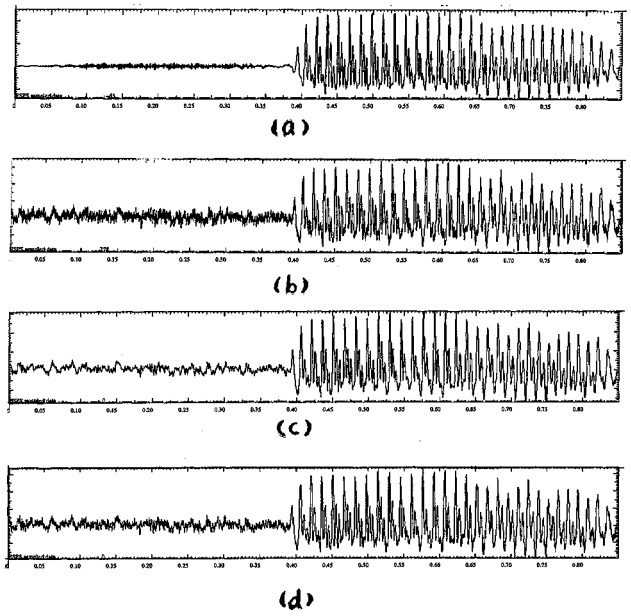

Fig. 2. **Examples of filter operation on the waveform of "say"; (a) original wave, (b) original wave plus noise (SNR=10dB), (c) filtered wave (Algorithm 1), (d) filtered wave (Algorithm 2).**

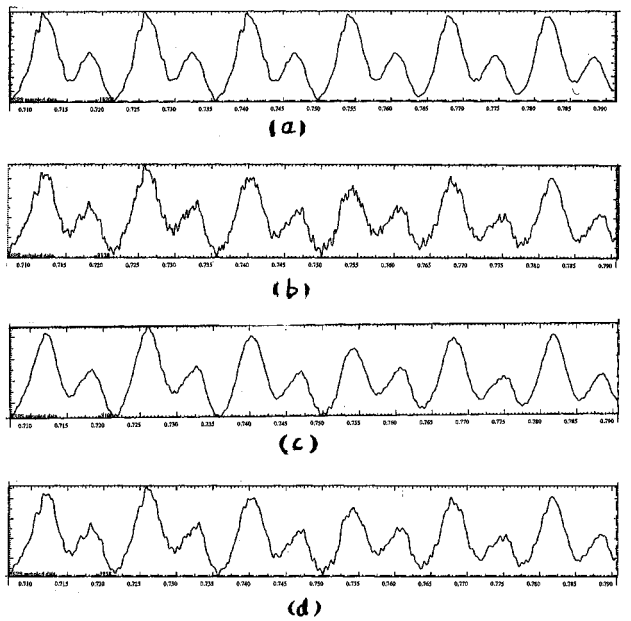

Fig. 3. **Examples of filter operation on the waveform of /i/ vowel; (a) original wave, (b) original wave plus noise (SNR=10dB), (c) filtered wave (Algorithm 1), (d) filtered wave (Algorithm 2).**

The effect of the filter is examined more closely in Figure 3 in which the waveform of the vowel /i/ taken from the end of the word 'say' is shown so that detail obscured by the scale of Figure 2 can be revealed. The vowel /i/ has a low frequency first formant and very high frequency second and third formants. The latter can be seen as a minor ripple on the major first formant wave in Figure 3a. The addition of the noise visually obscures the higher formant information (Fig.3b), but is largely removed using Algorithm 1 (Fig.3c). Unfortunately most of the higher formant ripple is also removed, leaving just a slight remnant of the noise ripple. Algorithm 2 retains more of the distinctive aspects of the higher formant ripple but also a greater amount of the noise ripple than Algorithm 1.

5. Conclusion

The results of the above experiments suggest that nonlinear and adaptive filtering techniques previously used to good effect in the image domain can lead to enhanced robustness of speaker identification in the presence of background noise. The mechanisms underlying the observed improvements remain to be explored in greater detail.

6. Acknowledgement

This research has been carried out on behalf of Harry Triguboff AM Research Syndicate.

References

[1] J.P. Openshaw and J.S. Mason, "Optimal Noise-masking of Cepstral Features for Robust Speaker Indentification", *ESCA Workshop on Automatic Speaker Recognition, Indetification and Verification*, pp.231-234, April, 1994.

[2] A. Varga, "Assessment for automatic speech identification: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech identification systems", *Speech Communication*, Vol.12, pp.247-251, 1993.

[3] Y. Gong and W.C. Treurniet, "Speech identification in noisy environments : A study", *CRC-TN 93-0.2*, June 1993.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-27, No.2, April 1979, pp.113-120.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *Proceedings of ICASSP'79*, pp.208-211, April 1979.

[6] J.S. Lim, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-26, No.4, August 1978 pp.354-358.

[7] H. Sheikhzadeh, H. Sameti, L. Deng and R.L. Brennan, "Comparative performance of spectral subtraction and HMM-based speech enhancement strategies with application to hearing aid design", *Proceedings of ICASSP'94*, Vol.1, pp.I-13-16, April 1994

[8] M.A. Ramalho, "A new speech enhancement technique with application to speaker identification", *Proceedings of ICASSP'94*, Vol.1, pp.I-29-32, April 1994

[9] Y. Ephraim, "Statistical-model-based speech enhancement systems", *Proceedings of the IEEE*, Vol.80, No.10, October 1992, pp.1524-1555.

[10] A. Erell, "Energy conditioned spectral estimation for identification of noisy speech", *IEEE Transactions on Speech and Audio Processing*, Vol.1, No.1, January 1993,pp.84-89.

[11] H. Tang, "A human-machine interactive system for efficient image restoration", *Proceedings of International Conference on Acoustics, Speech and Signal Processing 1994*, Adelaide, 18-22 April, Vol.5, pp.81-84, Australia, 1994.