Robust features for noise speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences

Author : Kuo-Hwei Yuo, Hsiao-Chuan Wang

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Trajectory filtering of autocorrelation sequence
- Procedure to compute RAS-MFCC
- Experiments

Introduction

- The idea is to filter the temporal trajectories of short-time one-sided autocorrelation sequences of speech such that the noise effect is removed.

- The filtered sequences are denoted by the relative autocorrelation sequences (RASs), and the mel-scale frequency cepstral coefficients (MFCC) are extracted from RAS instead of the original speech.

- This new speech feature set is denoted as RAS-MFCC.

Trajectory filtering of autocorrelation sequence

- The noisy speech signal is blocked into *M* frames of *N* samples, and is modeled as

$$y(m,n) = x(m,n) + w(m,n),$$
$$0 \le m \le M-1,\ 0 \le n \le N-1. \qquad (1)$$

- If the noise is uncorrelated with the speech, it follows that the autocorrelation of the noisy speech is the sum of autocorrelation of the clean speech $x(m,n)$ and autocorrelation of the noise $w(m,n)$.

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(m,k),$$
$$0 \le m \le M-1,\ 0 \le k \le N-1. \qquad (2)$$

k is the autocorrelation sequence index.

Trajectory filtering of autocorrelation sequence

- If the noise is stationary, the autocorrelation sequences of noise in all frames can be assumed to be identical and $r_{ww}(m,k)$ will depend only on autocorrelation index *k*.

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(k),$$
$$0 \le m \le M-1,\ 0 \le k \le N-1. \qquad (3)$$

- Differentiating both sides of Eq. (3) with respect to frame index *m* for all *k* yields

$$\frac{\partial r_{yy}(m,k)}{\partial m} = \frac{\partial r_{xx}(m,k)}{\partial m},$$
$$0 \le m \le M-1,\ 0 \le k \le N-1. \qquad (4)$$

- Eq. (4) demonstrates that, in each frame, the RAS of noisy speech is equal to the RAS of clean speech. This implies that the effect of noise is removed.

Trajectory filtering of autocorrelation sequence

- The RASs are approximated by

$$\frac{\partial r_{yy}(m,k)}{\partial m} \cong \frac{1}{T_L} \sum_{t=-L}^{L} t r_{yy}(m+t,k),$$

$$0 \le m \le M-1,\ 0 \le k \le N-1, \qquad (5)$$

$$\mathbf{T}_L = \sum_{t=-L}^{L} t^2 \qquad (6)$$

- Eq. (5) can be interpreted as a filtering process on the temporal autocorrelation trajectory using an FIR filter that has a transfer function given by Eq.

$$H(z) = \frac{1}{T_L} \sum_{t=-L}^{L} t z^t \qquad (7)$$

- Eq. (7) is a high-pass filter.

Procedure to compute RAS-MFCC

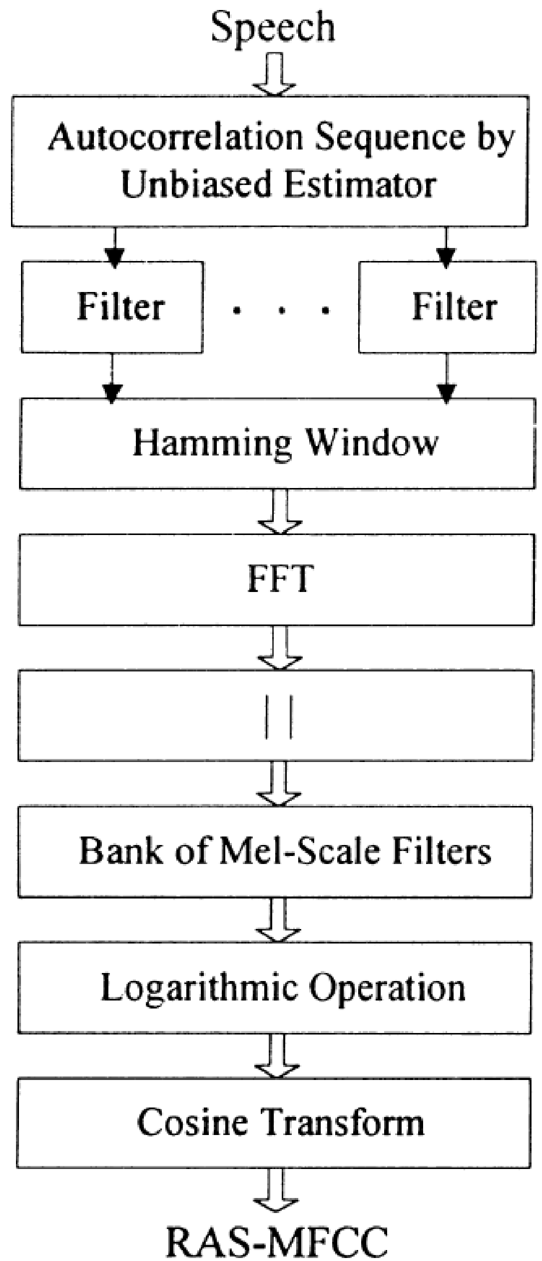
- The procedure for computing RAS-MFCC is summarized as follows:

  1. The original speech is segmented into overlapping frames with N sample data points per frame, and the N-point one-sided autocorrelation sequence for each frame is computed using the unbiased autocorrelation estimator given by

$$r_{yy}(m,k)=\frac{1}{N-k}\sum_{j=0}^{N-1-k} y(m,\, j)y(m,\, j+k),$$

$$0 \le k \le N-1. \qquad (8)$$

  2. The RAS's are obtained by processing all temporal trajectories of one-sided autocorrelation coefficients with the FIR high-pass filters given by Eq. (7).


Speech
Autocorrelation Sequence by Unbiased Estimator
Filter
Filter
Hamming Window
FFT
| |
Bank of Mel-Scale Filters
Logarithmic Operation
Cosine Transform
RAS-MFCC

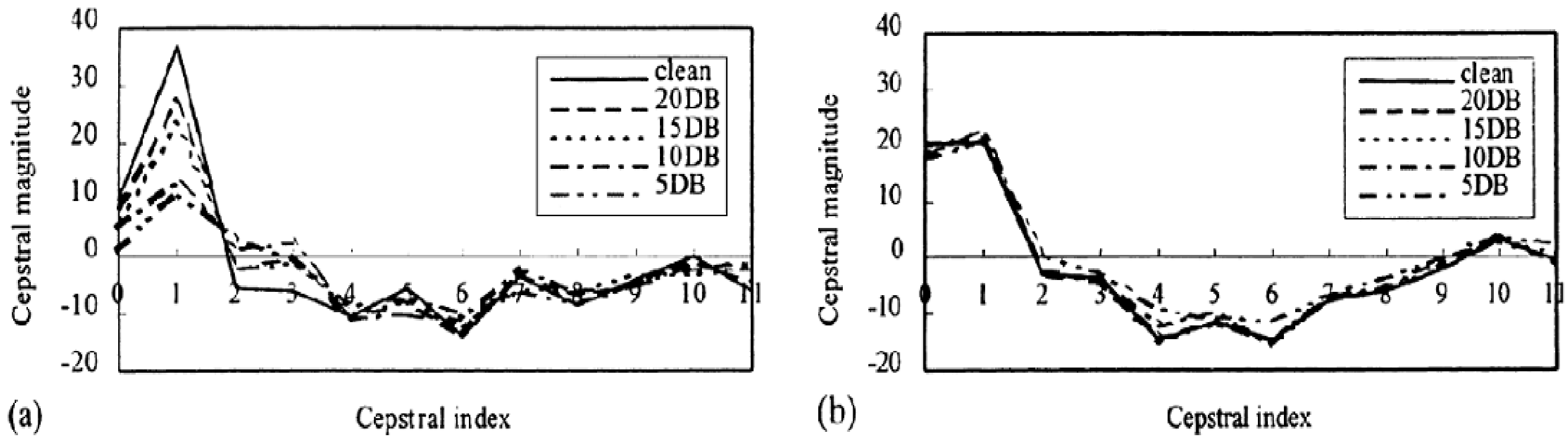


Fig. 2. Comparison of various feature types at different levels of SNR in white noise corruption: (a) Effect of additive white noise on MFCC; (b) effect of additive white noise on RAS-MFCC

Experiments

- A Mandarin isolated-digit database, collected from 100 speakers (50 males and 50 females) at an 8 kHz sampling rate in a noise-free environment, was the clean speech database.

- There are four experiments.

Experiments(1)

- This high-pass filter is given by Eq. (7), where L is a parameter in this filtering operation. In this experiment, the influences of the parameter L and estimator types on RAS-MFCC are examined.

- Table 1. The recognition rate using MFCC features is seriously degraded by white noise, while the RAS-MFCC features are robust to white noise.

Table 1
The recognition rates for RAS-MFCC at various parameters and comparison to MFCC with white noise corruption

SNR (dB)	Feature								
	MFCC	RAS-MFCC							
		Unbiased estimator				Biased estimator			
		Three frames ($L=1$)	Five frames ($L=2$)	Seven frames ($L=3$)	Nine frames ($L=4$)	Three frames ($L=1$)	Five frames ($L=2$)	Seven frames ($L=3$)	Nine frames ($L=4$)
Clean	0.957	0.938	0.932	0.900	0.885	0.953	0.933	0.889	0.852
20	0.784	0.927	0.912	0.883	0.878	0.938	0.924	0.878	0.850
15	0.602	0.903	0.893	0.857	0.852	0.911	0.897	0.867	0.833
10	0.461	0.859	0.859	0.834	0.831	0.822	0.828	0.835	0.823
5	0.319	0.735	0.775	0.781	0.788	0.607	0.681	0.730	0.750
0	0.130	0.467	0.580	0.620	0.653	0.354	0.509	0.542	0.508

Experiments(2)

- Table 2 shows the recognition rates using the different features for speech recognition in the presence of white noise corruption.

- RAS-MFCC outperforms the other features in noise and performs well even in severe noise conditions such as SNR at 0 dB, but in clean conditions RAS-MFCC is slightly less accurate than MFCC and LPCC.

Table 2
Comparison of recognition rates for the various feature types with white noise corruption

Feature type	SNR					
	Clean	20 dB	15 dB	10 dB	5 dB	0 dB
(a) The recognition rates for cepstral features with white noise corruption						
LPCC	0.951	0.703	0.517	0.378	0.276	0.130
MFCC	0.957	0.784	0.602	0.461	0.319	0.130
RAS-MFCC	0.932	0.912	0.893	0.859	0.775	0.580
(b) The recognition rates for cepstral and delta-cepstral features with white noise corruption						
LPCC	0.983	0.825	0.676	0.480	0.270	0.113
MFCC	0.985	0.871	0.780	0.621	0.373	0.108
RAS-MFCC	0.969	0.957	0.943	0.933	0.879	0.754

Experiments(3)

- This experiment compared RAS-MFCC with traditional MFCC paired with alternative noise compensation techniques.
- The experiment was conducted on four types of noise: white, F16, factory, and babble noises.

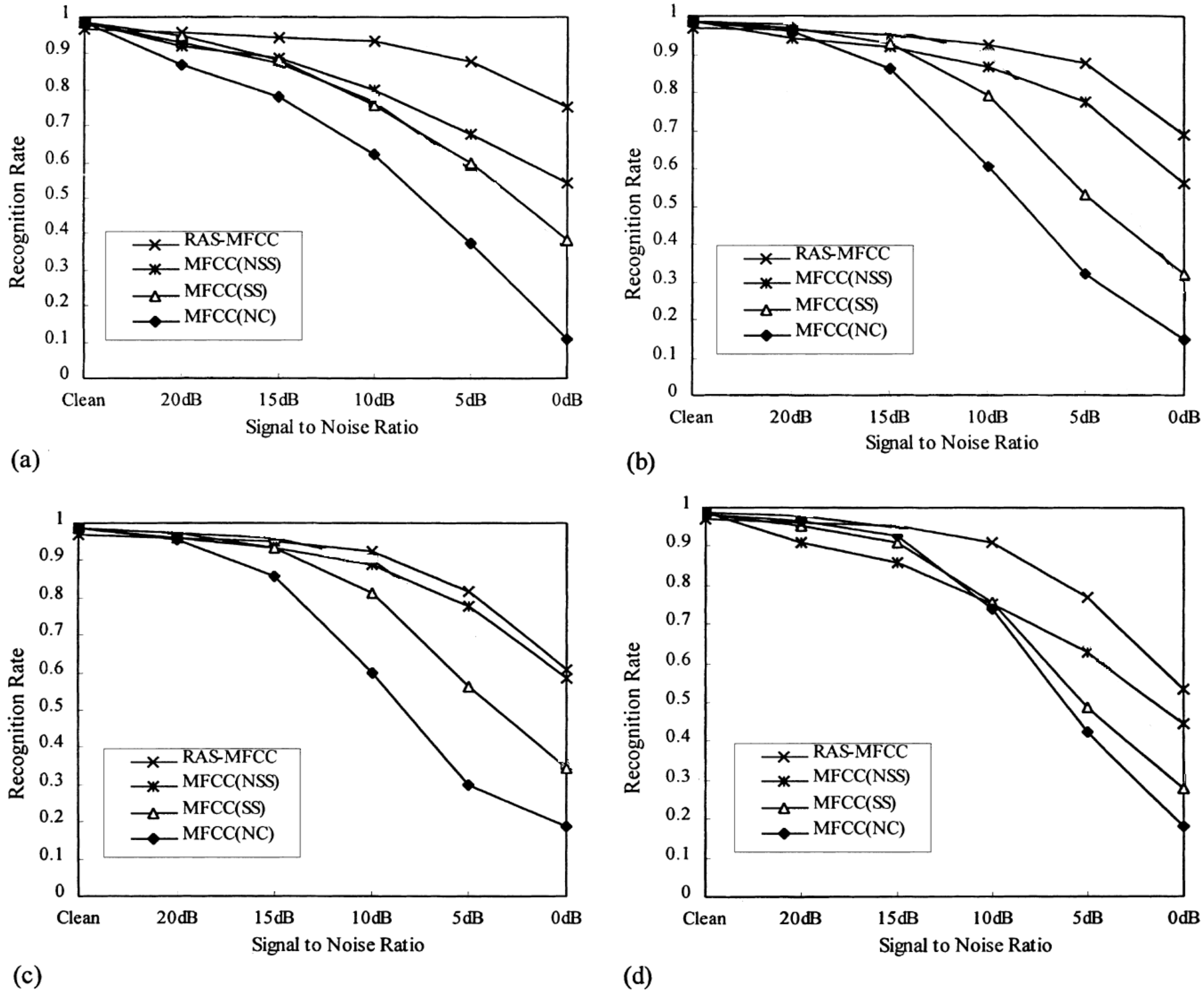


Fig. 5. Comparison of recognition rates for the various noise compensation methods with colored noise corruption: (a) white noise; (b) factory noise; (c) F16 noise and (d) babble noise.

Experiments(4)

- We evaluate three filters, H_{RAS}, H_{RASTA} and H_{Hirsch}, denoting the RAS filter, RASTA filter and Hirsch's filter, and apply these three filters to the DFT spectrum, subband, logarithmic subband, and autocorrelation domains.

$$H_{RAS}(z) = \frac{1}{10}(2z^2 + z - z^{-1} - 2z^{-2})$$

$$H_{RASTA}(z) = \frac{z^4(2 + z^{-1} - z^{-3} - 2z^{-4})}{10(1 - 0.98z^{-1})}$$

$$H_{Hirsch}(z) = 1 - \frac{\sum_{t=1}^{16}(0.94)^t z^{-t}}{\sum_{t=1}^{16}(0.94)^t}$$

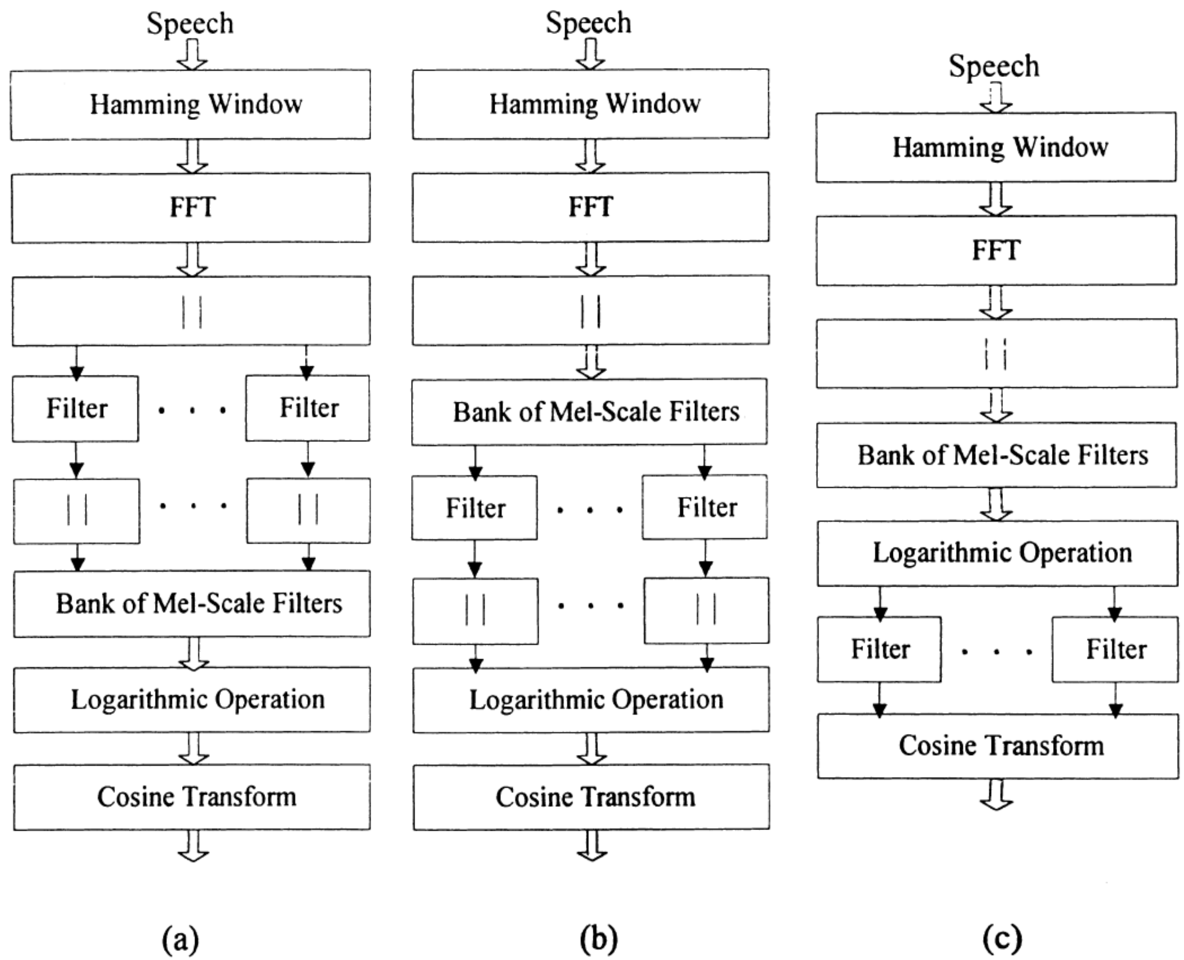


Fig. 6. Calculation of robust features based on temporal filtering in (a) DFT-magnitude domain, (b) subband domain and (c) logarithmic subband domain.

Table 3
The recognition rates for various temporal filtering techniques

SNR (dB)	Filter												
	No filter	DFT spectrum [a]			Subband [a]			Log subband [a]			Autocorrelation [a]		
		H_{RAS}	H_{RASTA}	H_{Hirsch}	H_{RAS}	H_{RASTA}	H_{Hirsch}	H_{RAS}	H_{RASTA}	H_{Hirsch}	H_{RAS}	H_{RASTA}	H_{Hirsch}
(a) with white noise corruption													
Clean	0.985	0.986	0.987	0.980	0.989	0.976	0.987	0.975	0.986	0.987	0.969	0.938	0.964
20	0.871	0.941	0.953	0.908	0.910	0.895	0.907	0.913	0.900	0.915	0.957	0.933	0.940
15	0.780	0.870	0.892	0.832	0.849	0.844	0.834	0.859	0.836	0.855	0.943	0.916	0.923
10	0.621	0.745	0.796	0.753	0.719	0.711	0.683	0.718	0.659	0.688	0.933	0.893	0.873
5	0.373	0.476	0.665	0.517	0.531	0.533	0.507	0.522	0.371	0.465	0.879	0.860	0.752
0	0.108	0.172	0.462	0.238	0.319	0.378	0.341	0.327	0.243	0.281	0.754	0.758	0.484
(b) with factory noise corruption													
Clean	0.985	0.986	0.987	0.980	0.989	0.976	0.987	0.975	0.986	0.987	0.969	0.938	0.964
20	0.960	0.968	0.971	0.965	0.954	0.897	0.948	0.932	0.944	0.962	0.963	0.930	0.963
15	0.863	0.917	0.917	0.900	0.913	0.813	0.890	0.884	0.885	0.912	0.950	0.920	0.950
10	0.605	0.747	0.778	0.722	0.807	0.683	0.789	0.744	0.681	0.792	0.925	0.888	0.916
5	0.322	0.540	0.551	0.471	0.552	0.476	0.551	0.554	0.392	0.479	0.874	0.773	0.831
0	0.146	0.258	0.344	0.261	0.322	0.331	0.334	0.263	0.208	0.191	0.688	0.568	0.554

[a] Refers to domain.