Temporal Structure Normalization of Speech Feature for Robust Speech Recognition

Author :Xiong Xiao, Eng Siong Chng, Haizhou Li

Professor : 陳嘉平

Reporter : 楊治鏞

Outline

- Introduction
- Normalization of temporal structure
- Experiment and results

Introduction

- In this paper, we will examine the normalization strategy and investigate its effect on the ASR.

- We design temporal filters to normalize the utterance-dependent feature PSD (i.e. the modulation spectrum) to a reference PSD function, which in effect is to normalize the feature's temporal structure.

Temporal Structure Varies With Noise

- We use the Mel-scaled filterbank cepstral coefficient (MFCC) as the feature for speech recognition.

- Let $x_k(n)$ be the cepstral coefficient of the n^{th} frame and k^{th} MFCC channel of an utterance.

- we have K time series, $x_1(n)$ to $x_K(n)$, where K is the number of channels.

Temporal Structure Varies With Noise

- The time series of these raw features are first processed by CMN and CVN, referred to as mean and variance normalization (MVN) hereafter, then normalized by the proposed temporal structure normalization (see Fig. 1).

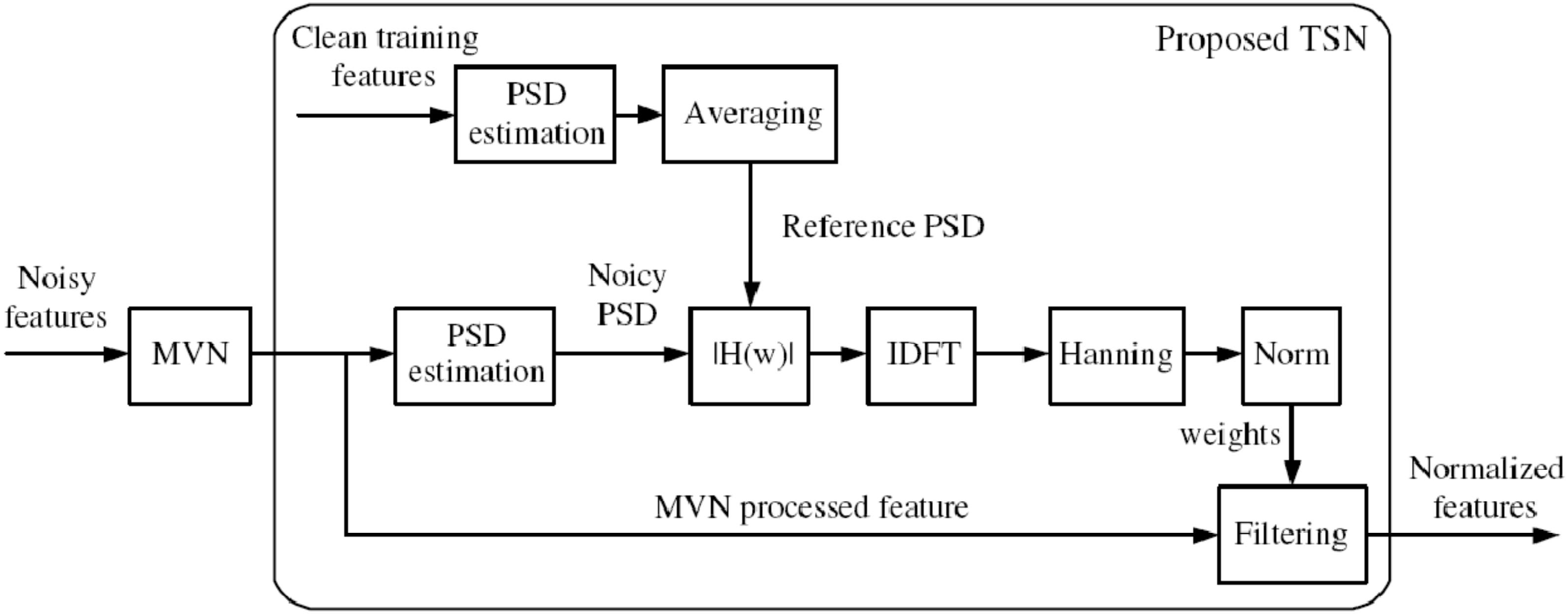
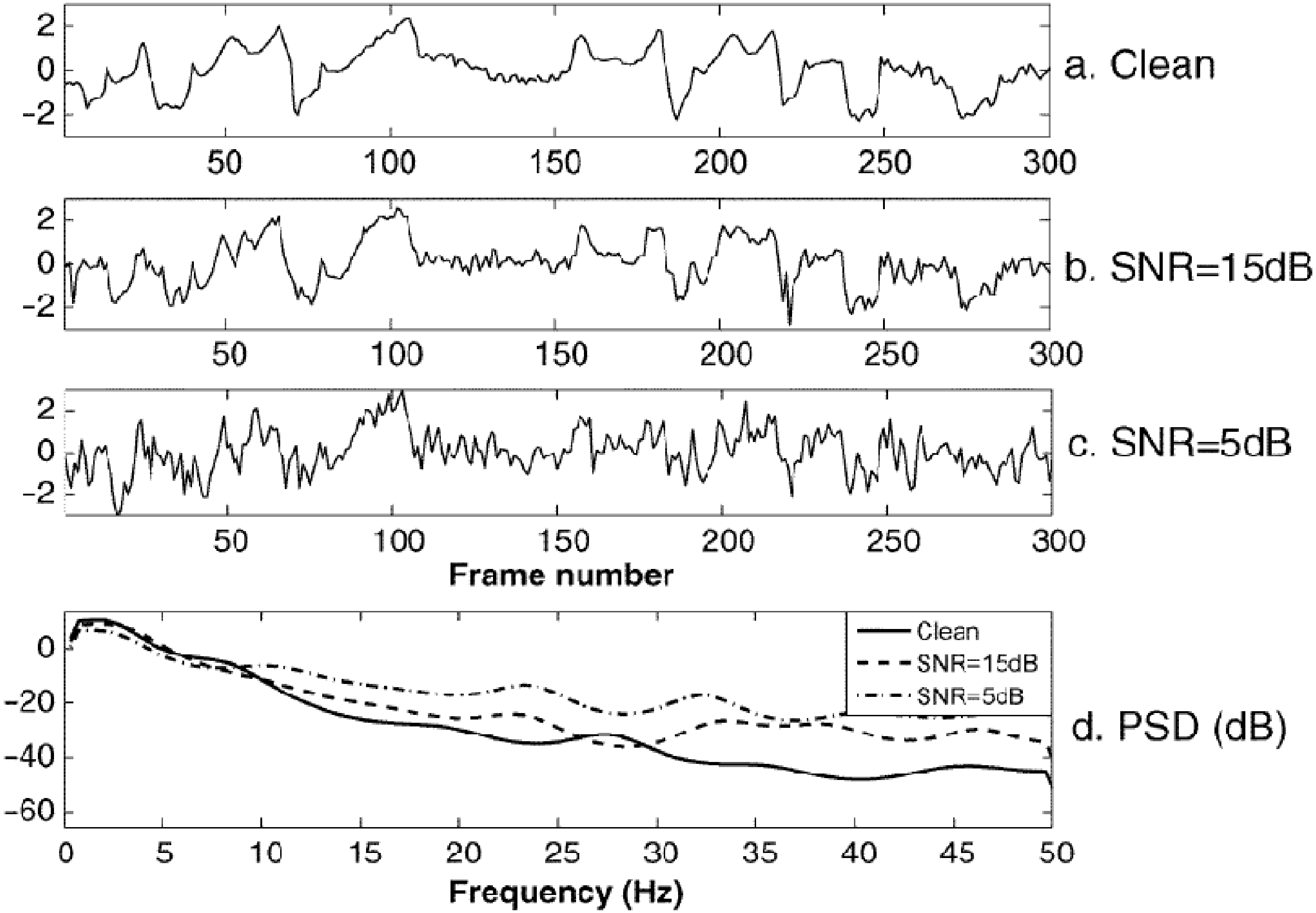


Fig. 1. The block diagram of the proposed framework

Temporal Structure Varies With Noise

- We examine a speech utterance that is corrupted by additive car noise.
- Fig. 2(a)–(c) shows the time series of the first MFCC feature $c1$ after MVN processing for three SNR levels.
- Fig. 2(d) illustrates their corresponding PSD functions.


2
0
-2
50
100
150
200
250
300
a. Clean
b. SNR=15dB
c. SNR=5dB
Frame number
0
-20
-40
-60
0
5
10
15
20
25
30
35
40
45
50
Frequency (Hz)
Clean
SNR=15dB
SNR=5dB
d. PSD (dB)

Normalizing the Temporal Structure Using the Square-root Wiener Filters

- Let $y_k(n)$ be the observed noisy speech feature series for the k^{th} channel.
- $x_k(n)$:clean
- $v_k(n)$:noise

$$y_k(n) = x_k(n) + v_k(n), \ \text{for}\ k = 1, \ldots, K$$

Normalizing the Temporal Structure Using the Square-root Wiener Filters

- Let $P_x^k(\omega)$, $P_y^k(\omega)$ and $P_v^k(\omega)$ be the PSD of $x_k(n)$, $y_k(n)$ and $v_k(n)$.

$$P_y^k(\omega) = P_x^k(\omega) + P_v^k(\omega)$$

- The magnitude response of the square-root Wiener filter is

$$|H_k(\omega)| = \sqrt{P_x^k(\omega) \big/ P_v^k(\omega)}$$

Training the reference PSD functions

- As the PSD of the clean feature $P_x^k(\omega)$ is unknown, we instead use an averaged PSD function $\overline{P}_x^k(\omega)$ to evaluate $|H_k(\omega)|$.

- We call $\overline{P}_x^k(\omega)$ the reference PSD functions and obtain them by averaging the clean feature’s PSD over multiple utterances.

Training the reference PSD functions

- Calculate the feature PSD of all channels of all training utterances $P_x^{k,m}(\omega)$, for $k = 1, \ldots, K$ and $m = 1, \ldots, M$, where M is the number of utterances used for training $\bar{P}_x^k(\omega)$

$$\bar{P}_x^k(\omega) = \frac{1}{M} \sum_{m=1}^{M} P_x^{k,m}(\omega), \qquad k = 1, \ldots, K$$

Normalizing Temporal Structure

- Find the filter's weights using the inverse discrete Fourier transform (IDFT).

  $$w_k(i) = IDFT(|H_k(\omega)|)$$

- Extract the central taps of $w_k(i)$ to form $w'_k(i)$.

- Apply Hanning window on $w'_k(i)$ to reduce truncation effect.

Normalizing Temporal Structure

- Normalize the sum of the weights to one to ensure that the filter's gain is unity at zero frequency.

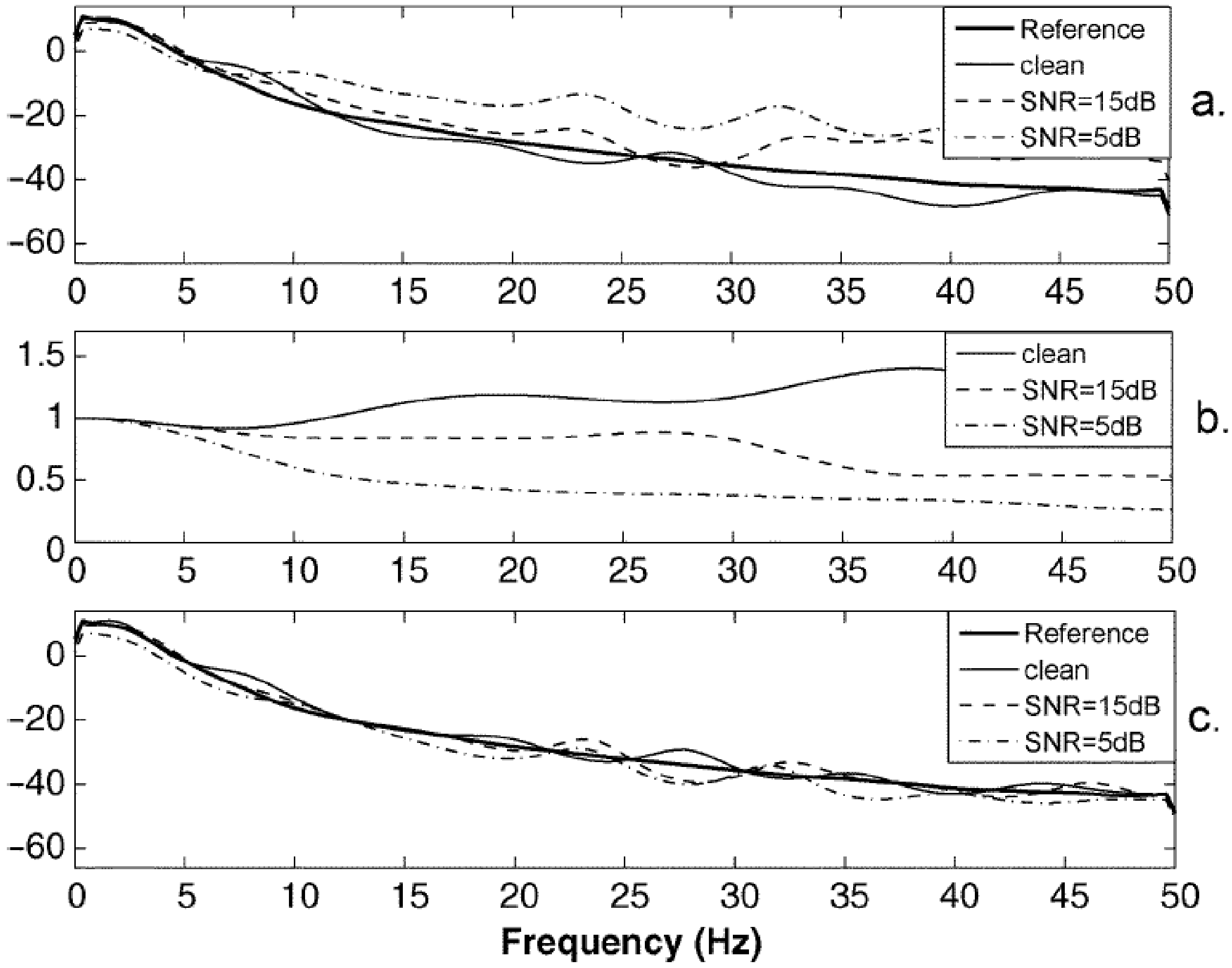
- The normalized features are estimated as

$$\hat{x}_k(n) = y_k(n) \otimes w'_k(i) \qquad k = 1, \ldots, K$$

Experiment Settings

- AURORA-2
- The training features for TSN1’s reference functions are processed by MVN only.
- Those for TSN2 are first processed by MVN and then smoothed by the ARMA filter used in MVA.

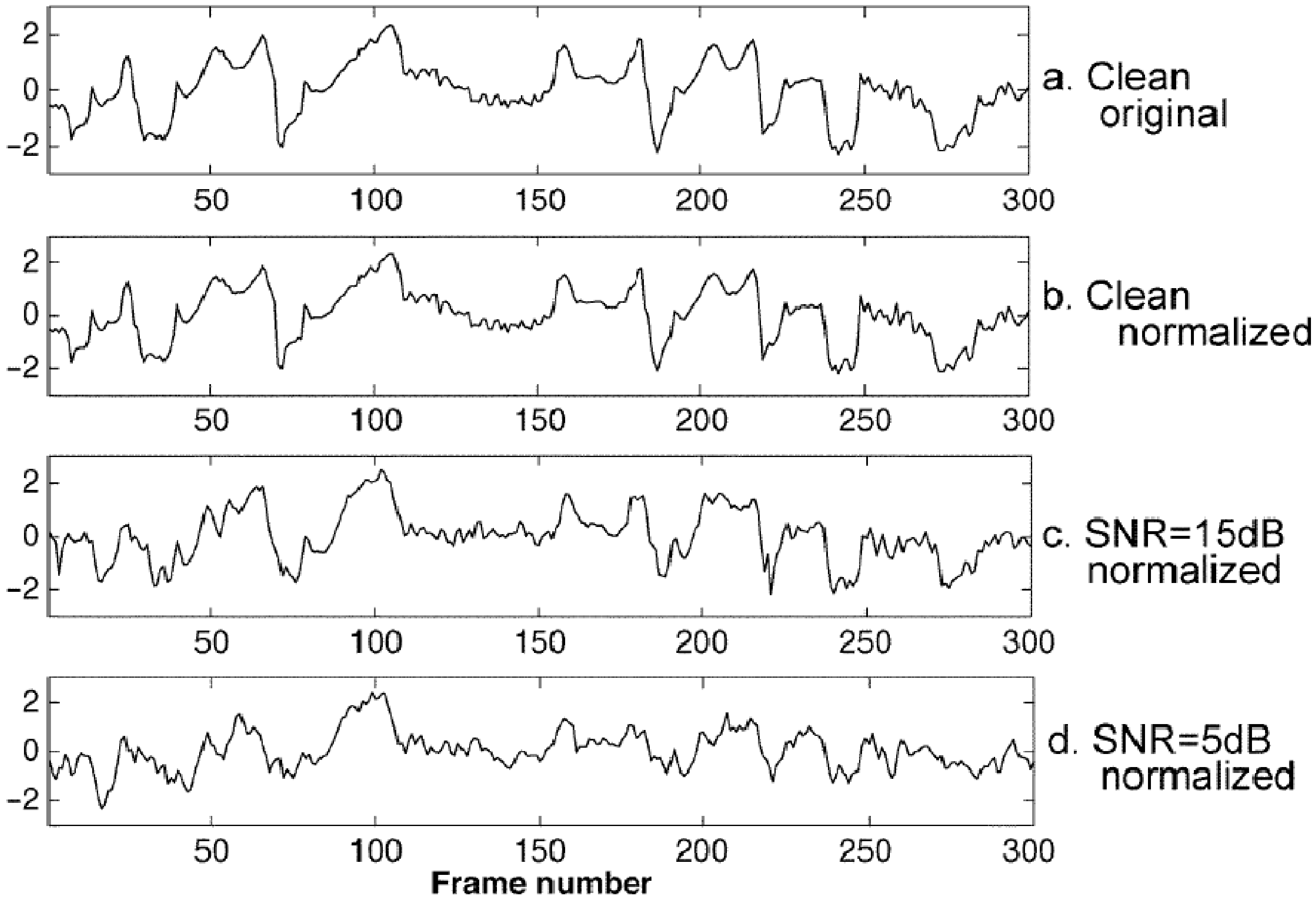
The normalization effects of TSN1



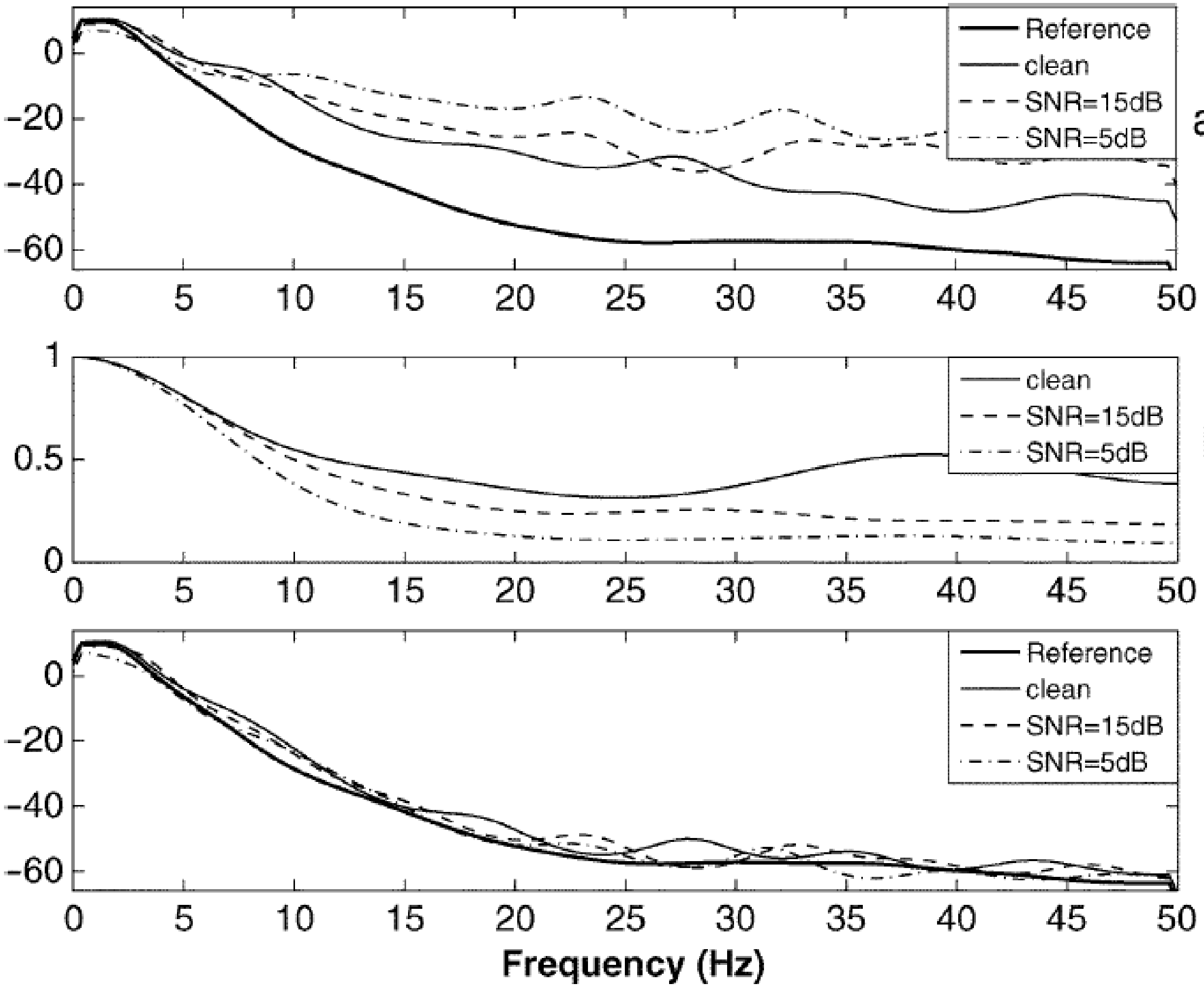
a. Noisy PSD (dB)

b. Transfer function

c. Normalized PSD (dB)


2
0
-2
50
100
150
200
250
300
a. Clean original
b. Clean normalized
c. SNR=15dB normalized
d. SNR=5dB normalized
Frame number

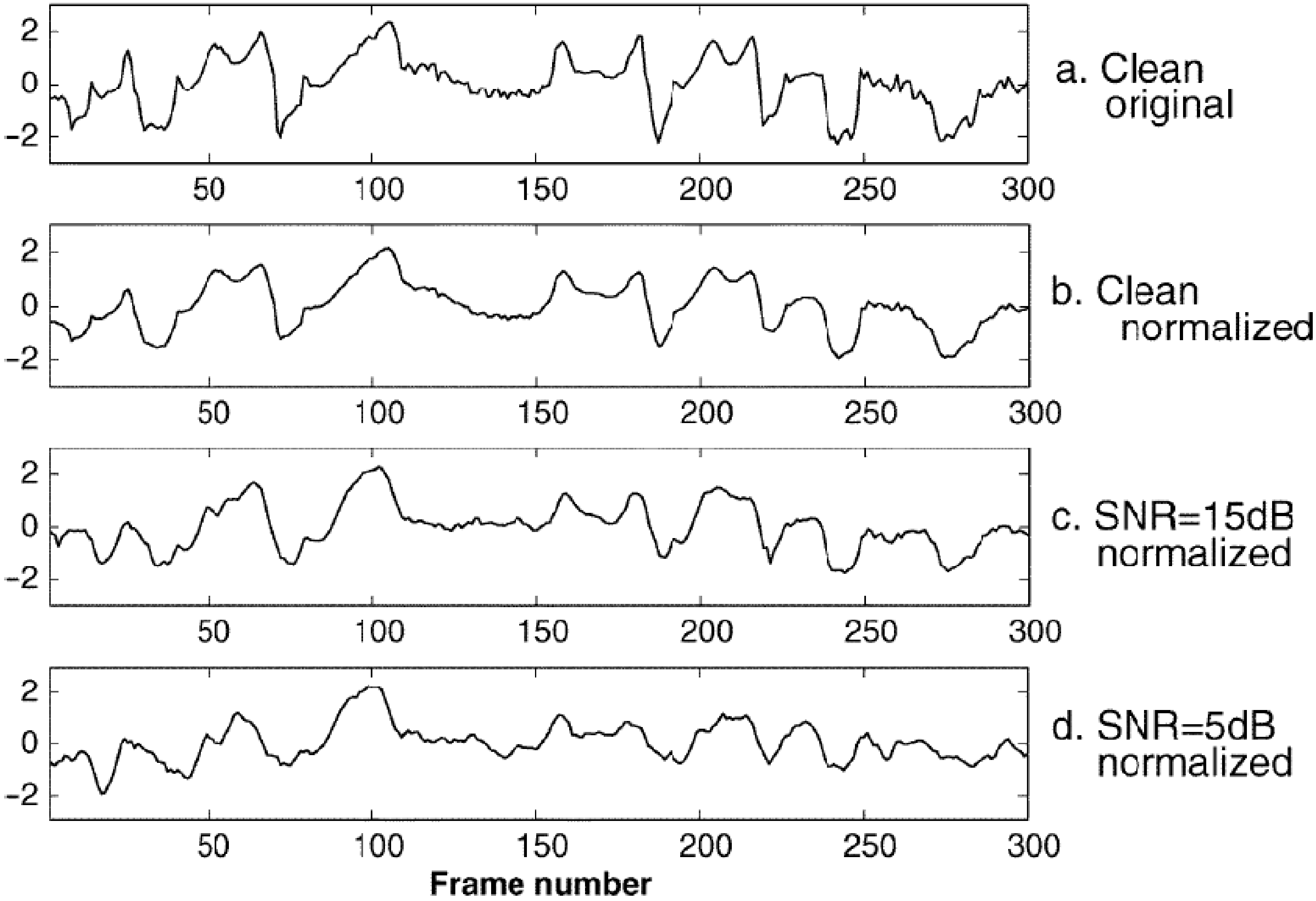
The normalization effects of TSN2



a. Noisy PSD (dB)

b. Transfer function

c. Normalized PSD (dB)


2
0
-2
50
100
150
200
250
300
a. Clean original
b. Clean normalized
c. SNR=15dB normalized
d. SNR=5dB normalized
Frame number

Recognition Results

- we compare the performance of the proposed normalization schemes with 4 other methods using the AURORA-2 framework.
  - a) MVN: CMN followed by CVN.
  - b) RASTA: MVN followed by RASTA filtering.
  - c) MVA (M=3): MVN followed by ARMA filtering.
  - d) LPF: MVN followed by LPF filtering.
  - e) TSN1: MVN followed by TSN1.
  - f) TSN2: MVN followed by TSN2.

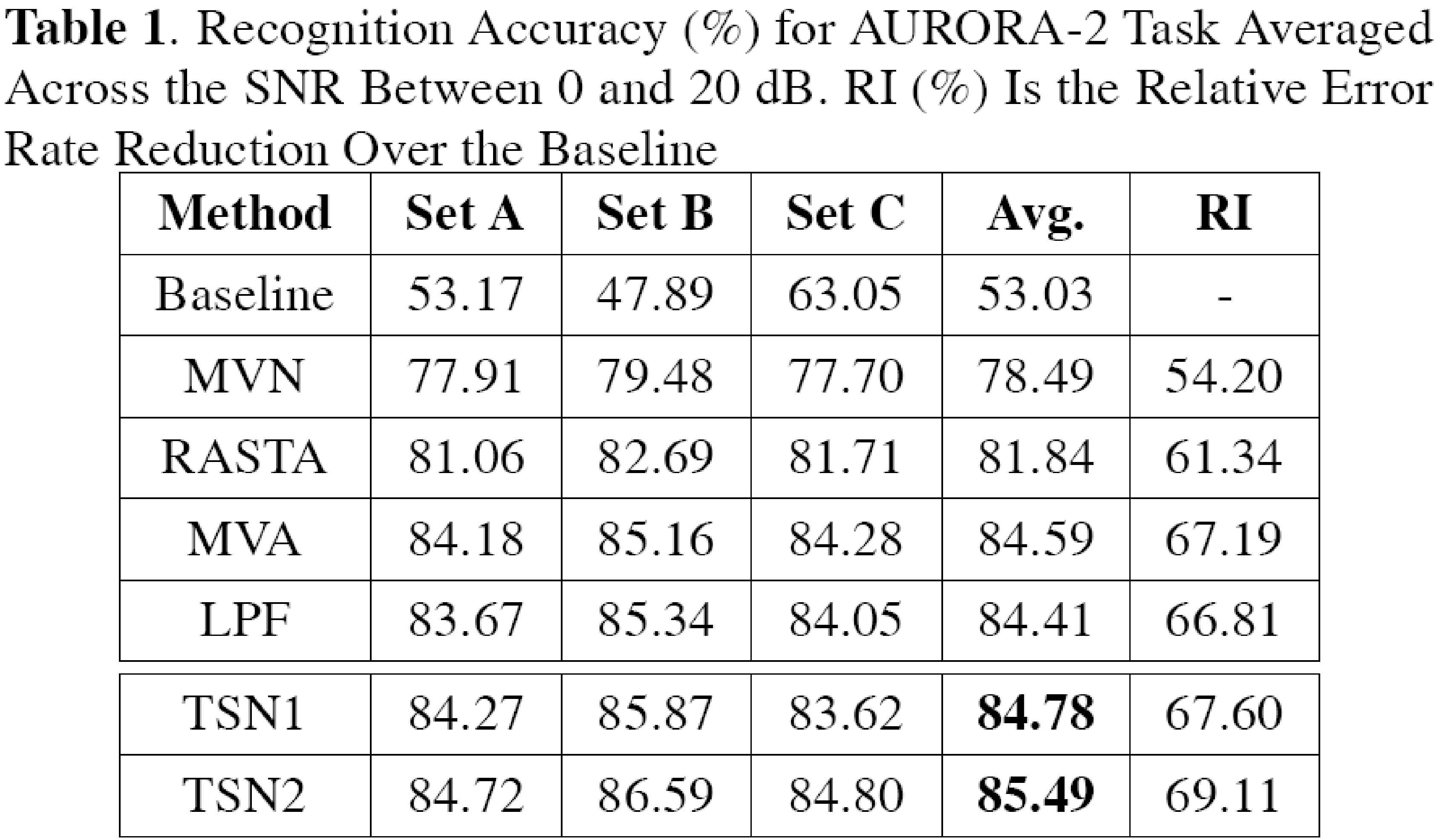


Table 1. Recognition Accuracy (%) for AURORA-2 Task Averaged Across the SNR Between 0 and 20 dB. RI (%) Is the Relative Error Rate Reduction Over the Baseline

Method	**Set A**	**Set B**	**Set C**	**Avg.**	**RI**
Baseline	53.17	47.89	63.05	53.03	-
MVN	77.91	79.48	77.70	78.49	54.20
RASTA	81.06	82.69	81.71	81.84	61.34
MVA	84.18	85.16	84.28	84.59	67.19
LPF	83.67	85.34	84.05	84.41	66.81
TSN1	84.27	85.87	83.62	**84.78**	67.60
TSN2	84.72	86.59	84.80	**85.49**	69.11

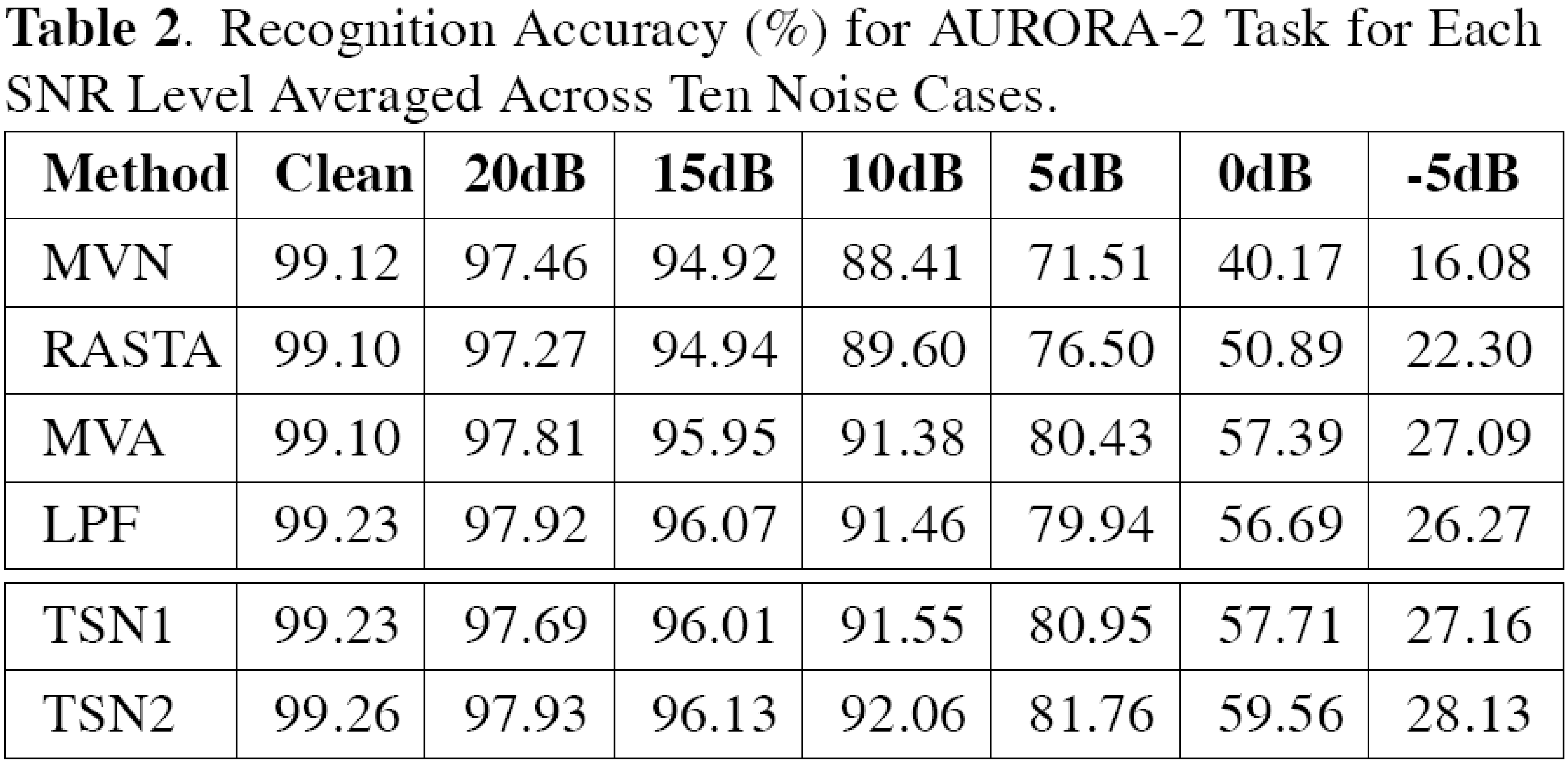


Table 2. Recognition Accuracy (%) for AURORA-2 Task for Each SNR Level Averaged Across Ten Noise Cases.

Method	Clean	20dB	15dB	10dB	5dB	0dB	-5dB
MVN	99.12	97.46	94.92	88.41	71.51	40.17	16.08
RASTA	99.10	97.27	94.94	89.60	76.50	50.89	22.30
MVA	99.10	97.81	95.95	91.38	80.43	57.39	27.09
LPF	99.23	97.92	96.07	91.46	79.94	56.69	26.27
TSN1	99.23	97.69	96.01	91.55	80.95	57.71	27.16
TSN2	99.26	97.93	96.13	92.06	81.76	59.56	28.13