Subband Temporal Modulation Spectrum Normalization for Automatic Speech Recognition in Reverberant Environments

Authors: *Xugang Lu, Masashi Unoki, Satoshi Nakamura*

Professor:陳嘉平
Reporter: 吳柏鋒

Outline

- Introduction
- Reverberation effect on temporal envelopes
- Proposed temporal modulation spectrum normalization algorithm
- Experimental

Introduction

- In this paper, we first investigated the reverberation effect on subband temporal envelopes(STE) by using the modulation transfer function (MTF).

- Based on the investigation, we proposed an algorithm which normalizes the subband temporal modulation spectrum (TMS) to reduce the diffusion effect of the reverberation.

Introduction

- During the normalization, both the subband TMS of the clean and reverberated speech are normalized to a <u>reference</u> TMS calculated from a clean speech data set for each frequency subband.

- Based on the normalized subband TMS, the <u>inverse</u> Fourier transform was done to restore the subband temporal envelopes by keeping their original phase information.

Reverberation effect on temporal envelopes

- Reverberation effect can be quantified by using the temporal modulation spectrum.
- Consider the reverberation effect by using the <u>smoothed</u> temporal modulation PSD to smooth out the details of the PSD due to linguistic context effect.

Reverberation effect on temporal envelopes

- The average of ensemble of the STEs can be quantified using the average of the power spectral density (PSD) of the STE as:

$$P_{xx}(\Omega) = \left\langle \frac{\left|A_x(\Omega)\right|^2}{N} \right\rangle ; P_{yy}(\Omega) = \left\langle \frac{\left|A_y(\Omega)\right|^2}{N} \right\rangle$$

suppose $a_x(t)$ and $a_y(t)$ are the subband temporal envelopes (STEs) of the clean and reverberation speech, and their Fourier transforms are $A_x(\Omega)$ and $A_y(\Omega)$, the Ω is the modulation frequency , N is the length of the STE and $\langle . \rangle$ is the ensemble average operator

Proposed temporal modulation spectrum normalization algorithm

- It is difficult to estimate the STMTF or the inverse filters.

- We attempt to normalize the subband temporal modulation PSDs of the clean and reverberated speech to a reference modulation PSD.

Proposed temporal modulation spectrum normalization algorithm

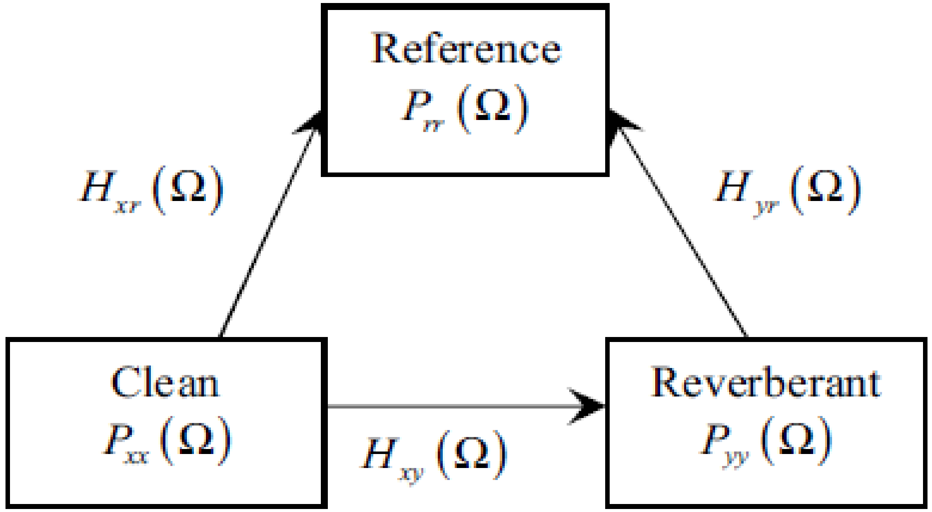


Figure 2: Normalization of the modulation power spectrum.

The $H_{xr}(\Omega)$ and $H_{yr}(\Omega)$ can be regarded as the PSDs of the modulation transfer filters between the clean, reverberated and the the reference environments.

Proposed temporal modulation spectrum normalization algorithm

- the reference PSD is set as the average PSD of the STE of a prior clean speech data set in each frequency band as:

$$P_{rr}(\Omega) = \frac{1}{M}\sum_{i=1}^{M} P^{i}{}_{xx}(\Omega)$$

Where $P^{i}{}_{xx}(\Omega)$ is the modulation PSD of the i-th clean utterance, i=1, 2, …, *M*, with *M* is the total number of the speech utterances.

Proposed temporal modulation spectrum normalization algorithm

- The PSD of the STMTFs between the clean and reverberated speech is:

$$H_{xr}(\Omega) = \frac{P_{rr}(\Omega)}{P_{xx}(\Omega)}; H_{yr}(\Omega) = \frac{P_{rr}(\Omega)}{P_{yy}(\Omega)}$$

$$\longrightarrow \quad H_{xy}(\Omega) = H_{xr}(\Omega)\, H_{yr}(\Omega)^{-1}$$

Proposed temporal modulation spectrum normalization algorithm

Experimental

- One clean training data set from AURORA-2J data corpus in used in this study.
- Using the speaker to microphone distance (SMD=400cm).

Experimental

- Using the smoothed temporal modulation PSD to smooth out the details of the PSD due to linguistic context effect.

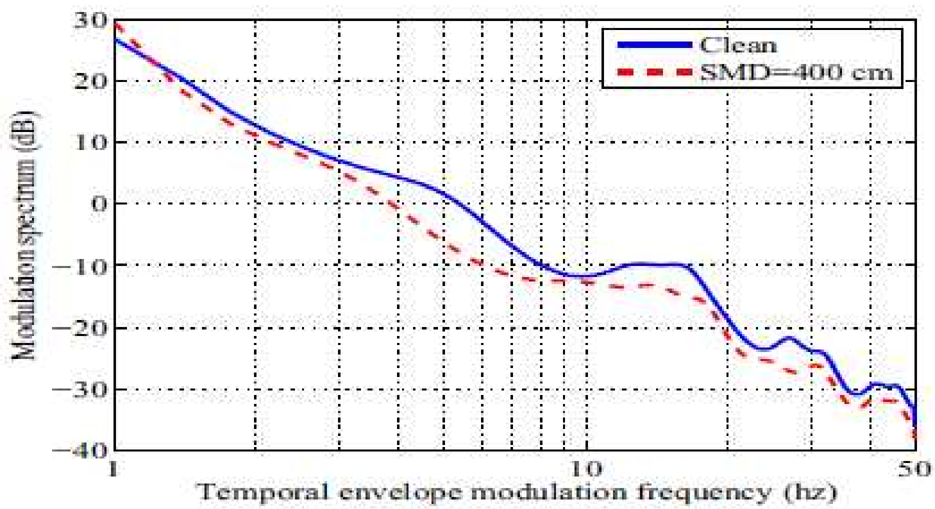


Figure 1: Temporal modulation spectrum of clean (solid) and reverberated (dashed) speech (with center frequency 1kHz).

Experimental

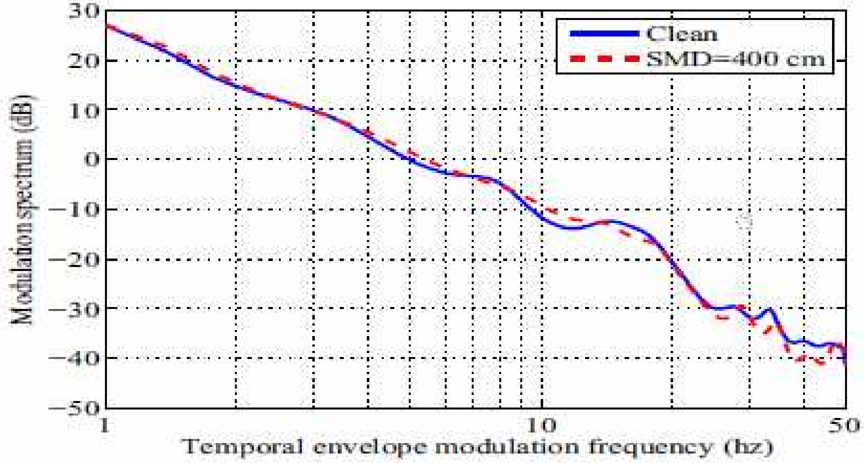
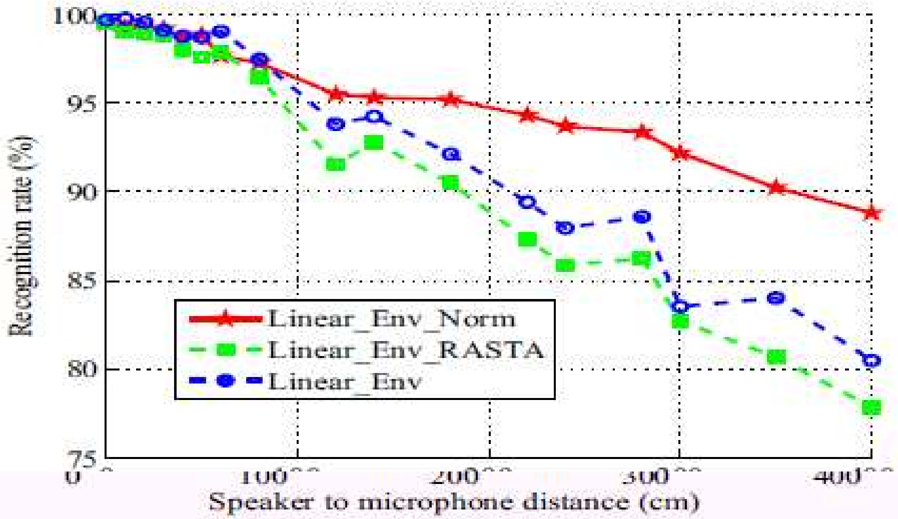


Figure 3: Normalized subband temporal modulation PSDs of the clean (solid) and reverberated (dashed) speech.

Experimental



gure 6: Recognition performance for filtering on the subband
mporal envelope.