HISTOGRAM-BASED SUBBAND POWER WARPING AND SPECTRAL AVERAGING FOR ROBUST SPEECH RECOGNITION UNDER MATCHED AND MULTISTYLE TRAINING

Electrical & Computer ENGINEERING

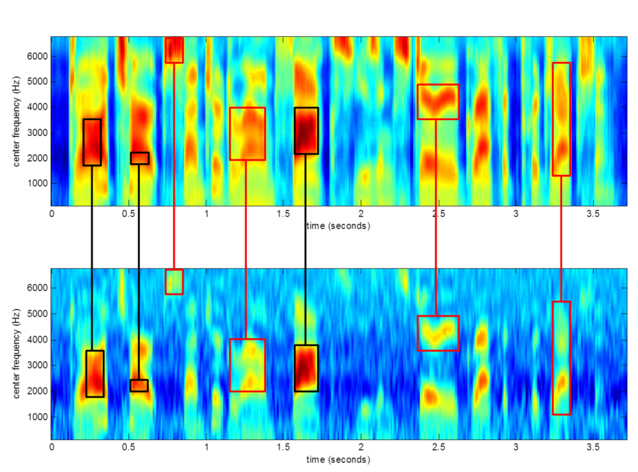
Mark J. Harvilla and Richard M. Stern
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

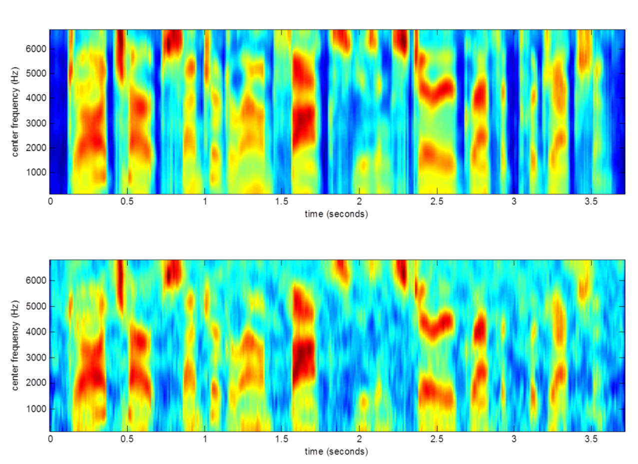
This paper describes a new algorithm that increases the robustness of speech recognition systems by matching the power histograms of the input in each frequency band to those obtained over clean training data, and then mixing together the processed and unprocessed spectra. It is shown that taking a weighted average between the processed and unprocessed power spectra contributes to further gains in recognition accuracy. Results are obtained for multiple speech recognition systems, noise types, and training conditions illustrating the broad utility of this approach. The algorithm is called **CSAWH** ("see-saw") for *Compensatory Spectral Averaging and Warping using Histograms*.

Motivation

- The spectral representation of noisy speech retains a residue of the local contrast observed with clean speech.
- The two figures in the left column below illustrate, highlighted in black, the time-frequency regions strong enough to be unaffected by the addition of noise; red highlighted regions indicate a residual contrast in the noise spectrum that can be approximately reconstituted through histogram normalization.
- The two figures in the right column show clean and noisy spectra after CSAWH processing.



Clean (top) and noisy (bottom) unprocessed spectrograms



Clean (top) and noisy (bottom) spectrograms after CSAWH processing

Histogram Matching

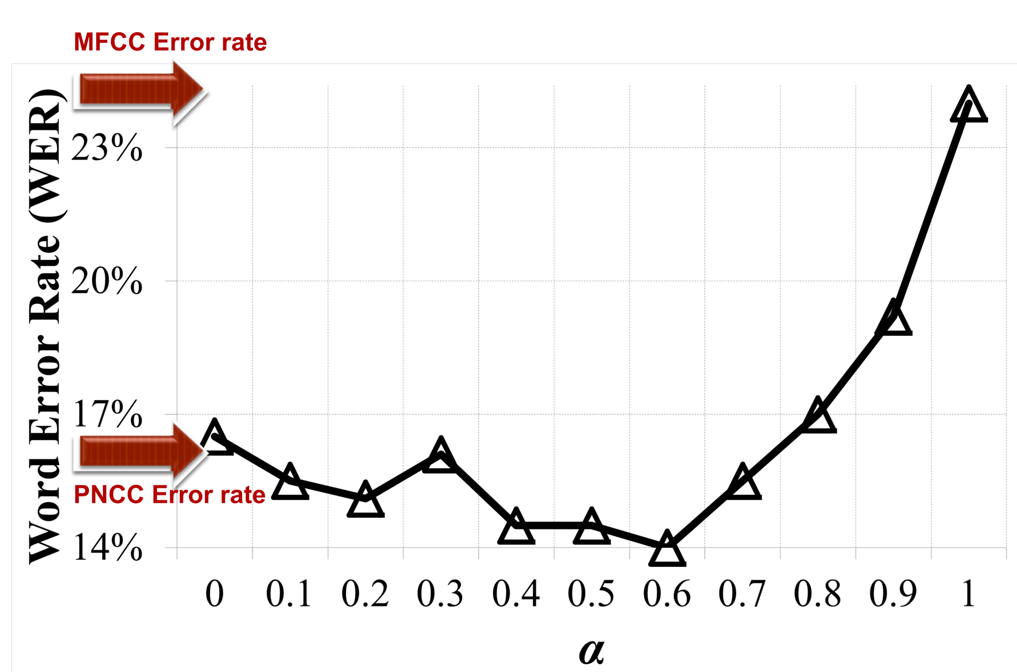
- 100-bin histograms of the log spectral magnitude are obtained over clean training data, one for each subband of the log spectrum.
- For each input utterance, conventional non-parametric histogram matching is performed to map the distribution of each subband of the input to the corresponding reference histogram.

This effectively normalizes the dynamic range by mapping the local maximum and minimum of each log spectral channel to the global maximum and minimum.

Spectral Averaging

$$P_{out}[n,k] = \alpha P_{orig}[n,k] + (1-\alpha)P_{proc}[n,k], 0 \leq \alpha \leq 1$$

- Spectral averaging is particularly helpful in the matched and multistyle training conditions, as well as when the data are filtered.
- The optimal value of α is sensitive to SNR, noise type, and training style.
- Spectral averaging can improve results up to an additional 15.5% relative (see figure below).

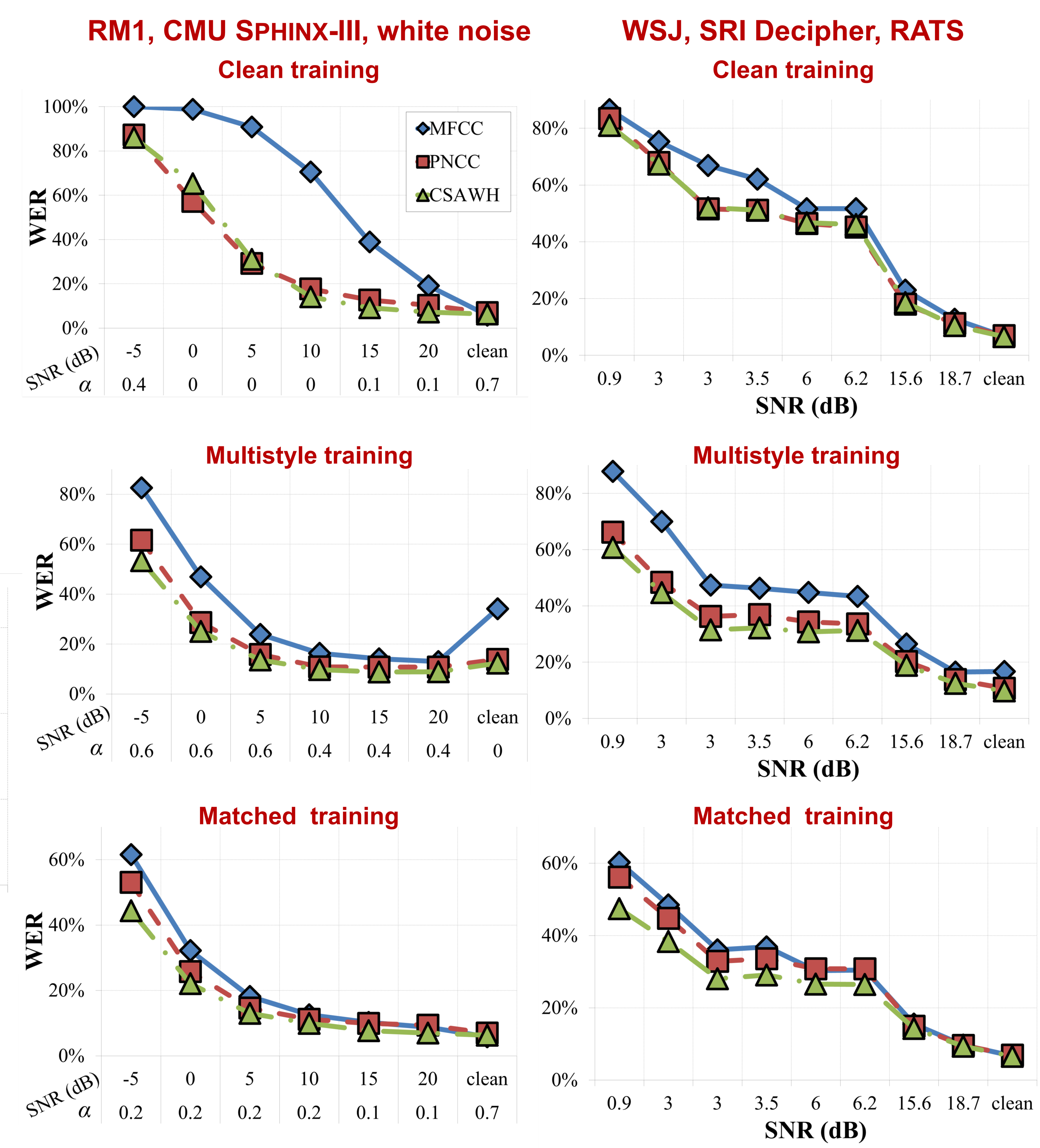


CSAWH performance for different values of α in 5-dB white noise under multistyle training

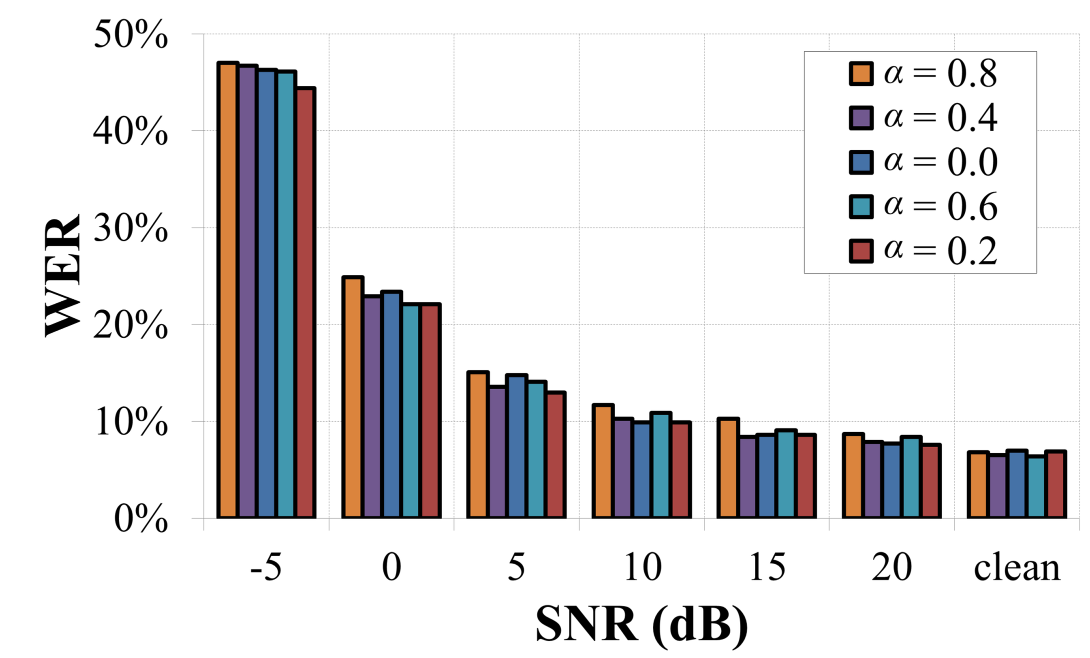
Experimental Results

Two primary experiments were run:

1. RM1 degraded by additive white Gaussian noise run on CMU's SPHINX-III recognizer.
   - The mixing parameter α was chosen manually for each SNR and training condition for optimal results.
   - Some results with α fixed are also shown.
2. WSJ degraded by RATS-like additive noise and linear channel filtering run on SRI DECIPHER; Columbia's `renoiser` tool was used to estimate noise parameters from real RATS data.
   - The mixing parameter α was fixed at 0.85 for all tests to gauge practical performance.



RM1, CMU Sphinx-3, white noise
matched training



CSAWH performance for different values of α under matched training; slight improvements are consistently observed for α = 0.2 in this case

Summary

- Histogram-based CSAWH processing performs comparably to PNCC under clean conditions.
- Histogram-based CSAWH processing in combination with spectral averaging achieves best performance for these tasks under multistyle and matched training conditions.

References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoust., Speech and Signal Processing, pp. 113–120, April 1979.
[2] C. Kim and R.M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in IEEE Int. Conf. on Acoust., Speech and Signal Processing, May 2010.
[3] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in IEEE Int. Conf. on Acoust., Speech and Signal Processing, May 1996.
[4] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in Int. Conf. on Spoken Language Processing, 2000, vol. 4, pp. 556–559.
[5] A. de la Torre et al., "Non-linear transformations of the feature space for robust speech recognition," in IEEE Int. Conf. on Acoust., Speech and Signal Processing, 2002, pp. 401–404.
[6] F. Hilger, Quantile based histogram equalization for noise robust speech recognition, Ph.D. thesis, Computer Science Department, RWTH Aachen University, Aachen, Germany, 2004.
[7] R.C. Gonzalez and R.E.Woods, Digital Image Processing, pp. 128–138, Pearson Prentice Hall, Upper Saddle Ridge, New Jersey, third edition, 2008.
[8] C. Kim, Signal processing for robust speech recognition motivated by auditory processing, Ph.D. thesis, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2010.
[9] D. Ellis, "RENOISER - Utility to decompose and recompose noisy speech files," http://labrosa.ee.columbia.edu/projects/renoiser/.