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

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 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, illustrated 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 reconstructed through histogram normalization.
- The two figures in the right column show clean and noisy spectra 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.

Experimental Results
Two primary experiments were run:
1. RM1 degraded by additive white Gaussian noise run on CMU’s PHINX-III recognizer.
   - The mixing parameter $\alpha$ was chosen manually for each SNR and training condition for optimal results.
   - Results are obtained for unprocessed power spectra contributes to further gains in recognition accuracy. Results are obtained for unprocessed power spectra.
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 $\alpha$ was fixed at 0.85 for all tests to gauge practical performance.

Spectral Averaging
$$P_{\text{proc}}[n,k] = \alpha P_{\text{orig}}[n,k] + (1 - \alpha) P_{\text{proc}}[n,k], 0 \leq \alpha \leq 1$$
- Spectral averaging is particularly helpful in the matched and multistyle training conditions, as well as when the data are filtered.
- The optimal value of $\alpha$ is sensitive to SNR, noise type, and training style.
- Spectral averaging can improve results up to an additional 15.5% relative (see figure below).

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