# 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 15213 USA
{mharvill, rms}@cs.cmu.edu

## 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. Before calculating prototype histograms over the training data, the power signals in each channel are normalized by the local maximum and minimum of the channel. In contrast, histograms calculated over the testing data are normalized by the global maximum and minimum of the power spectrum. This mode of normalization leads to a significant reduction in noise. Following the histogram-based processing, 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.

***Index Terms***— Robust speech recognition, histogram matching, spectral averaging, matched training, multistyle training

## 1. INTRODUCTION

With the continued development of large-vocabulary continuous speech recognition (LVCSR) systems based on hidden Markov modelling (HMM) technologies, automatic speech recognition (ASR) systems have become increasingly more accurate when trained and tested in noise-free environments. Unfortunately, ASR performance sharply deteriorates in the presence of noise or other types of signal degradation unless additional signal processing to enhance robustness is included.

Many algorithms have been developed to combat the effects of additive noise. Some of these algorithms are based on the estimation and subtraction of background noise (*e.g.* [1]), some mimic attributes of the human auditory system to enhance the properties of speech believed to be relevant to recognition (*e.g.* [2]), and still others use models of the environment to characterize the combined effects of additive noise and linear filtering (*e.g.* [3]). Nevertheless, most of these algorithms were developed under the presumption that the ASR system would be trained with clean speech, only to be exposed to noisy data in the testing stage. In recent years it has been observed that it can be advantageous to train very large systems using raw noisy data. Such *multistyle training* is becoming more common, necessitating the development of robustness algorithms that maintain good performance when trained in this fashion.

This paper describes a new noise-compensation approach called *Compensatory Spectral Averaging and Warping using Histograms* (*CSAWH,* pronounced "see-saw"), which is a histogram-based method that operates on the power of the speech signal in parallel subbands, warped according to the equivalent rectangular bandwidth (ERB) scale. This transformation is followed by a weighted spectral averaging of the processed and unprocessed power spectra, which tempers the effects of noise reduction imposed by the histogram-based transformation. In general, CSAWH processing proves to be particularly beneficial in matched and multistyle training conditions.

## 2. OVERVIEW OF HISTOGRAM-BASED METHODS

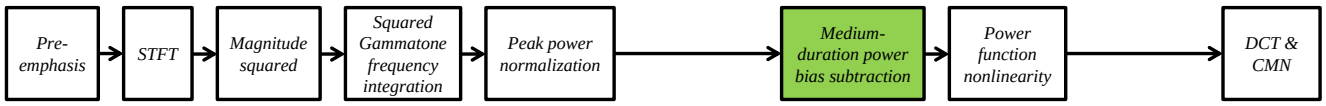
Balchandran and Mammone [4] originally proposed the use of histogram matching to invert the effects of nonlinear channel distortion effects in speaker identification systems. The first direct application of histogram matching for ASR was proposed by Dharanipragada and Padmanabhan [5] as a computationally inexpensive form of unsupervised speaker adaptation comparable in performance to maximum likelihood linear regression (MLLR). Using mel-frequency cepstral coefficients (MFCC), Dharanipragada and Padmanabhan applied histogram matching at the feature level with reference to histograms of the features obtained over clean training data.

In the following years, de la Torre *et al.* [6] and many others explored the application of histogram-based methods to *robust* speech recognition. Hilger's work [7] focused on a parametric implementation, *quantile-based histogram equalization*, using a power-law function applied at the output of the conventional mel-spaced filter bank in MFCCs. In contrast, de la Torre *et al.* considered the utility of histogram equalization applied to the cepstral features, underscoring the ability of the histogram equalization process to undo the nonlinear effects of additive noise in the cepstral domain. CSAWH modifies and extends these approaches by changing the way in which the histograms are computed and normalized, and through the inclusion of the subsequent weighted spectral averaging.

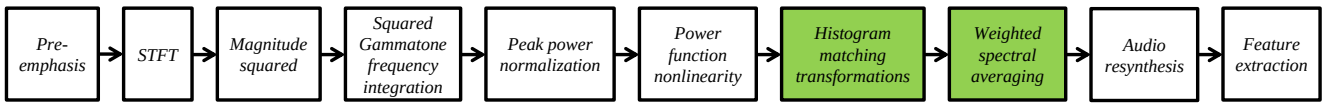
## 3. MOTIVATION FOR HISTOGRAM-BASED TRANSFORMATIONS

In previous work, Kim [2] described *power-normalized cepstral coefficients* (PNCC), which are more robust to noise and reverberation than MFCC and perceptual linear prediction (PLP) features. PNCC

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. We thank Chanwoo Kim for many useful discussions.




(a) PNCC




(b) CSAWH

**Fig. 1**. The major functional blocks of PNCC and CSAWH; shaded blocks indicate structures associated with noise reduction. Peak power normalization refers to normalizing the spectrogram by the global $98^{th}$ percentile.

feature extraction is generally similar in nature to MFCC feature extraction, differing primarily in the use of a Gammatone filter bank instead of the triangular filter bank, a power-law nonlinearity instead of the logarithmic nonlinearity, and the important addition of a stage of "medium-time" noise compensation. Figure 1 compares the processing used in PNCC features and CSAWH features. As can be seen, much of CSAWH processing is similar to PNCC.

CSAWH was motivated by the observation that the output of PNCC processing without the shaded medium-time processing blocks (called S-PNCC by Kim) still retains a residue of the local contrast with respect to frequency and time that was observed in the clean speech. This suggests that the effects of noise can be at least partially inverted by applying a nonlinear warping function to the power values in each frequency band, so that the relative energies of the noisy signal are more closely matched to those of clean speech.

## 4. DESCRIPTION OF CSAWH PROCESSING

The desired nonlinear warping can be imposed automatically using nonparametric histogram matching. We perform histogram matching as described in [8]. Briefly, given an input cumulative distribution function (CDF) $C_X$ for random variable $X$ and a desired CDF $C_Y$, $X$ can be transformed to $Y$ using the function

$$f(x) = C_Y^{-1}(C_X(x)) \tag{1}$$

In our work, the prototype CDFs are effectively scaled so that their corresponding histograms have no null bins at the edges. On the other hand, the histograms representing the input signal levels may have null bins at their edges, which enables substantial noise reduction.

### 4.1. Initial processing

Figure 1(b) exhibits the processing flow of CSAWH. The input signal is passed through a high-pass pre-emphasis filter and the short-time Fourier transform (STFT) $X[n, m]$ is calculated using a window length of 51.2 ms at 78.125 frames per second. We have empirically determined that a window length of 51.2 ms yields best results, which is consistent with the observation that windows of duration of 50 – 150 ms are helpful in reducing the variance of estimates of the noise. As noted in [2], this "medium-time analysis" provides a more accurate characterization of the statistical attributes of noise, which vary more slowly than the characteristics of speech.

After calculation of the STFT, subband power signals $P[n, k]$ are generated by summing the squared magnitude of the STFT in each time frame, weighted by the squared magnitude response of each of 40 Gammatone filters. After applying peak power normalization as described in [2], a power-function nonlinearity of the form $P_{orig}[n, k] = P[n, k]^{1/15}$ is applied. Though similar to the conventional log nonlinearity, some researchers have found that a power-law nonlinearity provides greater recognition accuracy (*e.g.* [7, 9]).

### 4.2. Histogram-based power warping

Forty 100-bin histograms are obtained for the clean training data, representing the output of each of the Gammatone filter bank channels. As seen in Fig. 1, the histograms from the clean speech are calculated immediately after the power-function nonlinearity. Each histogram is obtained by dividing the range of the data in the channel into 100 bins of uniform width. Consequently, the true widths and bin centers generally vary from channel to channel. Since the bin values for each channel span the range of the channel, there are no null histogram bins at the edges.

For each test utterance, 100 equally-spaced bin centers are used for the calculation of the input histograms. These bin centers are calculated by dividing the *global* range of the spectrum over all Gammatone channels. For a given frequency band, the input histogram will generally have null bins at its edges. The matching of the test histogram to the prototype histogram is highly nonlinear because of this mismatch in ranges. The power warping moves the channel-specific maximum and minimum values from the test utterance toward the global maximum and minimum, and redistributes the values in between accordingly, which tends to reduce the impact of the noise.

### 4.3. Weighted spectral averaging

In some cases, the nonlinear warping imposed by the histogram matching causes noisy segments of the signal to be amplified rather than suppressed. This most typically occurs when processing noisy filtered speech. Due to the nature of the processing, a filtered channel will naturally be amplified. Similarly, the histogram matching may also partially suppress low-energy regions of speech. We have found that one simple but effective way to balance the potential contributions of the unprocessed and undistorted (but noisy) signal versus the processed and de-noised (but possibly distorted) sig-

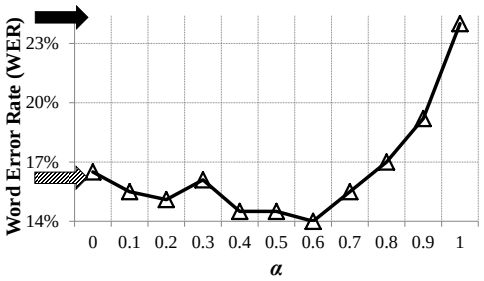



**Fig. 2**. CSAWH performance for different values of $\alpha$ in 5-dB white noise on RM1 and CMU Sphinx-3 under multistyle training. The solid arrow indicates baseline MFCC performance, the striped arrow indicates PNCC. Here, the optimal $\alpha = 0.6$ provides 15.5% relative improvement over the performance at $\alpha = 0$, where no spectral averaging is performed.

nal is to compute a spectrally-weighted linear combination of the processed and unprocessed spectrograms.

*Post-process weighted spectral averaging* refers to the weighted linear combination of the original power spectrum, $P_{orig}$, with the processed power spectrum, $P_{proc}$:

$$P_{out}[n,k] = \alpha P_{orig}[n,k] + (1-\alpha)P_{proc}[n,k], 0 \leq \alpha \leq 1 \quad (2)$$

The weighting parameter $\alpha$ is currently determined empirically for a given type of noise and training style, as will be discussed in Sec. 5.1.

### 4.4. Audio resynthesis and feature extraction

Although CSAWH operates on an STFT-based parametric representation, we have found that better recognition accuracy is typically obtained by resynthesizing an enhanced waveform and then computing conventional cepstral-based features, rather than deriving cepstral parameters directly without the waveform resynthesis [10]. The direct implementation of resynthesis enables CSAWH to be coupled with any conventional feature extraction algorithm. Audio resynthesis is accomplished using the overlap-add (OLA) algorithm, as described in Sec. 3 of [10].

## 5. EXPERIMENTAL RESULTS AND CONCLUSIONS

This section describes the recognition accuracy obtained using CSAWH for speech in various types of noise and using different training styles. We made use of CMU Sphinx-3 [11], and the DARPA Resource Management (RM1) database for demonstrative optimization tests.[1] SRI DECIPHER, as described in [12], and the Wall Street Journal (WSJ) task were used for evaluation tests.[2]

[1]A bigram language model (LM) and three-state HMM-based acoustic model (AM) with eight Gaussian mixtures per GMM is used on a subset of the RM1 database containing 1600 training utterances and 600 test utterances; first-pass results are reported.

[2]A bigram LM and three-state HMM-based AM with 32 Gaussians per GMM is trained with a subset of WSJ1 and tested on WSJ eval94 comprising 35,990 and 424 utterances, respectively; third-pass results after MLLR and trigram rescoring are reported.

| training style | rel. improvement over MFCC (%) | rel. improvement over PNCC (%) |
|---|---|---|
| clean | 48.29 | 0.95 |
| multistyle | 42.83 | 13.39 |
| matched | 26.10 | 15.73 |

**Table 1**. Approximate best-case relative improvements in average word error rate (WER) using CSAWH compared to MFCC and PNCC baselines in AWGN for the RM1 corpus using the CMU Sphinx-3 recognizer. The best value of $\alpha$ was manually selected for each SNR and training style.

### 5.1. Optimizing the weighting parameter $\alpha$

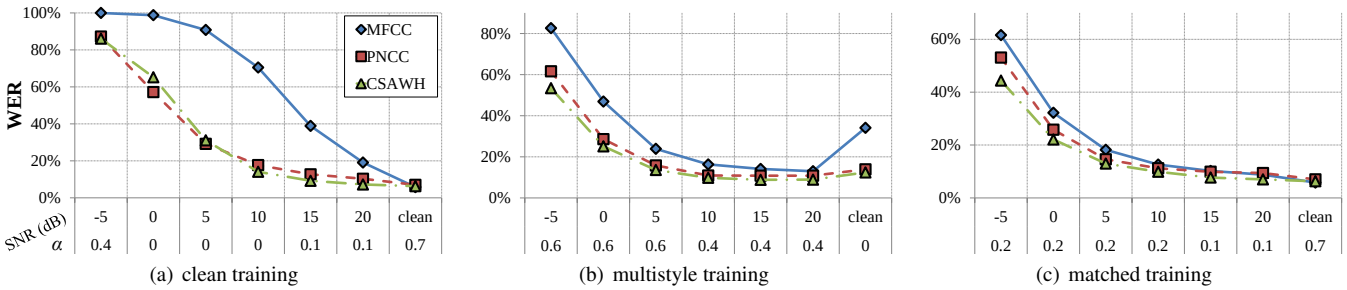
The optimal value of the weighting parameter $\alpha$ in Eq. (2) varies depending on the training style, the signal-to-noise ratio (SNR), and the noise type. This section illustrates the best-case performance of CSAWH in additive white Gaussian noise (AWGN), given the empirically-determined optimal value of $\alpha$ for each SNR and training style. To obtain these optimal values of $\alpha$, we ran Sphinx-3 on RM1 corrupted with white noise at seven different SNRs ($-5, 0, 5, 10, 15, 20$, and $\infty$ dB). Results were obtained for values of $\alpha$ between 0 and 1 inclusive, incrementing by 0.1 for each of the training conditions: clean, multistyle, and matched. The training data were partitioned such that the same number of utterances would be used to train under multistyle and clean conditions, with utterances from each speaker approximately uniformly distributed over the different SNRs. The results for 5-dB SNR in multistyle training are shown in Fig. 2. These results suggest that, given an accurate mechanism with which to determine $\alpha$, post-process spectral averaging will give rise to significant improvements in multistyle and matched training ASR.

### 5.2. Experiments using SRI DECIPHER

As noted in Section 4.3, it would be impractical to select the best $\alpha$ manually for each SNR, noise type, training condition, etc., as was done in Section 5.1. Without a method for blindly selecting nearly-optimal values of $\alpha$, which is the focus of our current work, $\alpha$ must be fixed over all training styles and SNRs. This section shows that CSAWH is still useful when $\alpha$ is fixed over all training styles to a value that is best overall, but not optimized for any *particular* noise type, training condition, or SNR.

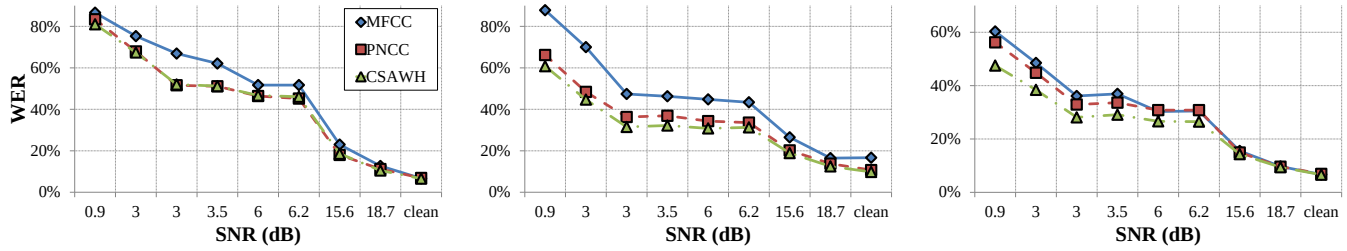
The noise samples used in these tests were developed to approximate eight highly-degraded communications channels that are representative of data from the DARPA RATS program (this task is more realistic than white noise). These channels were modelled as clean speech degraded by a time-invariant linear filter and subjected to arbitrary additive noise; the `renoiser` tool for MATLAB [13] was used to estimate the filter and noise parameters. The frequency response of the channel filter is typically high-pass, and the additive noise is generally not stationary. Since the test SNRs reflect direct estimates of the SNRs of each of the eight RATS voice channels, they are not uniformly spaced and are sometimes repeated.

The results in Figure 4 and Table 2 indicate that the algorithm is inherently useful, even without an optimal method of blindly deriving $\alpha$ in real-time. Moreover, the results of Secs. 5.1 and 5.2, taken together, are encouraging as they confirm the algorithm's portability across distinct ASR systems, and its effectiveness in the face of different types of noise, filtering, and system training.




(a) clean training (b) multistyle training (c) matched training

**Fig. 3**. ASR results using the CMU Sphinx-3 recognizer for the RM1 corpus in white noise; the top row of the horizontal axis labels lists the SNR in dB, the bottom row indicates the optimal value of $\alpha$ used for that SNR.




(a) clean training (b) multistyle training (c) matched training

**Fig. 4**. ASR results using SRI DECIPHER for the WSJ corpus in RATS-like noise; each case reflects CSAWH results with fixed $\alpha = 0.85$.

| training style | rel. improvement over MFCC (%) | rel. improvement over PNCC (%) |
|---|---|---|
| clean | 12.96 | 0.42 |
| multistyle | 31.77 | 9.29 |
| matched | 17.49 | 13.05 |

**Table 2**. Relative average WER improvement using CSAWH over comparable baselines in RATS-like noise for the WSJ corpus using SRI DECIPHER. A pre-selected, fixed value of $\alpha = 0.85$ is used for all training conditions, distortion channels, and SNRs.

## 6. 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] R. Balchandran and R.J. Mammone, "Non-parametric estimation and correction of non-linear distortion in speech systems," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, May 1998, vol. 2, pp. 749–752.

[5] 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.

[6] 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.

[7] F. Hilger, *Quantile based histogram equalization for noise robust speech recognition*, Ph.D. thesis, Computer Science Department, RWTH Aachen University, Aachen, Germany, 2004.

[8] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, pp. 128–138, Pearson Prentice Hall, Upper Saddle Ridge, New Jersey, third edition, 2008.

[9] C. Kim, *Signal processing for robust speech recognition motivated by auditory processing*, Ph.D. thesis, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2010.

[10] C. Kim, K. Kumar, and R.M. Stern, "Robust speech recognition using a small power boosting algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop*, December 2009.

[11] CMU Sphinx Speech Consortium, "CMU sphinx open source toolkit for speech recognition," http://cmusphinx.sourceforge.net/wiki/download/.

[12] H. Murveit *et al*., "Large-vocabulary dictation using SRI's DECIPHER speech recognition system: progressive search techniques," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, 1993, vol. 2, pp. 319–322.

[13] D. Ellis, "RENOISER - Utility to decompose and recompose noisy speech files," http://labrosa.ee.columbia.edu/projects/renoiser/.