An Improved Histogram Equalization Approach for Robust Speech Recognition

Shih-Hsiang Lin, Yao-Ming Yeh, Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University

2006/09/08

Outline

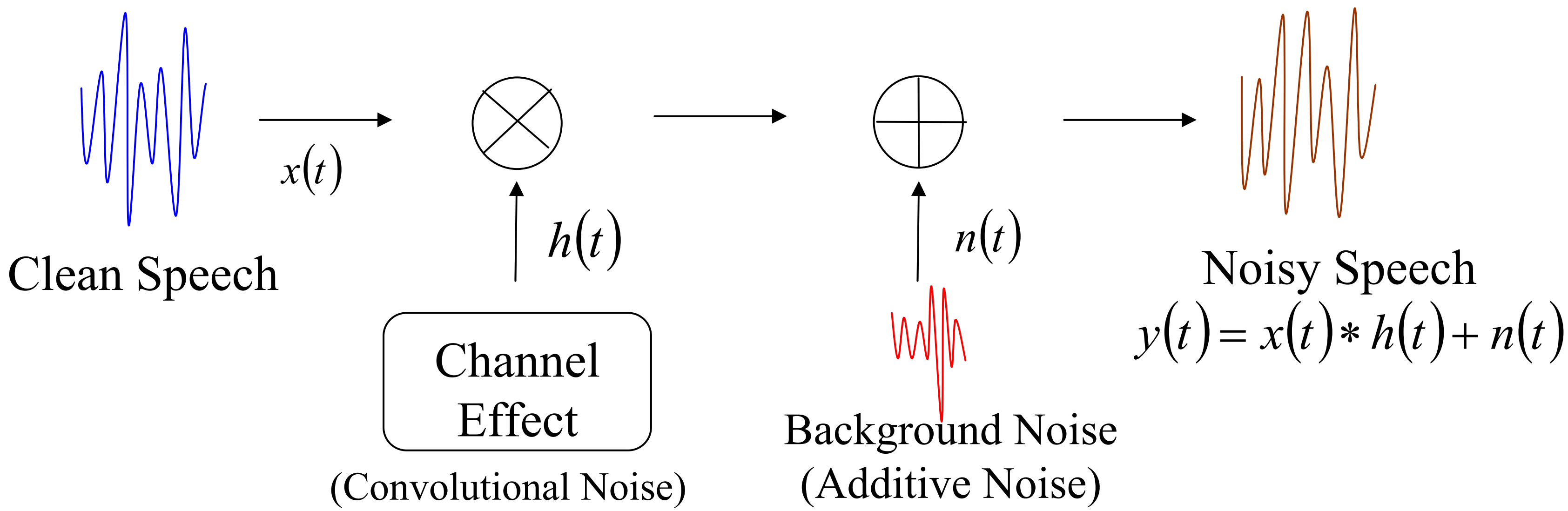
- Introduction
- Review of Conventional Histogram Equalization (HEQ) Approaches
- Proposed Polynomial-Fit Histogram Equalization (PHEQ) Approach
- Integration with Other Robustness Techniques
- Experimental Setup and Results
- Conclusions and Future Work

Introduction

- Varying environmental effects lead to severe mismatch between the acoustic conditions for the training and test speech data
  - Accordingly, performance of an automatic speech recognition (ASR) system would dramatically degrade
- Techniques dealing with this issue generally fall into three categories
  - Speech Enhancement
    - Spectral Subtraction (SS), Wiener Filter (WF), etc.
  - Robust Speech Feature or Feature Normalization
    - Cepstral Mean Subtraction (CMS), Cepstrum Mean and Variance Normalization (CMVN), etc.
  - Acoustic Model Adaptation
    - Maximum a Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR), etc.

Introduction (cont.)

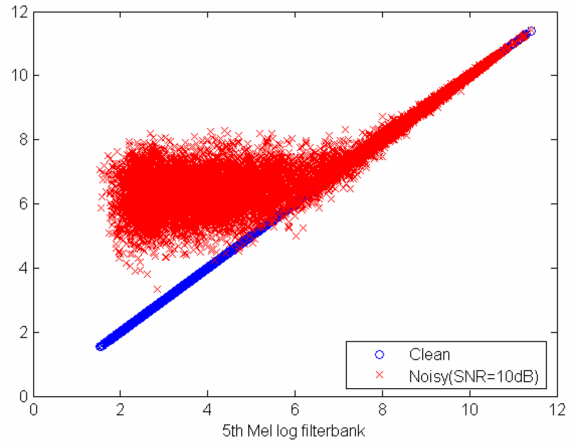
- A Simplified Distortion Framework

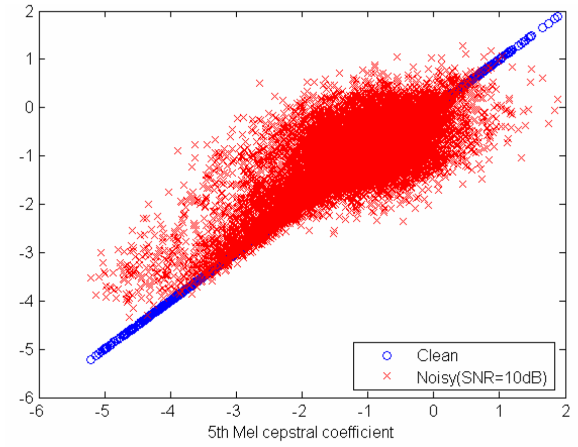


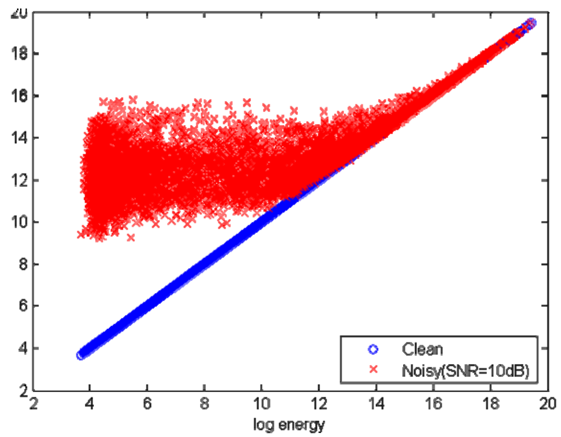
  - Channel effects are usually assumed to be constant while uttering
  - Additive noises can be either stationary or non-stationary

Introduction (cont.)

- Non-linear Environmental Distortions







  - Clean speech was corrupted by 10dB subway noise
    - Not only linear but also non-linear distortions were involved

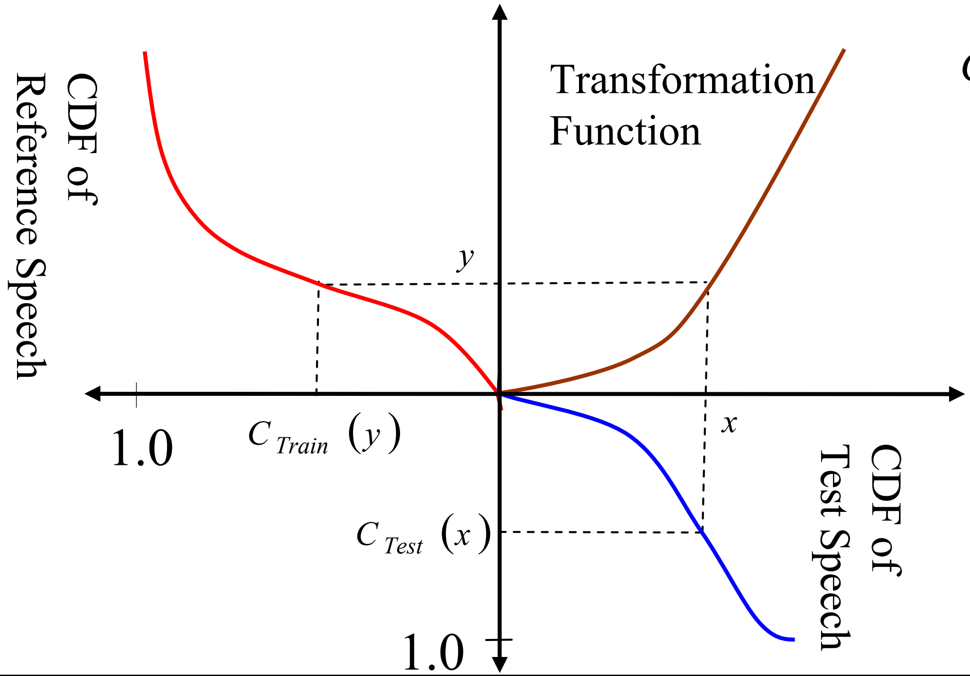
Introduction (cont.)

- Constraint of the linear property of conventional CMN and CMVN approaches
  - Linear distortions can be effectively dealt with
    - However, non-linear environmental distortions can not be adequately compensated

- Recently, histogram equalization (HEQ) approaches have been widely investigated for the compensation of non-linear environmental effects
  - HEQ attempts to not only match speech features' means/variances but also completely match the features' distributions of training and test data
  - Superior performance gains have been demonstrated

Roots of HEQ

- HEQ is a general non-parametric method to make the cumulative distribution function (CDF) of some given data match a reference one
  - E.g., the equalization of the CDF of test speech to that of training (reference) speech



$$
\begin{aligned}
C_{Test}(x) &= \int_{-\infty}^{x} p_{Test}(x')dx' \\
&= \int_{-\infty}^{F(x)} p_{Test}\left(F^{-1}(y')\right)\frac{dF^{-1}(y')}{dy'}dy' \\
&= \int_{-\infty}^{y} p_{Train}(y')dy'|_{y=F(x)} \\
&= C_{Train}(y)
\end{aligned}
$$

Practical Implementation of HEQ

- Due to a finite number of speech features being considered, the cumulative histograms are used instead of the CDFs

- HEQ can be simply implemented by table-lookup (THEQ)
  - e.g. {(Quantile *i*, Restored Feature Values)}
  - To achieve better performance, the table sizes cannot be too small
    - The needs of huge disk storage consumption
    - Table-lookup is also time-consuming

Quantile-based Histogram Equalization (QHEQ)

- QHEQ attempts to calibrate the CDF of each feature vector component of the test data to that of training data in a quantile-corrective manner
  - Instead of full matching of cumulative histograms
- A parametric transformation function is used

$$H(x) = Q_K \left(\alpha \left(\frac{x}{Q_K} \right)^{\gamma} + (1-\alpha) \left(\frac{x}{Q_K} \right) \right)$$

- For each sentence, the optimize parameters α and β should be obtained from the quantile correction step

$$\{\alpha, \gamma\} = \arg \min_{\{\alpha,\gamma\}} \left(\sum_{k=1}^{K-1} \left(H(Q_k) - Q_k^{train} \right)^2 \right)$$

  - Exhaustive online grid search is required: time-consuming

Polynomial-Fit Histogram Equalization (PHEQ)

- We propose to use least squares regression for the fitting of the inverse function of CDFs of training speech
  - For each speech feature vector dimension of the training data, a polynomial function can be expressed as follows, given a pair of y_i and corresponding CDF $C_{Train}(y_i)$

$$G\left(C_{Train}(y_i)\right) = \tilde{y}_i = \sum_{m=0}^{M} a_m \left(C_{Train}(y_i)\right)^m$$

  - The corresponding squares error

$$E'^2 = \sum_{i=1}^{N} \left(y_i - \sum_{m=0}^{M} a_m \left(C_{Train}(y_i)\right)^m \right)^2$$

  - Coefficients a_m can be estimated by minimizing the squares error

PHEQ (cont.)

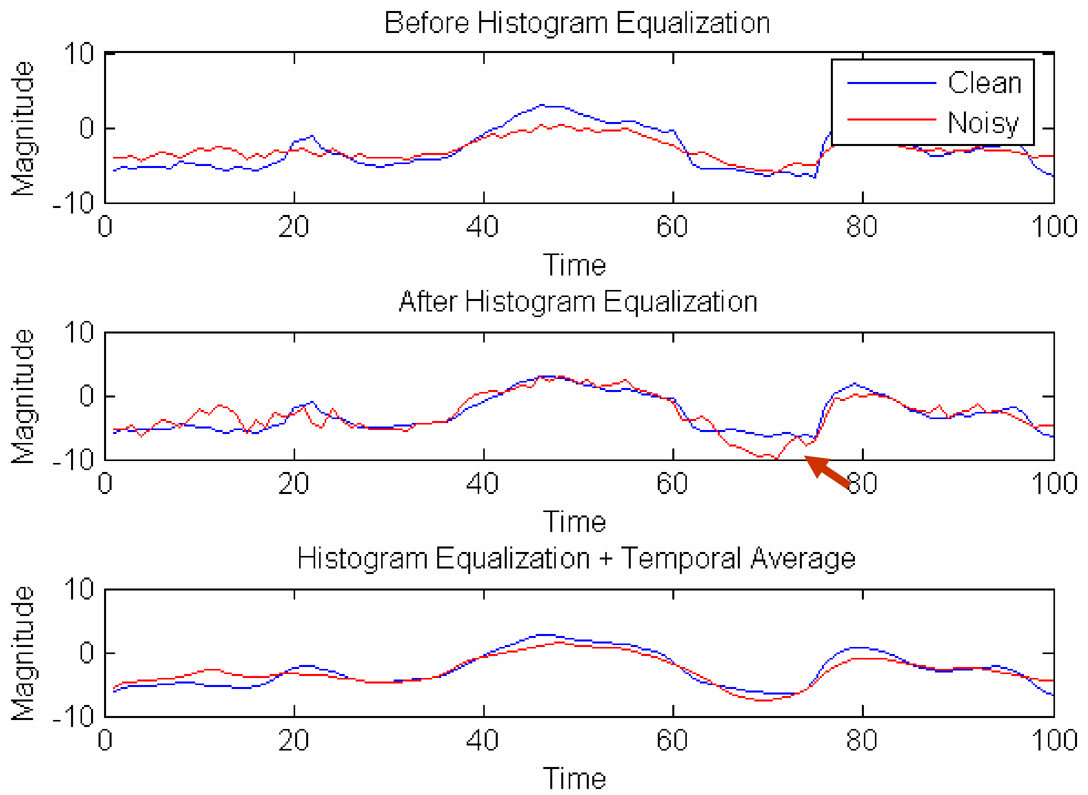
- Implementation details
  - For each feature vector dimension, $Y_{1,N} = [y_1, y_2, \ldots y_N]$, the CDF value of each frame can be estimated using the following steps
    - $Y_{1,N}$ are sorted in an ascending order
    - The corresponding CDF value of each frame is approximated by

$$C(y_i) \approx \frac{S_{pos}(y_i)}{N}$$

      - Where $S_{pos}(y_i)$ is an indication function, indicating the position of y_i in the sorted data
  - During recognition
    - The CDF value $C(y_i)$ of each test frame is estimated and taken as the input to the corresponding inverse function *G* to obtain a restored feature component

Polynomial-Fit Histogram Equalization (cont.)

- Though, as will be indicated, PHEQ are effective
  - Some undesired sharp peaks or valleys caused by non-stationary noises or occurred during equalization can not be well compensated by HEQ



Temporal Averaging (TA)

- Several approaches using the moving averages of temporal information were also investigated
  - Non-Casual Moving Average

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=-L}^{L} \tilde{y}_{(t+i)}}{2L+1} & if\ L < t \le T - L, \\ \tilde{y}_t & otherwise \end{cases}$$

  - Casual Moving Average

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=0}^{L} \tilde{y}_{(t-i)}}{L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases}$$

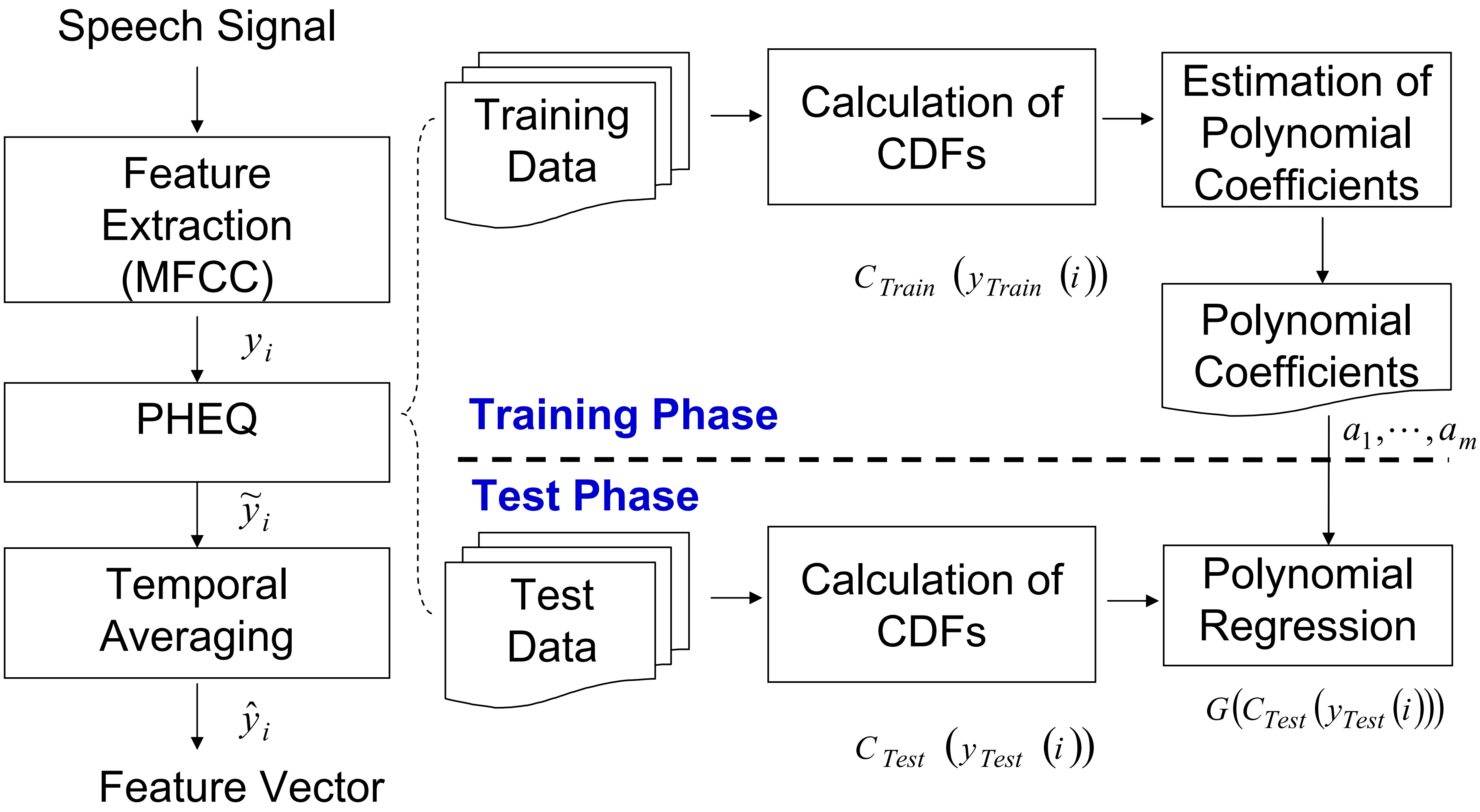
  - Non-Casual Auto Regression Moving Average

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t+j)}}{2L+1} & if\ L < t \le T - L, \\ \tilde{y}_t & otherwise \end{cases}$$

  - Casual Auto Regression Moving Average

$$\hat{y}_t = \begin{cases} \dfrac{\sum_{i=1}^{L} \hat{y}_{(t-i)} + \sum_{j=0}^{L} \tilde{y}_{(t-j)}}{2L+1} & if\ L < t \le T, \\ \tilde{y}_t & otherwise \end{cases}$$

Block Diagram of Proposed Approach



Experimental Setup

- The speech recognition experiments were conducted under various noise conditions using the Aurora-2 database and task
  - Front-end speech analysis
    - 39-dimensional feature vectors were extracted at each time frame, including 12 MFCCs + log Energy, and the corresponding delta and acceleration coefficients
  - Back-end recognizer
    - HTK recognition toolkit for training of acoustic models
    - Each digit acoustic model was a left-to-right continuous density HMM with 16 states (3 diagonal Gaussian mixtures per state)
    - Two additional silence models were defined
      - Short pause: 1 state (6 Gaussians)
      - Silence: 3 states (6 Gaussians per state)

Experimental Results: PHEQ

Word Error Rate (WER)		Polynomial Order			
		3	5	7	9
Clean Condition Training	All training data	22.39	21.54	21.08	21.30
	1000 quantiles	21.80	21.46	21.13	21.16
	100 quantiles	22.68	21.31	20.75	20.55
	10 quantiles	23.42	22.20	22.54	23.42
Multi Condition Training	All training data	10.80	10.34	10.43	10.54
	1000 quantiles	10.48	10.32	10.40	10.45
	100 quantiles	10.73	10.45	10.36	10.45
	10 quantiles	11.65	10.61	10.79	11.58

Average word error rates (WERs) w.r.t different numbers of training data and different polynomial orders which were used in the estimation of the inverse functions of CDFs

– WER is slightly improved when the order of the polynomial regression becomes higher
– 100 quantiles and 7-th polynomial order were used in the following experiments

Experimental Results: PHEQ-TA

Word Error Rate (WER)		Span Order					
		0	1	2	3	4	5
Clean Condition Training	Non-Casual MA	20.75	17.75	16.83	17.26	18.15	19.66
	Casual MA	20.75	19.23	18.28	17.44	17.12	17.28
	Non-Casual ARMA	20.75	17.83	16.90	16.38	16.99	17.34
	Casual ARMA	20.75	17.93	16.84	19.20	17.44	19.20
Multi Condition Training	Non-Casual MA	10.36	9.88	9.88	10.24	10.94	11.69
	Casual MA	10.36	10.13	9.74	9.76	9.78	10.12
	Non-Casual ARMA	10.36	9.88	9.78	9.84	9.94	10.11
	Casual ARMA	10.36	9.95	9.71	10.84	9.76	10.68

Average word error rates (WERs) w.r.t combine PHEQ with different temporal averaging techniques and different span orders

- Non-Casual ARMA can yield better performance
- In clean-condition training, it can provide a relative improvement of about 20% compared with that of using PHEQ alone

Experimental Results: PHEQ-TA

		Word Error Rate (WER)			
		Set A	Set B	Set C	Average
Clean Condition Training	MFCC	41.06	41.52	40.03	41.04
	AFE	38.69	44.25	28.76	38.93
	CMVN	27.73	24.60	27.17	26.37
	MS+VN+ARMA(3)	18.38	16.14	21.81	18.17
	THEQ	19.72	18.57	19.24	19.16
	QHEQ	23.53	21.90	22.36	22.64
	PHEQ	20.98	20.17	21.43	20.75
	PHEQ-TA	16.83	15.10	20.02	16.78

Average word error rates (WERs) of different feature normalization approaches

- PHEQ provides significant performance boosts for the baseline MFCC system
- It is also better than CMVN, and competitive to HEQ and QHEQ

Experimental Results: PHEQ-TA (cont.)

		Word Error Rate (WER)			
		Set A	Set B	Set C	Averag e
Multi Condition Training	MFCC	14.78	16.01	19.33	16.18
	AFE	10.64	10.76	12.85	11.13
	CMVN	12.70	12.45	14.52	12.98
	MS+VN+ARMA(3)	9.49	10.37	10.06	9.95
	THEQ	10.02	10.41	10.34	10.24
	QHEQ	10.20	10.75	10.76	10.53
	PHEQ	9.91	9.41	13.14	10.36
	PHEQ-TA	9.41	9.53	11.21	9.82

Average word error rates (WERs) of different feature normalization approaches

– In multi-condition training, PHEQ also provides consistently better results as that is done in clean-condition training

Integration with Other Robustness Techniques

- Finally, we integrated our proposed feature normalization approach with two conventional feature de-correlation and compensation techniques
  - **Heteroscedastic Linear Discriminant Analysis (HLDA) and Maximum Likelihood Linear Transform (MLLT)**
    - HLDA and MLLT were conducted directly on the Mel-frequency filter bank outputs
    - HLDA is used for dimension reduction and MLLT is used for feature decorrelation
  - **Stereo-based Piecewise LInear Compensation (SPLICE)**
    - The piecewise linearity is intended to approximate the true nonlinear relationship between clean and corresponding noisy utterances
    - Provide accurate estimates of the bias or correction vectors without the need for an explicit noise model
    - SPLICE is a frame-based bias removal algorithms

Integration with Other Robustness Techniques (cont.)

		Word Error Rate (WER)			
		Set A	Set B	Set C	Average
Clean Condition Training	HLDA-MLLT+CMVN	21.63	21.37	21.59	21.52
	HLDA-MLLT+PHEQ-TA	15.98	15.96	15.91	15.96
	SPLICE+CMVN	16.34	14.95	21.18	16.75
	SPLICE+PHEQ-TA	13.40	13.41	17.08	14.14
Multi Condition Training	HLDA-MLLT+CMVN	9.49	9.51	10.40	9.68
	HLDA-MLLT+PHEQ-TA	9.06	8.87	8.55	8.88
	SPLICE+CMVN	10.40	11.00	13.80	11.32
	SPLICE+PHEQ-TA	9.54	10.88	12.18	10.60

Average word error rates (WERs) achieved by combing different normalization and de-correlation approaches

– Either the feature de-corrleation technique, like HLDA-MLLT, or the feature compensation technique, like SPLICE, can achieve significant performance gains when being combined with PHEQ-TA

Conclusions and Future Work

- The HEQ approaches for feature normalization were extensively investigated and compared
  - Wd have proposed the use of data fitting schemes to efficiently approximate the inverse of the CDF of the training speech for HEQ
  - Further investigation of PHEQ is currently undertaken
- Different moving average methods were also exploited to alleviate the influence of sharp peaks and valleys
- The combinations with the other feature de-correlation and compensation techniques indeed demonstrated very encouraging results