Exploring the Use of Speech Features and Their Corresponding Distribution Characteristics for Robust Speech Recognition

Author : Shin-Hsiang Lin, Berlin Chen, Yao-Ming Yeh

Professor:陳嘉平
Reporter:葉佳璋

Outline

- Introduction
- Cluster-based polynomial-fit histogram(CPHEQ)
- Polynomial-fit histogram Equalization(PHEQ)
- Experiment

Introduction

- The performance of current automatic speech recognition systems often deteriorates radically when input speech is corrupted by noise.
- Broadly speaking, the existing methods can be classified into two categories
  - Directly on the basis of the feature domain
  - Consider some specific statistical feature characteristics

Introduction

- In this paper, we propose a cluster-based polynomial-fit histogram equalization(CPHEQ) approach.
- It can makes use of both the speech feature and corresponding distribution characteristics for speech compensation.

Cluster-Based Polynomial-Fit Histogram

- The basic idea behind CPHEQ stems from two diverse approaches
  - Stereo-based piecewise linear compensation for environments(SPLICE)
  - HEQ
- SPLICE attempts to use a GMM to characterize the noisy feature space, and each Gaussian component represents one specific distortion condition.

Cluster-Based Polynomial-Fit Histogram

- Theoretically, SPLICE can deal with either linear or nonlinear distortions when the number of Gaussian components increase to infinite.
- In order to avoid this shortcoming, we add the idea of HEQ.

Cluster-Based Polynomial-Fit Histogram

- HEQ uses nonlinear transformation functions to compensation nonlinear distortions.
- We propose the use of CPHEQ, which combines the merits of both SPLICE and HEQ as well as overcomes their individual defects.

Cluster-Based Polynomial-Fit Histogram

- For CPHEQ, we first use the speech data to train a GMM model.
- The GMM is expressed as follows:

$$p\left(Y_t\right) = \sum_{k=1}^{K} p\left(k\right) p\left(Y_t \mid k\right) = \sum_{k=1}^{K} p\left(k\right) N\left(Y_t; \mu_k, \Sigma_k\right)$$

  - K :the mixture number of GMM
  - Y_t :noisy feature vector
  - μ_k, Σ_k :mean vector and diagonal covariance matrix

Cluster-Based Polynomial-Fit Histogram

- We assume that the compensated feature vector $\tilde{X}_t$ can be derived by

$$\tilde{X}_t = E\left[X_t \mid Y_t\right] = E\left[E\left[X_t \mid Y_t, k\right]\right] = \sum_{k=1}^{K} p\left(k \mid Y_t\right) E\left[X_t \mid Y_t, k\right]$$

- $p\left(k \mid Y_t\right)$ is the posterior probability given by

$$p\left(k \mid Y_t\right) = \frac{p\left(Y_t \mid k\right) p\left(k\right)}{\sum_{k'=1}^{K} p\left(Y_t \mid k'\right) p\left(k'\right)}$$

Cluster-Based Polynomial-Fit Histogram

- In decorrelated feature domain, it is reasonable to assume that the feature vector components y_t of Y_t are independent of each other.
- The restored value of y_t given the kth mixture is defined as follows:

$$\tilde{x}_{t,k} = E\left[x_t \mid Y_t, k\right] \approx E\left[x_t \mid y_t, k\right]$$

Cluster-Based Polynomial-Fit Histogram

- We introduce the idea originating from HEQ to approximate the conditional expectation.

$$\tilde{x}_{t,k} \approx G_k\left(CDF\left(y_t\right)\right)$$

  - $CDF\left(y_t\right)$ is the CDF value of the feature component y_t .
  - $G_k(\bullet)$ is the inverse function, which map each CDF value onto its corresponding predefined feature value.

Cluster-Based Polynomial-Fit Histogram

- For the feature vector component sequence $y_1, \cdots, y_t, \cdots, y_L$ of a specific dimension can be computed through the follow two steps:
  - Step1: The sequence is first sorted in ascending order according to the feature vector components.
  - Step2: The CDF value of a feature vector component is then given as

$$CDF(y_t) \approx \frac{S_{pos}(y_t) - 0.5}{L}$$

Where $S_{pos}(y_t)$ is a function that return the rank of y_t in ascending order of $y_1, \cdots, y_t, \cdots, y_L$.

Cluster-Based Polynomial-Fit Histogram

- We use a polynomial function to approximate the inverse function of CDF for efficiency and storage requirement.
- In the training phase, the coefficients a_{km} of the polynomial function $G_k\left(CDF\left(y_t\right)\right)=\sum_{m=0}^{M} a_{km}\left(CDF\left(y_t\right)\right)^m$ can be estimated by MMSE defined by

$$E_k^2=\sum_{t=1}^{T-1}\left(p\left(k\mid Y_t\right)\times\left(x_t-\sum_{m=0}^{M} a_{km}\left(CDF\left(y_t\right)\right)^m\right)\right)^2$$

  - T:the total number of frame in training data
  - y_t, x_t :the feature vector component for noisy speech and its corresponding clean counterpart.

Cluster-Based Polynomial-Fit Histogram

- In the test phase, the restored value of can be obtained by

$$\tilde{x}_t = \sum_{k=1}^{K} \left(p\left(k \mid Y_t\right) \times \left(\sum_{m=0}^{M} a_{km} \left(CDF\left(y_t\right)\right)^m \right) \right)$$

- In order to reduce the computation time, we use the maximum a posterior probability (MAP) criterion

$$E_k^2 = \sum_{t=1}^{T-1} \left(\delta\left(k \mid Y_t\right) \times \left(x_t - \sum_{m=0}^{M} a_{km} \left(CDF\left(y_t\right)\right)^m \right) \right)^2$$

$$\tilde{x}_t = \sum_{k=1}^{K} \left(\delta\left(k \mid Y_t\right) \times \left(\sum_{m=0}^{M} a_{km} \left(CDF\left(y_t\right)\right)^m \right) \right)$$

$$\delta\left(k \mid Y_t\right) = \begin{cases} 1, & if \quad k = \arg\max_{k'} p\left(k' \mid Y_t\right) \\ 0, & otherwise. \end{cases}$$

Polynomial-Fit Histogram Equalization

- We present a variant of CPHEQ, named polynomial histogram equalization (PEQ).
- Only a single global transformation is utilized to obtain the restored value $\tilde{x}_t$ of the noisy feature vector component y_t.

$$\tilde{x}_t = G_k\left(CDF\left(y_t\right)\right) = \sum_{m=0}^{M} a_{km}\left(CDF\left(y_t\right)\right)^m$$

$$E_k^2 = \sum_{t=1}^{T-1}\left(x_t - \sum_{m=0}^{M} a_{km}\left(CDF\left(x_t\right)\right)^m\right)^2$$

Experiments and Results

- Aurora-2 database and task.
  - Variable length continuous English digital string spoken into a close-talking microphone.
  - Eight different types of real-word additive noises are artificially added into Test Set A and Test Set B.
  - Two types of training scenarios: clean training and multi-condition training.

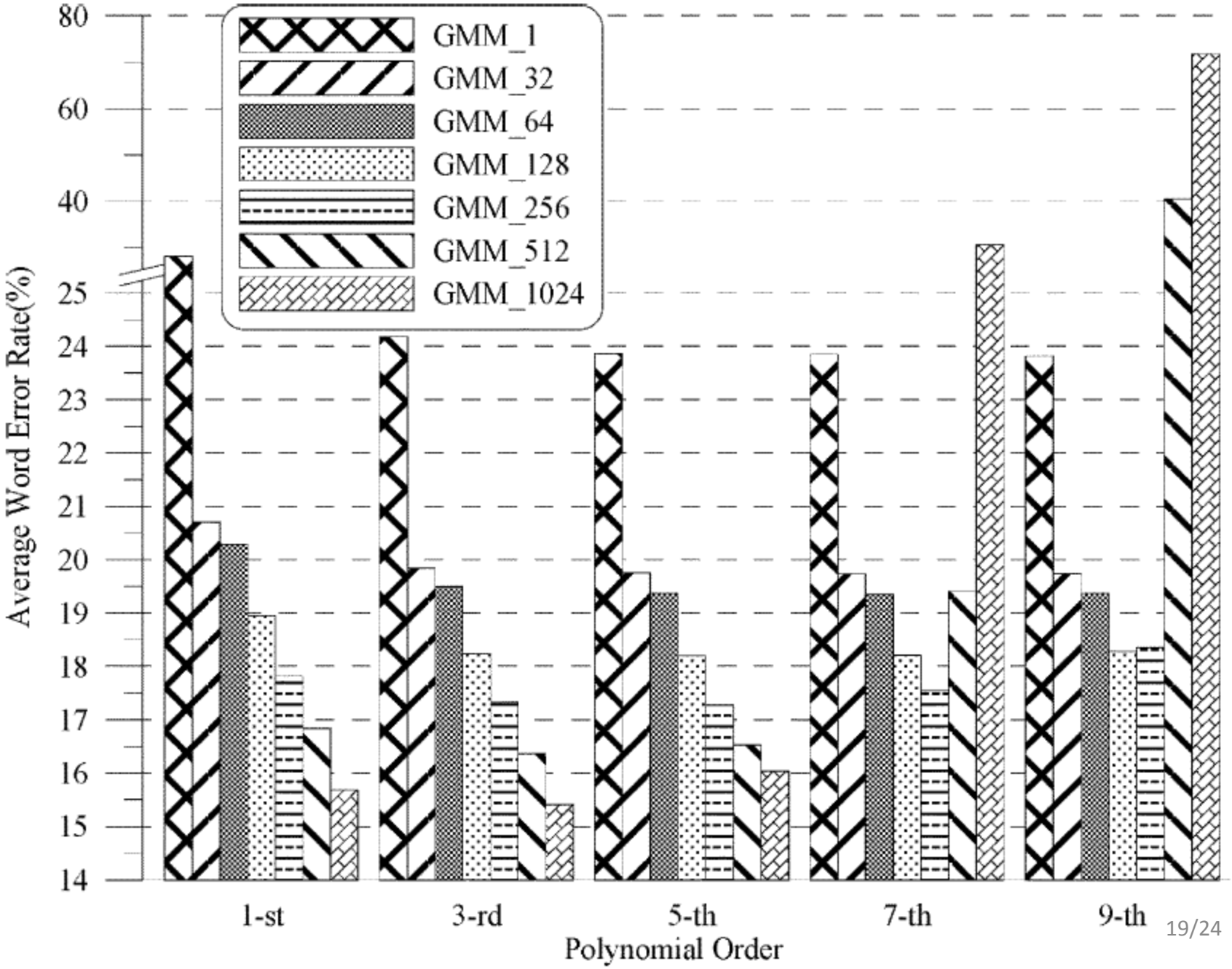
Experiment on CPHEQ

- We first evaluate the performance of CPHEQ when different criteria are used to obtain the polynomial.
- Multi-condition training set.
- Noise type identical to set A with an SNR range of 5-20dB.
- The number of GMM set 32 to 1024 and the order of polynomial set three.

Experiment on CPHQ

COMPARISON OF THE AVERAGE WER RESULTS (%) OF THE HARD- AND SOFT-DECISION APPROACHES USED FOR DERIVING THE POLYNOMIAL FUNCTIONS OF CPHEQ

	Number of Mixtures					
	32	64	128	256	512	1024
Hard	19.84	19.49	18.24	17.33	16.36	15.41
Soft	19.88	19.46	18.23	17.31	16.33	15.40


GMM_1
GMM_32
GMM_64
GMM_128
GMM_256
GMM_512
GMM_1024
Average Word Error Rate(%)
80
60
40
25
24
23
22
21
20
19
18
17
16
15
14
1-st
3-rd
5-th
7-th
9-th
Polynomial Order

Experiment on CPHQ

- The performance tends to degrade substantially when the order becomes too large.
- These phenomena may be explained by the reason that only **a limited set of training data was used in this study**(the fact of the curse of dimensionality).
- And the use of higher-order polynomial function might have **led to oscillations between the exact-fit value**.

Experiments on PHEQ

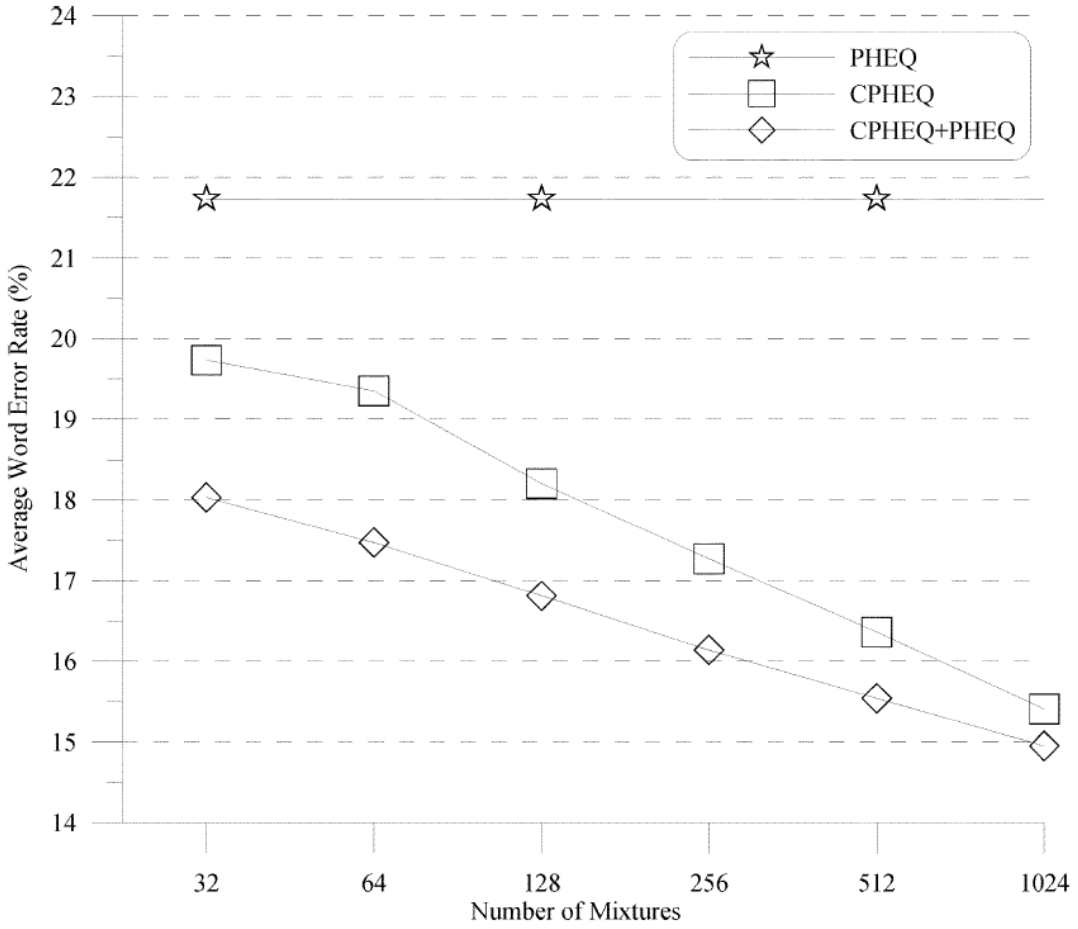
AVERAGE WER RESULTS (%) OBTAINED WITH RESPECT TO DIFFERENT POLYNOMIAL ORDERS USED IN THE ESTIMATION OF THE TRANSFORMATION FUNCTIONS OF PHEQ

	Number of Polynomial Orders				
	1-st	3-rd	5-th	7-th	9-th
PHEQ	23.25	21.80	21.46	21.13	21.16

Experiments on PHEQ

- A smaller mixture number may be insufficient to delineate the noise characteristic.
- We try to combine CPHEQ with PHEQ through a simple linear interpolation of these two methods, to overcome this shortcoming.
- The interpolation weights were preliminarily set equal 0.5 and fixed during the test phase.

Experiments on PHEQ



	Test Set A	Test Set B	Test Set C	Average
MFCC	14.78	16.01	19.33	16.18
AFE	7.03	7.95	8.27	7.65
SPLICE	11.03	11.47	16.86	12.17
PHEQ	9.91	9.41	13.14	10.36
CPHEQ	10.29	9.81	12.04	10.44
CMS+CPHEQ	8.49	9.22	10.80	9.24
AFE+CPHEQ	7.24	8.06	7.87	7.69