Local Projection and Support Vector Based Feature Selection in Speech Recognition

Author : Antonio Miguel
Alfonso Ortega
Professor: 陳嘉平
Reporter : 許峰閤

Outline

- Introduction
- Feature Extraction
- Local Projection
- Support Vector
- Experiments

Introduction

- This is a method that provide robustness to mismatch conditions by using local time-frequency projection and feature selection.

- The support vector provides the most representative example which have influence in the error rate in mismatch condition.

Feature Extraction

- Inorder to obtain robust feature vectors, dynamic features are usually used.

First: $$c_t = (W_F)^T o_t$$

o_t the log-filterbank output vector size B x 1

c_t the C component cepstrum vectors

t is time index

Feature Extraction

- The dynamic features can be obtain in matrix form:

$$X_t = (W_F)^T O_t W_T$$

$$O_t = (o_{t-\frac{L}{2}}, \ldots, o_{t+\frac{L}{2}})$$

- L is the sliding widow length
- W_T is the time projection matrix L x S
- S is the dynamic stream

- The notation used to describe the feature extraction is discussed both temporal and frequency projection in a compact way.

- It can be expressed in a 2D mask:

$$(X)_{s,c} = \sum_{b=1}^{B} (W_F)_{b,c} \sum_{l=1}^{L} (O)_{b,l} (W_T)_{l,s}$$

$$= \sum_{b,l} (W_F)_{b,c} (W_T)_{l,s} (O)_{b,l} = \sum_{b,l} (W_{2D}^{s,c})_{b,l} (O)_{b,l}$$

c is the ceptrum index

S is the dynamic stream index

Feature Extraction

- This approach is called DCT2, has two benefits in terms of pattern recognition:

1. The classifier can be more simple.

2. Helps to reduce the variability due to small scale acoustic events.

Local Projections

- Some alternatives to the DCT transform can reduce the impact of narrow-band noise, like the frequency projection.

- The local projection can be define by concatenation in a feature vector of a number of partial subband DCT compression.

Feature Selection

- When the number of features keeps growing, there is a point where the accuracy starts to decrease.

- Compute the mutual information with respect to an informative variable like the component in the mixture in the trainning set.

Feature Selection

- The last would be to decide a vector size and select the informative feature based on the mutual information metric.

- This method is based for support vectors which reduced the WER.

Support Vectors

- The support vectors will be used to compute the mutual information metric.

$$\widehat{W} = \underset{W}{\arg\max}\, P(X \mid W)P(W) = \underset{W}{\arg\max}\, F(X \mid W)$$

- $F(X \mid W)$ is called discriminant function
- $x_t \in X$ set of feature vectors in development set
- $w_i \in W$ set of transcription

Support Vectors

$$d(X_i) = F(X_i \mid w_i) - \max_{\hat{w}_j \neq w_i} F(X_i \mid \hat{w}_j)$$

$$= \min_{\hat{w}_j \neq w_i} [F(X_i \mid w_i) - F(X_i \mid \hat{w}_j)]$$

$\hat{w}_j$ can be all the units, but we use the n-best output

If $d(X_i)$ <0 , the subsequence is not correctly recognition

Support Vectors

- Define the support vector set as

$$S = \{X_i \mid X_i \in X \;\; and \;\; \sigma_1 \geq d(X_i) \geq \sigma_2\}$$

- We can use these data to calculate the information

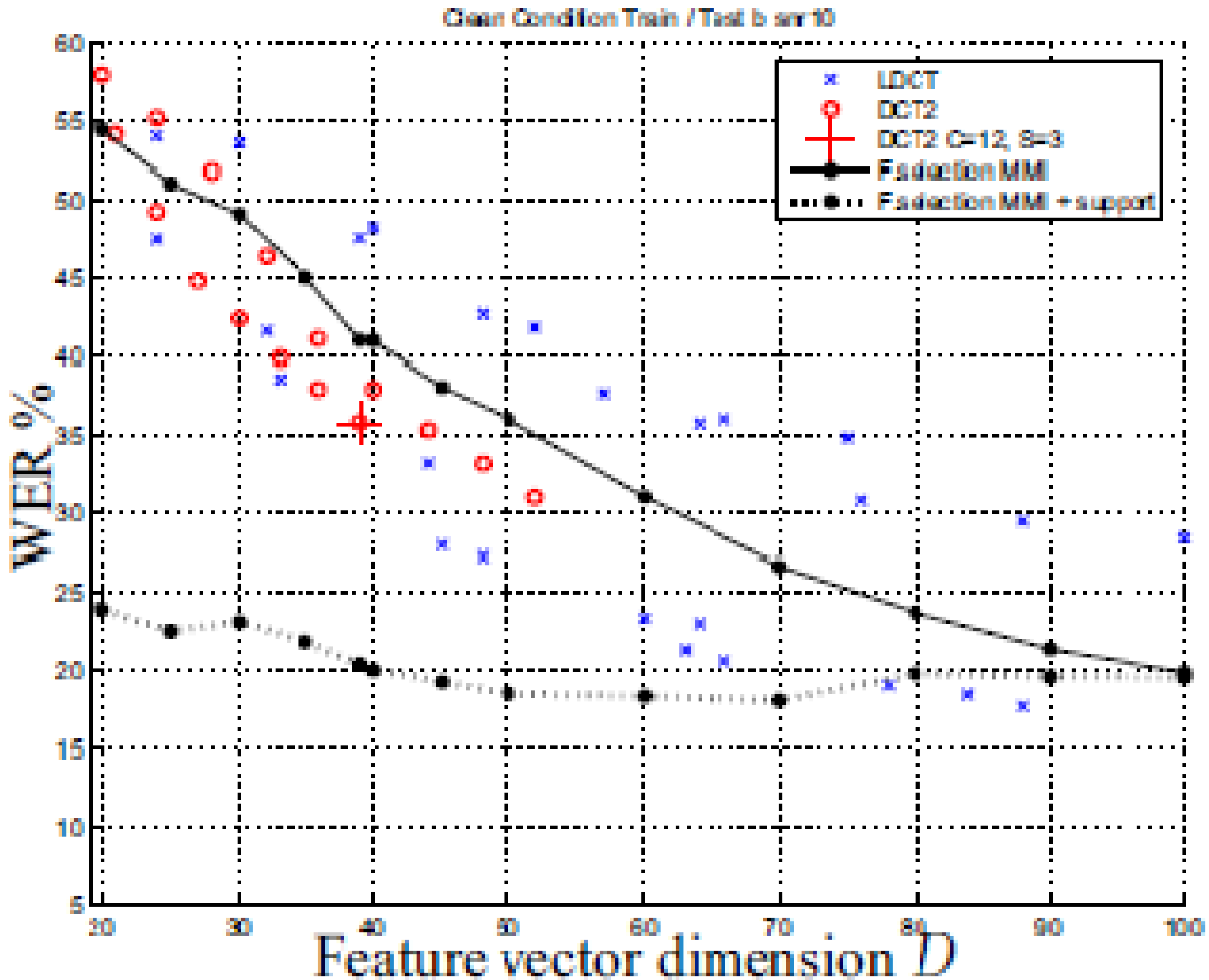
$$\hat{I}(X, Z \mid S)$$

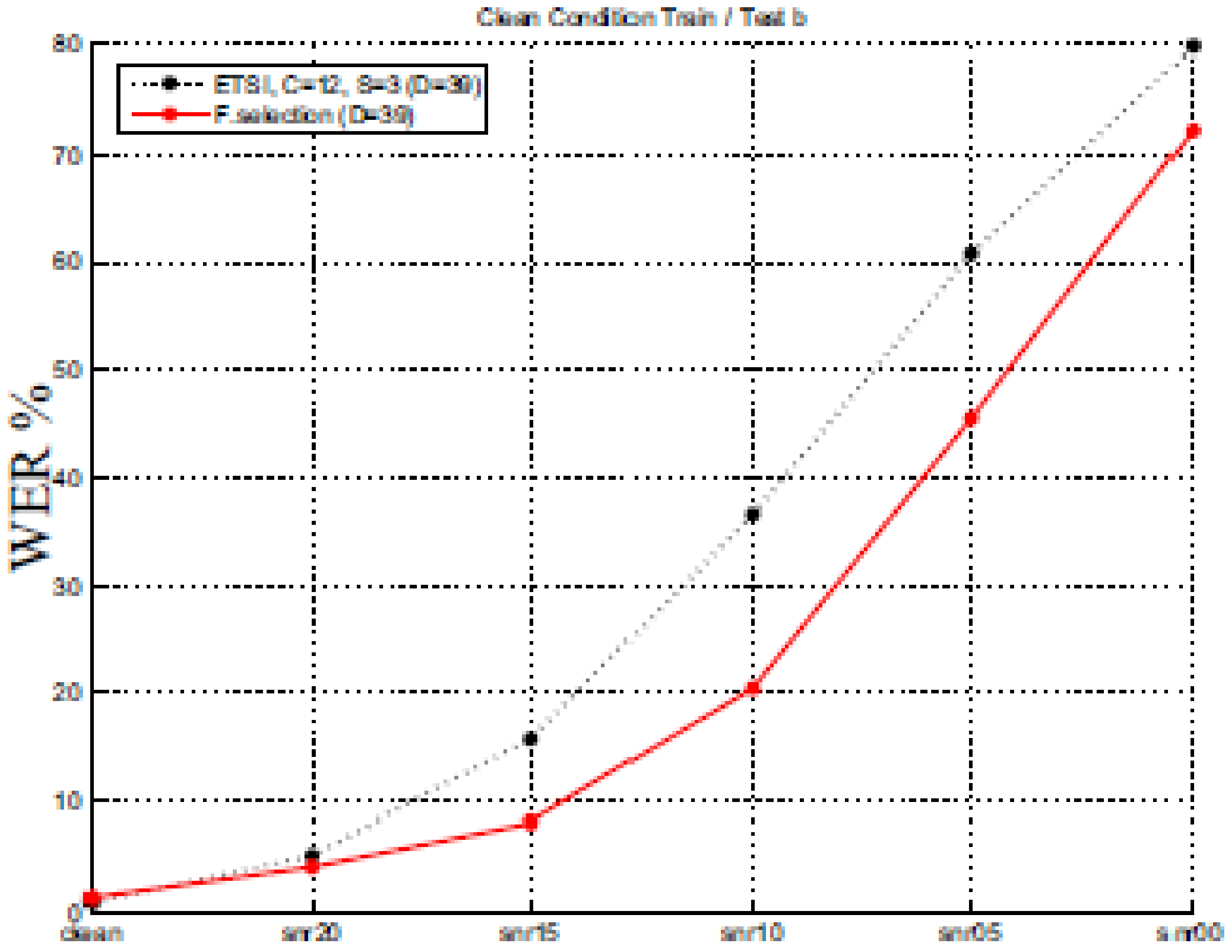
Z : Phoneme lable variable

X : Feature being analyzed

Experiments

- Performed on Aurora 2
- Compare with LDTCs and DCT2
- The 39 dimension MFCC for C=12 , S=3


Clean Condition Train / Test b snr10
WER %
Feature vector dimension D
LDCT
DCT2
DCT2 C=12, S=3
F.selection MMI
F.selection MMI + support



Clean Condition Train / Test b
ETSI, C=12, S=3 (D=39)
F.selection (D=39)
WER %
0
10
20
30
40
50
60
70
80
clean
snr20
snr15
snr10
snr05
snr00