Noise-Robust Speech Features Based on Cepstral Time Coefficients

Ja-Zang Yeh, and *Chia-Ping Chen*

Presented at Rocling 2009

Outline

- Introduction
- Cepstral Time Coefficients
- Experiments
- Conclusion

Outline

- **Introduction**
- Cepstral Time Coefficients
- Experiments
- Conclusion

Introduction

- A front-end of a speech recognition system may consist of several stages.

  - In the early stage : spectral subtraction and Wiener filter

  - In the middle stage : pre-emphasis and hamming window

  - In the post-processing stage: normalization, temporal information integration.

Introduction(Cont.)

- We investigate novel features based on simple transformation post-processing methods.

  - Insert a window of static cepstral vectors in a matrix and then apply the *discrete cosine transform*(DCT)

  - Coefficients after DCT is called *cepstral time coefficients*

  - Resultant matrix is called *cepstral time matrix* (CTM)

  - Further apply normalization and routines for delta and acceleration feature extraction

  - Combined with the static MFCC features

Outline

- Introduction
- **Cepstral Time Coefficients**
- Experiments
- Conclusion

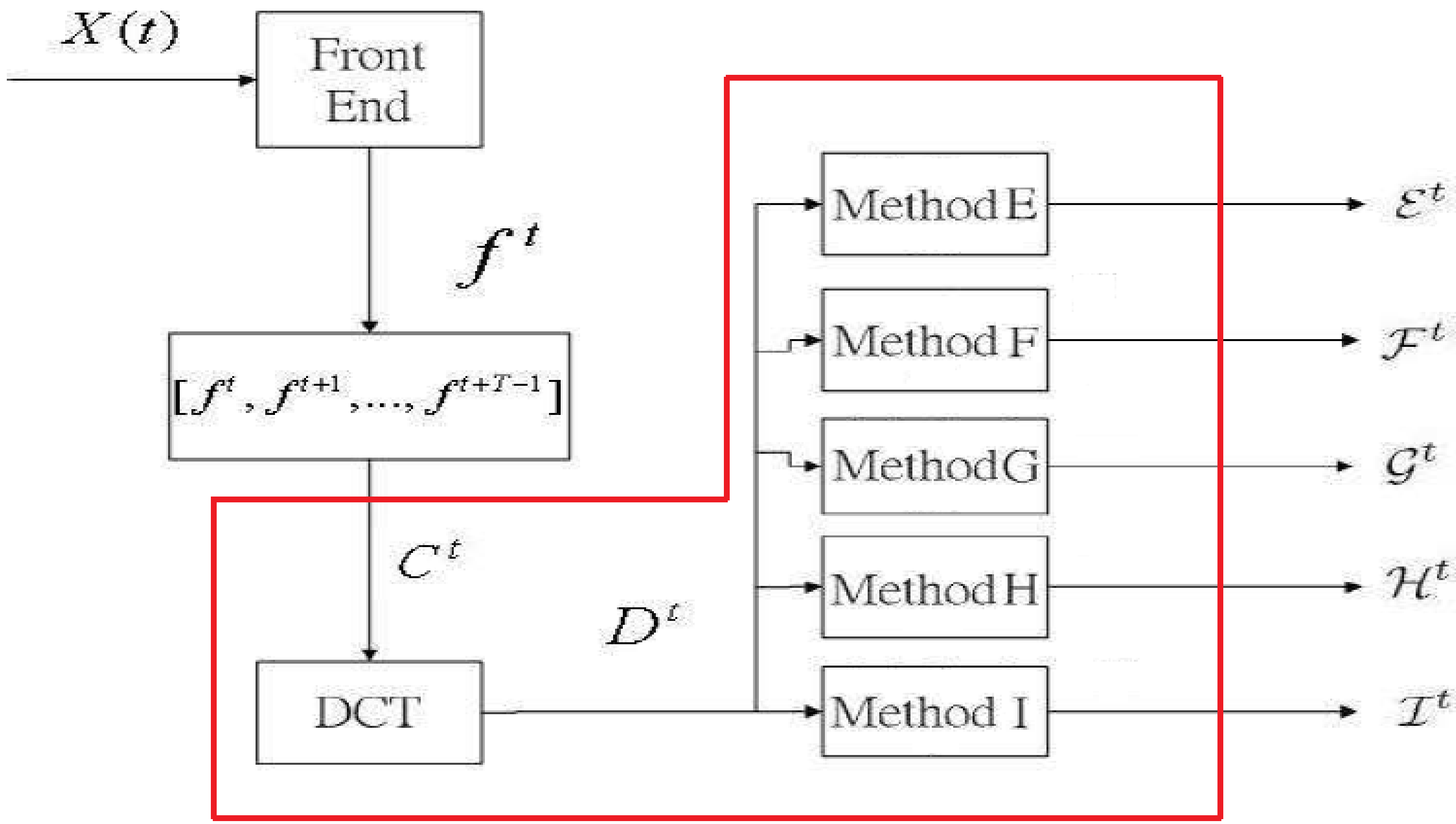


Figure 1: The block diagram of the proposed feature transformation methods.

Cepstral Time Coefficients

- We first insert a fixed number of adjacent feature vectors in a matrix

$$C^t \triangleq \begin{pmatrix} C^t_{11} & \cdots & C^t_{1T} \\ \vdots & \ddots & \vdots \\ C^t_{K1} & \cdots & C^t_{KT} \end{pmatrix} \triangleq \left[f^t \ f^{t+1} \cdots \ f^{t+T-1} \right]$$

K : the feature vector dimension

f^t : feature vector of frame t

C^t : the matrix whose column vectors are the T consecutive feature vectors starting from frame t

Cepstral Time matrix

- The cepstral time matrix at frame t, D^t, is related to C^t by the discrete-cosine transform(DCT).

$$D_{i:}^{t} = DCT(C_{i:}^{t})$$

$D_{i:}^{t}$: the i-th **row** of matrix $\boldsymbol{D}^t$
D_{in}^{t} : the n-th cepstral time coefficient (CTC) of channel i at frame t

- Our matrix index starts from 1 instead of 0

$$D_{in}^{t} = \sum_{\tau=1}^{T} C_{i\tau}^{t} \cos\left(\frac{(2\tau-1)(n-1)\pi}{2T}\right)$$

Method E

- Dividing the first column of D^t by the number of frames , while leaving other columns unchanged.

$$\begin{cases} E_{:1}^{t} = D_{:1}^{t} / T \\ E_{:n}^{t} = D_{:n}^{t}, \quad n \neq 1 \end{cases}$$

- Then we apply the delta and acceleration feature extraction steps.

$$\begin{cases} \breve{E}_{:2}^{t} = E_{:2}^{t} - E_{:1}^{t} \\ \breve{E}_{:3}^{t} = E_{:3}^{t} - 2E_{:2}^{t} + E_{:1}^{t} \end{cases}$$

Method E(Cont.)

- We add the $\breve{E}_{:2}^{t}$ and $\breve{E}_{:3}^{t}$ to the static MFCCs, resulting in a feature vector of

$$\mathcal{E}^{t} = \begin{bmatrix} C_{:1}^{t} \\ \breve{E}_{:2}^{t} \\ \breve{E}_{:3}^{t} \end{bmatrix}$$

Method F

- Normalize the feature values in the first column to the range of $[-1,1]$. Let F^t be defined by

$$\begin{cases} F_{:1}^t = D_{:1}^t / N^t \\ F_{:n}^t = D_{:n}^t, \quad n \neq 1 \end{cases}$$

N^t : the maximum magnitude in the first column.

i.e. $N^t = \underset{d}{Max} \left| D_{d1}^t \right|$

- The remaining operations are similar to Method E.

$$\begin{cases} \breve{F}_{:2}^t = F_{:2}^t - F_{:1}^t \\ \breve{F}_{:3}^t = F_{:3}^t - 2F_{:2}^t + F_{:1}^t \end{cases}$$

Method F(Cont.)

- We add $\breve{F}^{t}_{:2}$ and $\breve{F}^{t}_{:3}$ to the static MFCCs, resulting in a feature vector of

$$\mathcal{F}^{t} = \begin{bmatrix} C^{t}_{:1} \\ \breve{F}^{t}_{:2} \\ \breve{F}^{t}_{:3} \end{bmatrix}$$

Method G

- In Method G, we add the first and second columns of D^t to the static MFCC vector,

$$\mathcal{G}^t = \begin{bmatrix} C_{:1}^t \\ D_{:1}^t \\ D_{:2}^t \end{bmatrix}$$

Method H

- In Method H, we add the second and third columns of D^t to the static MFCC vector,

$$\mathcal{H}^t = \begin{bmatrix} C^t_{:1} \\ D^t_{:2} \\ D^t_{:3} \end{bmatrix}$$

Method I

- In Method I, we simply use the zeroth, first, and second cepstral time coefficients,

$$\mathcal{I}^{t} = \begin{bmatrix} D^{t}_{:1} \\ D^{t}_{:2} \\ D^{t}_{:3} \end{bmatrix}$$

Outline

- Introduction
- Cepstral Time Coefficients
- **Experiments**
- Conclusion

Evaluation corpus

- Evaluated by Aurora 3 corpus.
- Aurora 3 consisting of digit-string utterances in Danish, German, Finnish and Spanish.
- It provides a platform for fair comparison between systems of different front-ends.

Back end

- We use HTK as the back end to run experiments.
- Implemented by HMM, and use word-level model.
- Each word model consists of 16 emitting states, and each consists of 3 Gaussian components.
- The silence model consists of 3 sate, and each state consists of 6 Gaussian components.

Experiments set

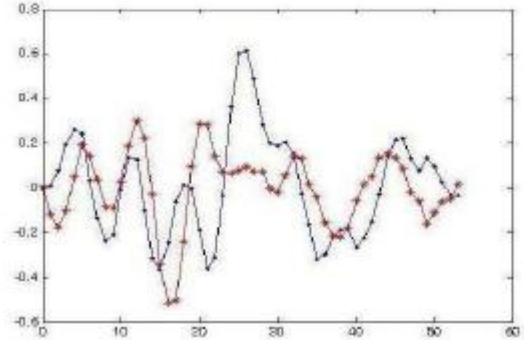
- We first evaluate and decide to use $T = 15$.
- For the static features we use 12 MFCC features and the log energy, making $K = 13$.
- Our baseline, it simply uses the MFCC, delta and acceleration features.

Relative improvement

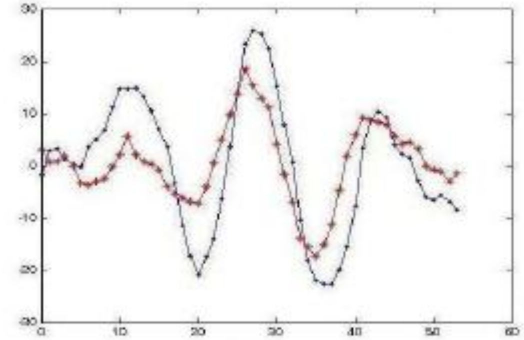
	German	Spanish	Finnish	Danish
Method E	-12.4	16.2	16.5	16.3
Method F	-10.5	22.4	10.8	16.3
Method G	-58.1	-29.0	-42.9	-19.2
Method H	7.5	26.6	25.4	23.2
Method I	-10.8	19.8	8.5	13.1

Discussion

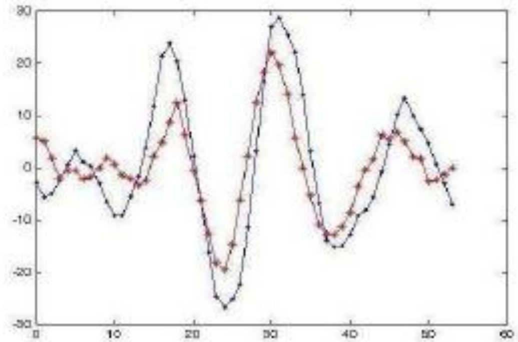
- Method E and F are similar, and they have similar performance level.
- Method G and H concludes that zeroth CTC is detrimental of recognition.
- We try scheme of normalizing and dividing.


0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
10
20
30
40
50
60

(B)


30
20
10
0
-10
-20
-30
0
10
20
30
40
50
60

(F)


30
20
10
0
-10
-20
-30
0
10
20
30
40
50
60

(H)

Outline

- Introduction
- Cepstral Time Coefficients
- Experiments
- **Conclusion**

Conclusion

- In this paper, we use five difference feature sets based on the cepstral time coefficients.
- The combination of raw MFCC and the second and the third columns of CTM yields the best.