HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | Speech Recognition Using Stochastic DTW |
| Author(s) | Yuxin, Zhang; Yoshikazu, Miyanaga; Siriteanu, Constantin |
| Relation | グリーン回路とシステムに関する国際ワークショップ = International Workshop on Green Circuits and Systems. 2010年10月25日（月）. 北海道大学大学院情報科学研究科大会議室（11階17号室）, 札幌市. |
| Issue Date | 2010-10-25 |
| Doc URL | https://hdl.handle.net/2115/44262 |
| Type | conference presentation |
| File Information | ZhangYuxin.pdf |

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

# Speech Recognition Using Stochastic DTW

Zhang Yuxin  Yoshikazu Miyanaga  and  Siriteanu Constantin

Hokkaido University, Japan

# Background

- Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) algorithms have been applied widely to speech recognition
- HMM has been the dominant technique in speech recognition

Table 1: Performance of HMM and DTW

| | HMM | DTW |
|---|---|---|
| Training | High | Zero |
| Complexity | Difficult | Easy |
| Accuracy | High | Low |

- How to get the more higher accuracy by DTW?



PRESENTATION

# New method for Speech recognition using DTW

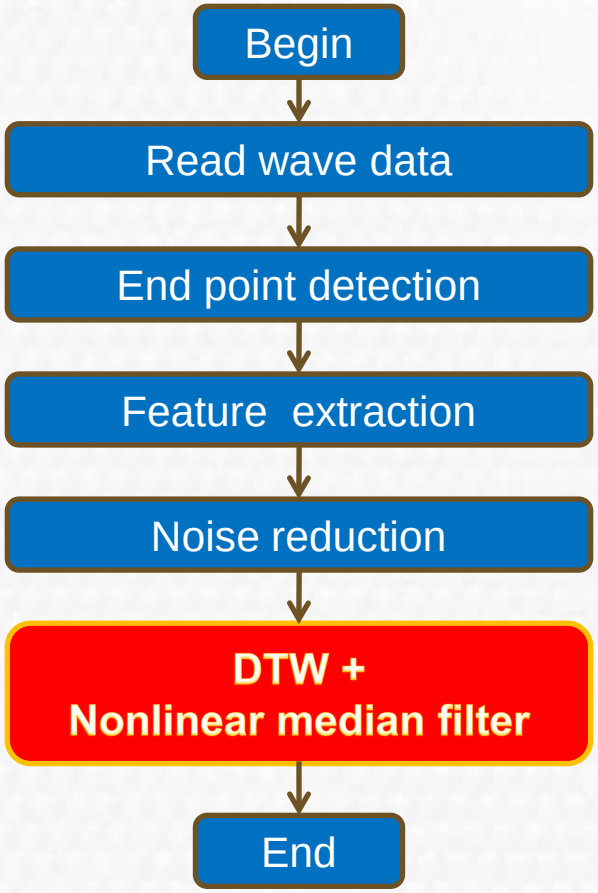



Fig.1 Flowchart of speech recognition using DTW



PRESENTATION

# End point Detection (1)

- Short time energy *E*:

$$E = \sum_{m=-\infty}^{+\infty} [x(m)(m-n)]^2 \qquad (1)$$

- Maximun energy of non-speech $\tau$:
  - n: number of frame (n = 5)
  - α: weight factor (α = 1.5)

$$\tau = \alpha \times \frac{1}{n} \sum_{i=1}^{n} E(i) \qquad (2)$$

- Noise energy level of frame *F(i)*:
  - λ: forgetting factor (0 ≤ λ ≤ 1)

$$F(i) = \lambda F(i-1) + (1-\lambda) E(i) \qquad (3)$$



PRESENTATION

# End point Detection(2)

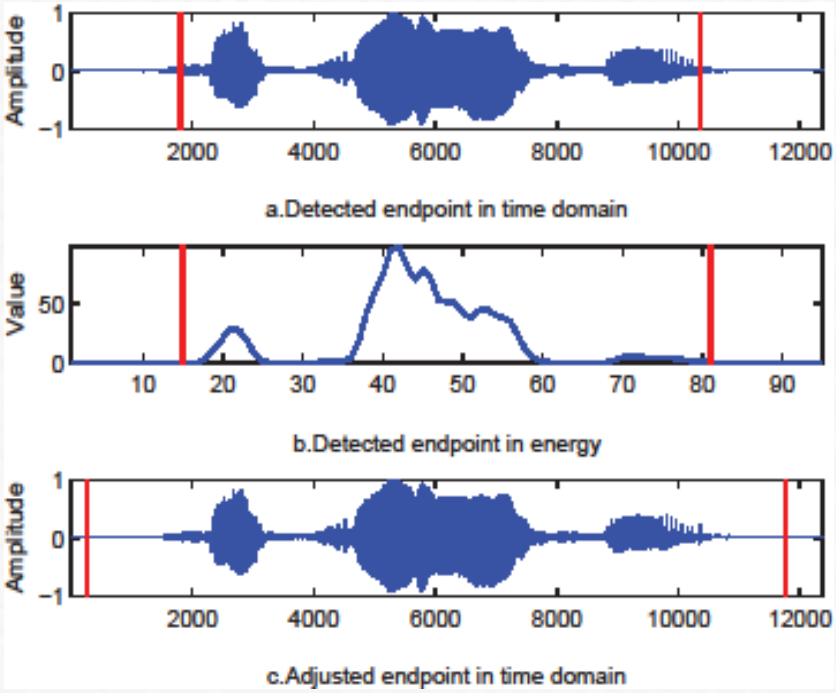



Fig.2 Clean speech data with endpoint detection

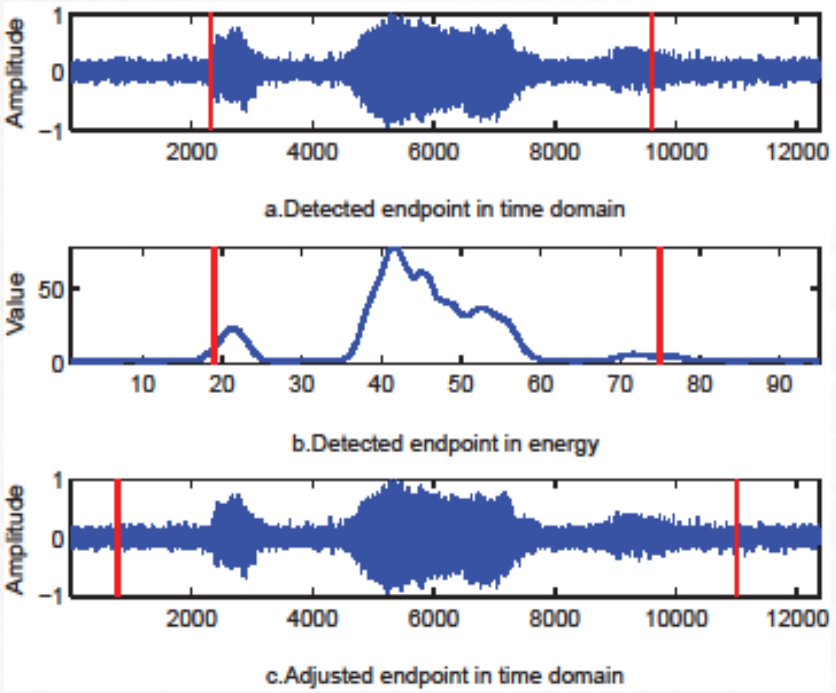



Fig.3 Speech data with endpoint detection at SNR =10 dB



PRESENTATION

# Noise Reduction (RSF+DRA)

- Most of noise spectrum energy concentrates around the direct current (DC) component .
- Some relatively noise of lower energy at high frequency
- Most of the noise energy is comprised in the band [0,1] Hz
- Band [1,16] Hz is important

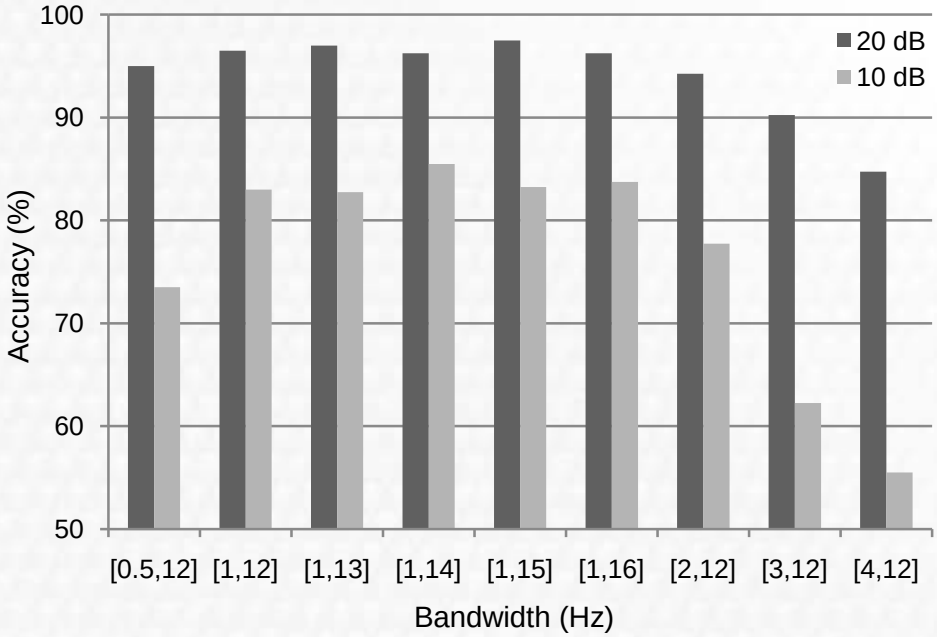


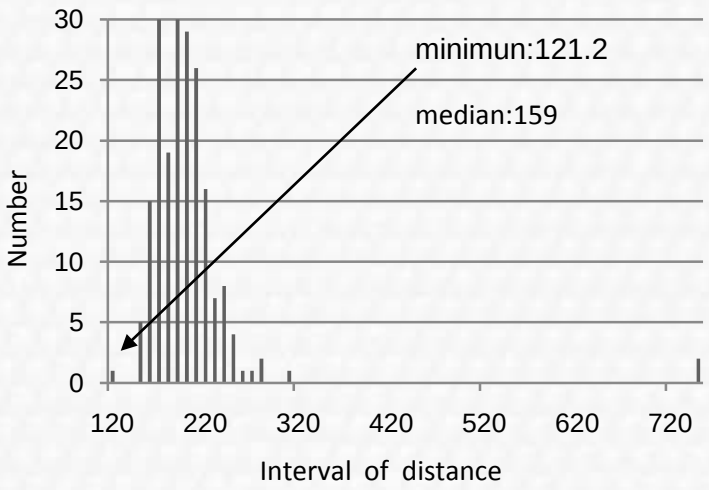

Fig.4 Recognition rate vs. bandwidth for RSF+DRA



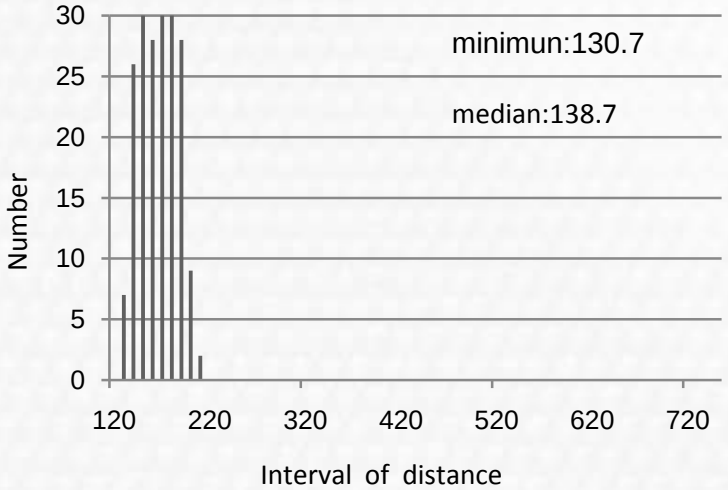
PRESENTATION

# DTW with nonlinear median filter

- Accuracy of recognition is not so high  if only DTW is used to recognize.
- Nonlinear median filter can improve the accuracy of recognition




a. Distribution of Distance to  "Date"




b. Distribution of Distance to  "Hatinohe"

Fig.5 Distributions of all distances for which the test word ”Hatinohe” matches ”Date” and ”Hatinohe”.



PRESENTATION

# Experiment parameter

- The test words: 50 Japanese words. Every word has 100 waveforms spoken by 100 persons.
- The reference words: 100 Japanese words and every word has 50 waveforms spoken by 50 persons.

Table 2: Parameter of experiment

| Recognition task | Isolated 100 words |
|---|---|
| Speech data | 100 Japanese region names from JEIDA |
| Sampling | 11.025kHz (16 bit) |
| Window Length | 23.2ms (256 points) |
| Frame Period | 11.6ms (128 points) |
| Bandwidth of bandpass filter | 1~16Hz |
| Feature extraction | 38 dimensional MFCC |
| Noise varieties | White and Babble noise |

# Experiment Result

Table 3: Recognition rate with end point detection

| Experiment methods | | Without end-point detection (%) | | With end-point detection (%) | |
|---|---|---|---|---|---|
| | | White 10 dB | White 20 dB | White 10 dB | White 20 dB |
| A | All data with CMS&DRA | 67.68 | 86.7 | 78.42 | 93.58 |
| B | All data with RSF&DRA | 70.54 | 87.38 | 77.36 | 92.76 |
| C | All data with DRA | 44.76 | 77.18 | 53.56 | 85.1 |
| D | Reference data with RSF&DRA<br>Test data with CMS&DRA | 65.1 | 84.34 | 73.38 | 91.76 |
| E | All data without noise reduction | 14.7 | 65.7 | 19.34 | 72.12 |



PRESENTATION

Table 4: Recognition rate with nonlinear median filter

| Methods | Without median filter (%) | | | | With median filter (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | white | | Babble | | white | | Babble | |
| | 10 dB | 20 dB | 10 dB | 20 dB | 10 dB | 20 dB | 10 dB | 20 dB |
| A | 78.42 | 93.58 | 71.9 | 90.54 | 84.18 | 96.92 | 77.74 | 93.28 |
| B | 77.36 | 92.76 | 70.42 | 89.16 | 85.04 | 97.08 | 77.38 | 92.82 |
| C | 53.56 | 85.1 | 56.52 | 85.38 | 58.3 | 90.14 | 65.66 | 89.3 |
| D | 73.38 | 91.76 | 68.1 | 88.26 | 82.4 | 96.5 | 76.06 | 92.48 |
| E | 19.34 | 72.12 | 22.94 | 73.84 | 24.2 | 77.46 | 28.94 | 79.06 |



PRESENTATION

# Conclution

- The endpoint detection is necessary to the DTW
- RSF/DRA is best among four methods in the white noise.
- The accuracy of DTW is improved by nonlinear median filter
- All recognition rates almost are close to that one using HMM with the same noise reduction in the 10dB and 20dB SNR.



PRESENTATION

# Thank you