An Improved Dynamic Time Warping Algorithm Employing Nonlinear Median Filtering

Zhang Yuxin and Yoshikazu Miyanaga

Hokkaido University, Sapporo, 060-0814, Japan
Email: zyx@icn.ist.hokudai.ac.jp miya@ist.hokudai.ac.jp

Performance of HMM and DTW

- 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

Speech recognition system

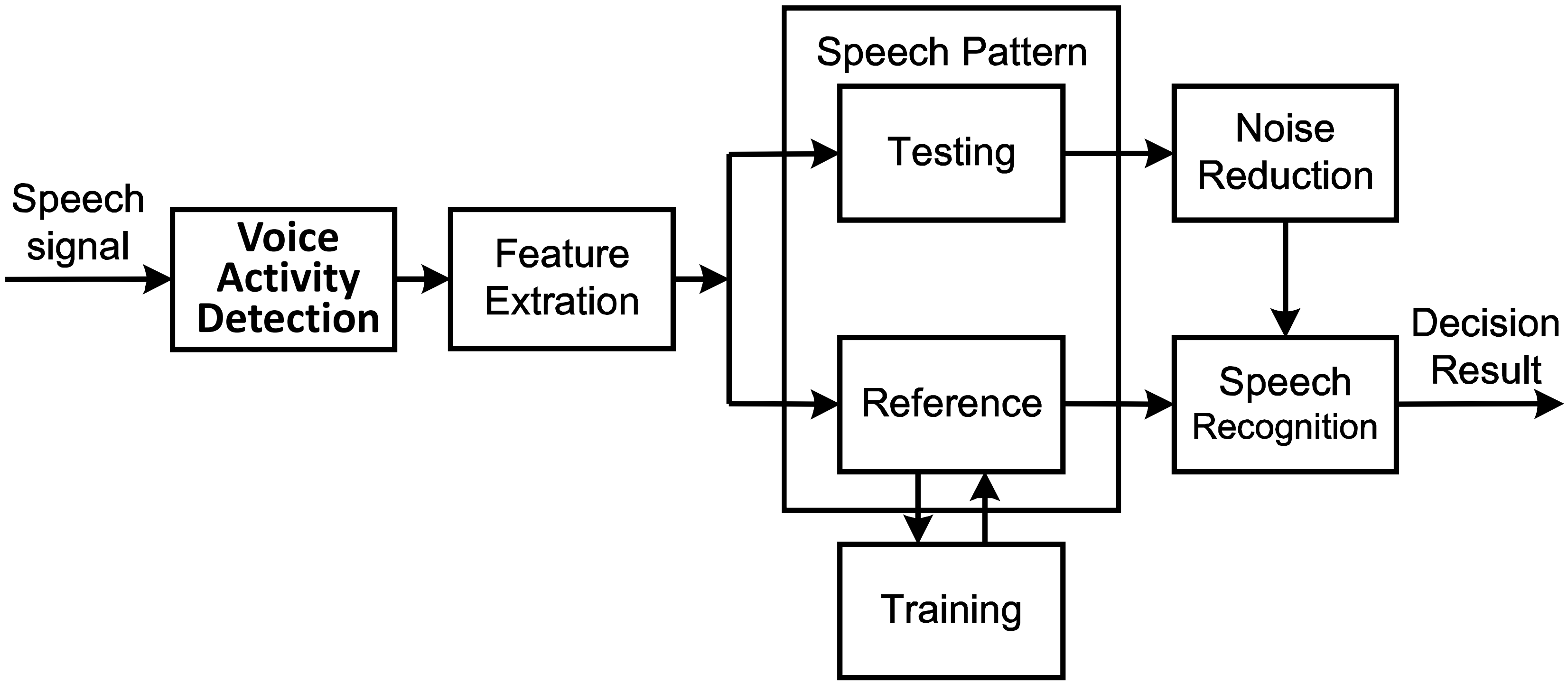


Fig. 1: ASR system diagram

Voice Activity Detection (VAD)

- Short time energy *E*:

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

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

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

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

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

Running Spectrum Filtering (RSF)

- Most of noise spectrum energy concentrates around the direct current (DC) component .
- Some relatively noise of lower energy at high frequency
- Assuming the feature vector of a noisy speech is $S=[S(1),S(2),\ldots,S(t),\ldots,S(T)]$, $S(t)=[f_1(t), f_2(t),\ldots, f_k(t)]$.
- RSF function is:

$$S(t) = Filter(S(t)) \quad (4)$$

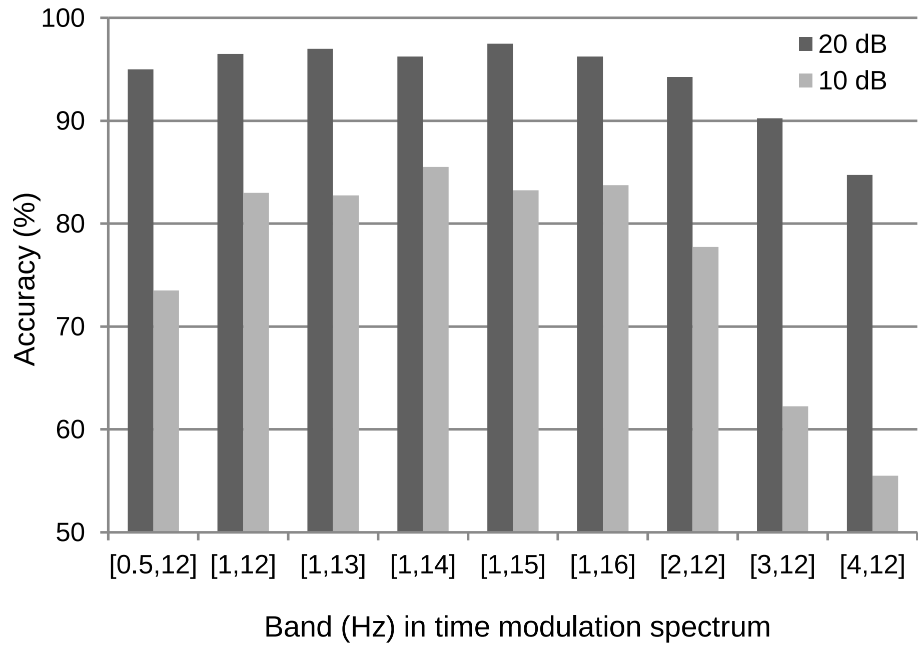


Fig.4: Recognition rate vs. band for RSF+DRA

Cepstral Mean Subtraction (CMS)

- The white noise is usually uniformly distributed in the whole spectrum.
- If the each coefficient subtract the average of every channel, then the average of noisy speech can be reduced to almost zero.
- CMS function is

$$f_i(t) = f_i(t) - \frac{1}{k}\sum_{j=1}^{k} f_j(t) \qquad (5)$$

Dynamic Range Adjustment (DRA)

- DRA algorithm: $f_i(t) = f_i(t) / \max_{i=1,\dots,k} \left| f_i(t) \right|$ (6)

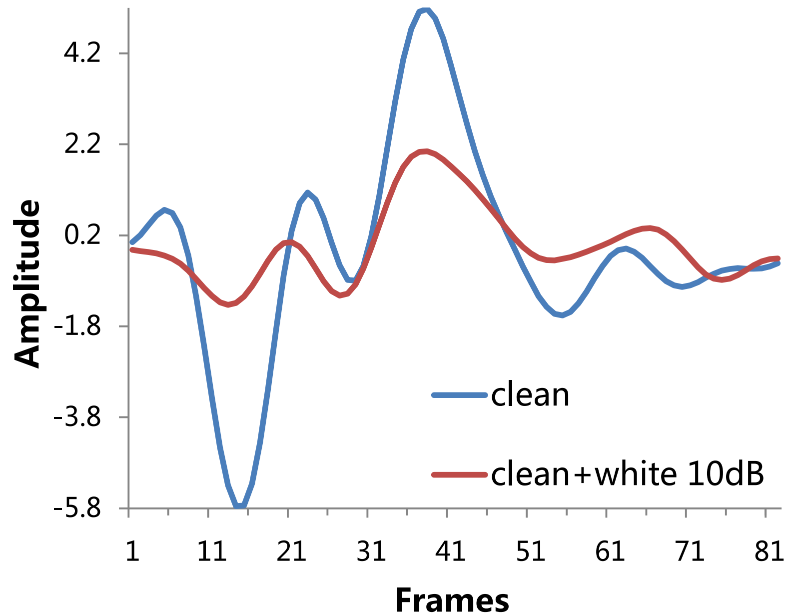


Fig.5: The No.3 dimension data of MFCC after RSF

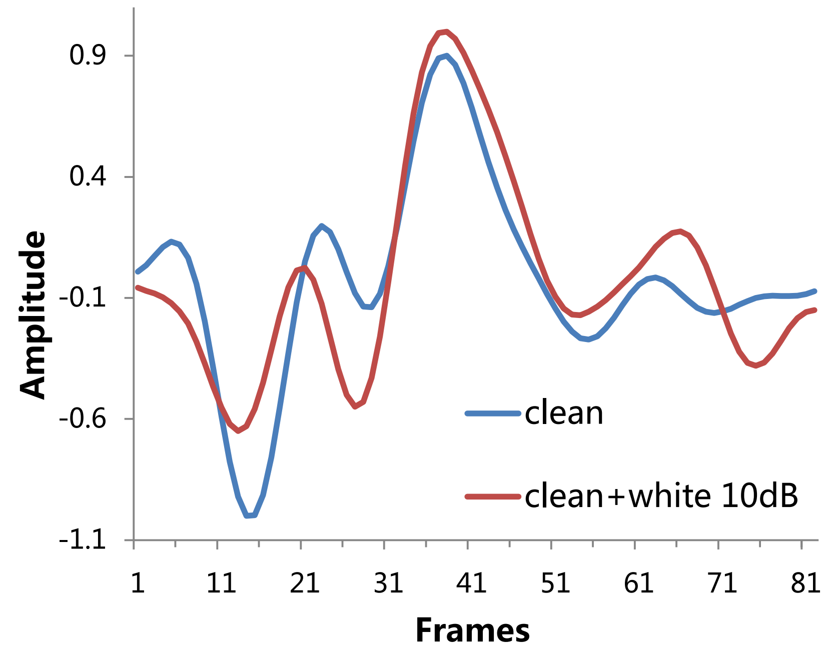
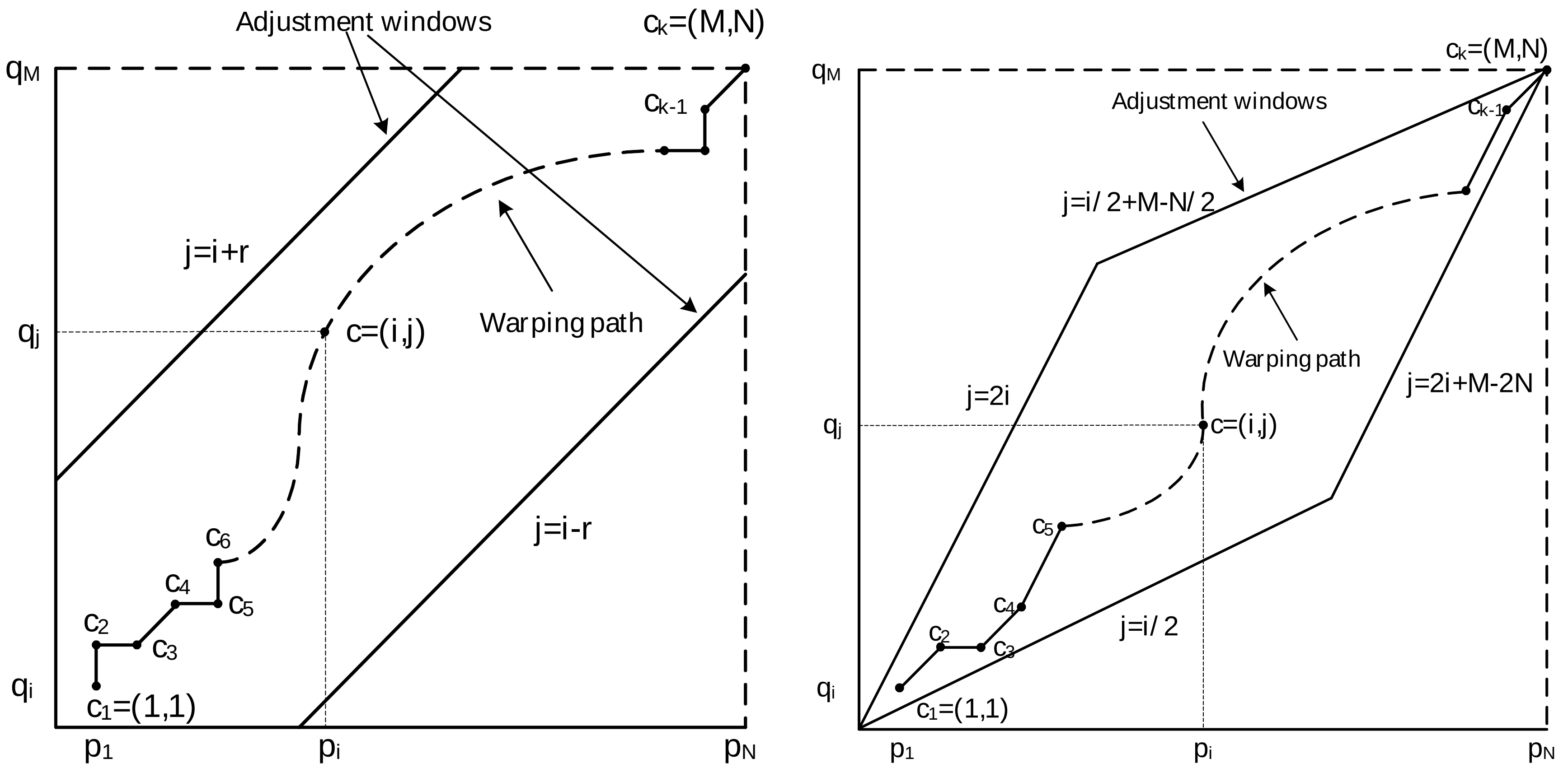


Fig.6: The No.3 dimension data of MFCC after RSF and DRA

DTW algorithms



(a) the Sakoe-Chuba Band (b) Itakura Parallelogram

Fig.7: Two of the most commonly used constraints.

The Accuracy of DTW and HMM

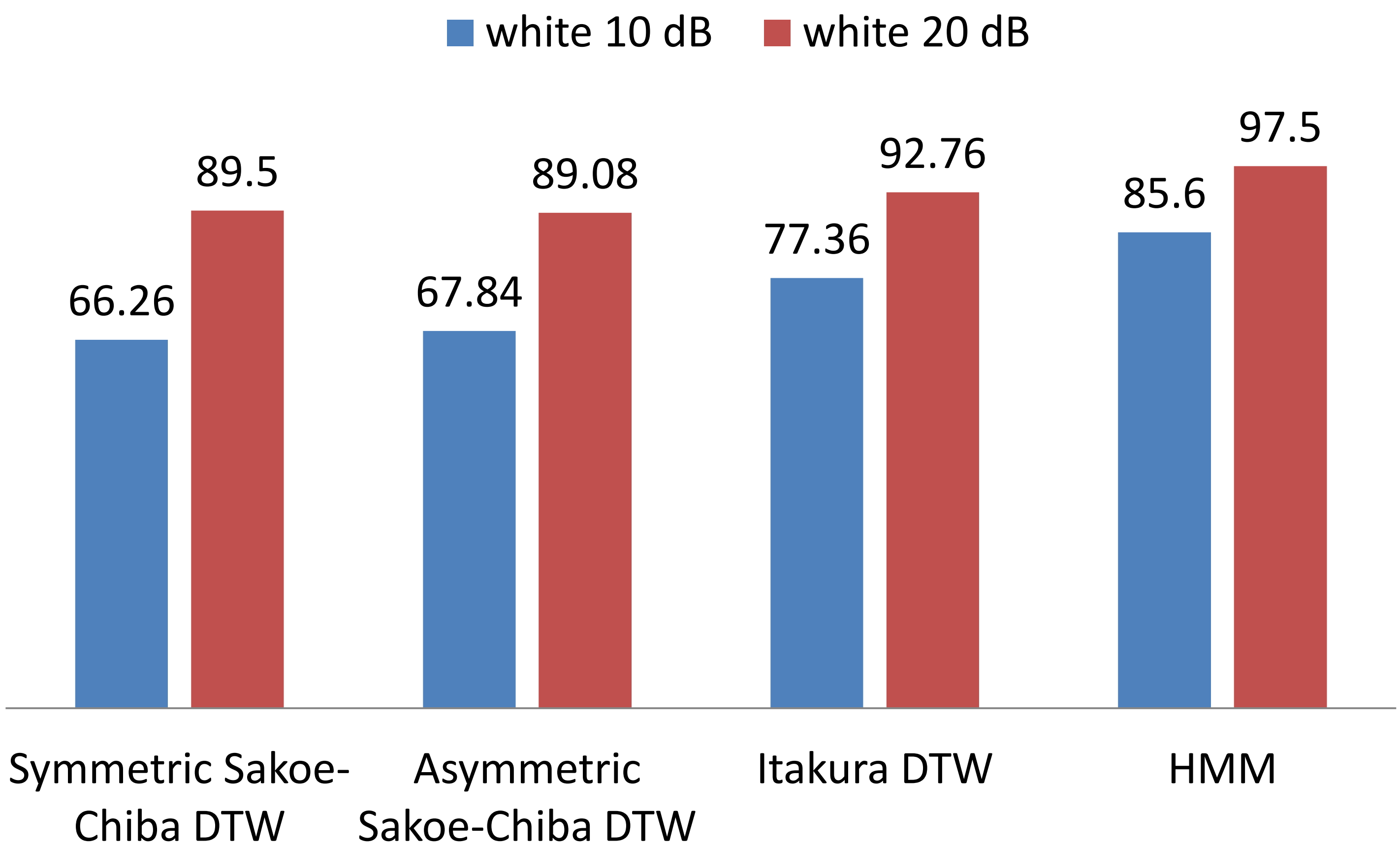
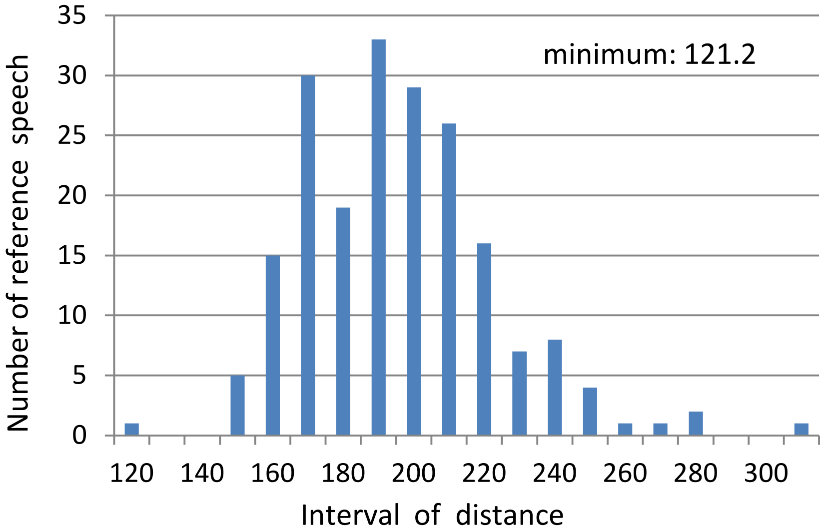
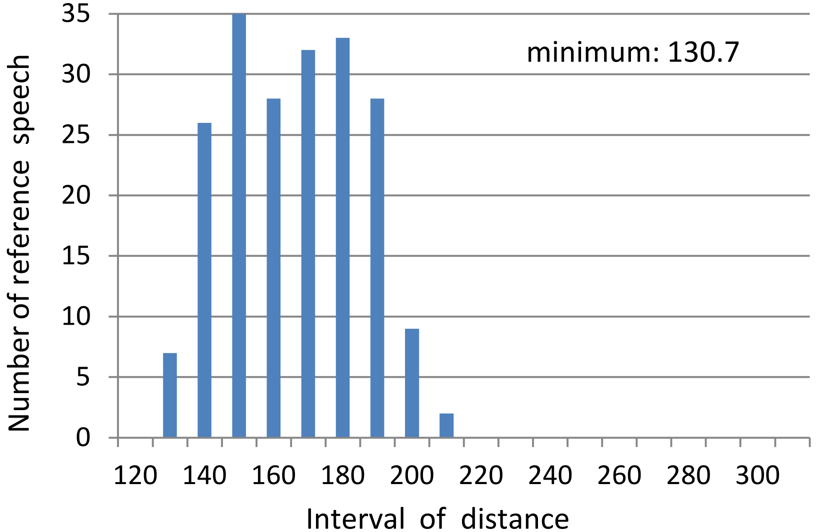


Fig.7: The accuracy of four ASR algorithms by RSF and DRA in white noise , SNR = 10, 20 dB.

DTW with Nonlinear Median filter(1)



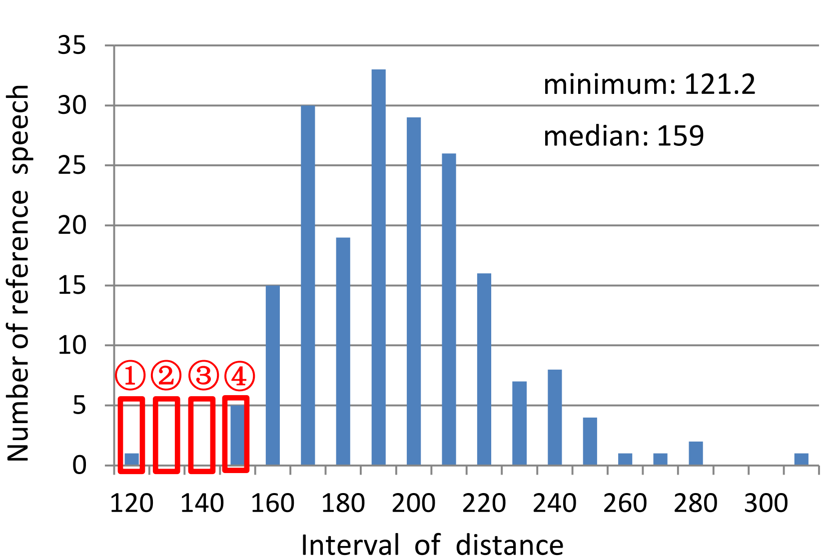
a. Distribution of distance to "Date"



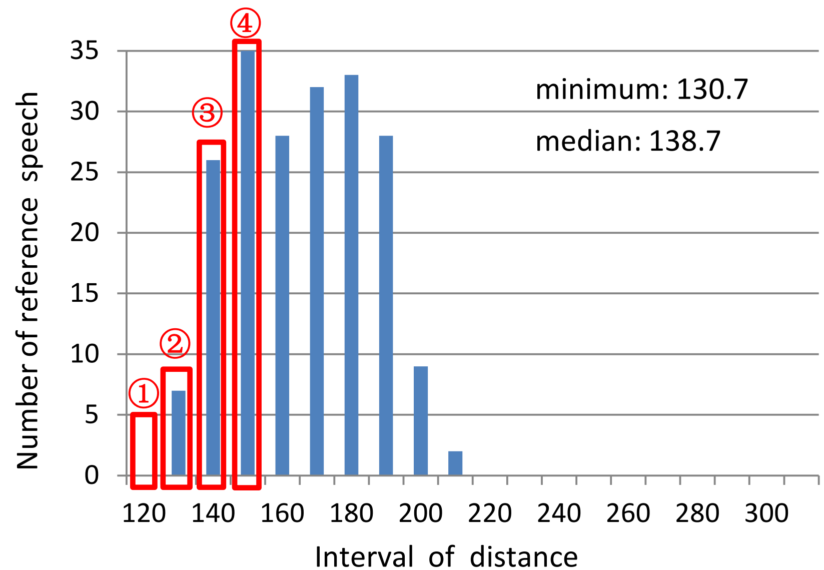
b. Distribution of distance to "Hatinohe"

Fig.8: Distributions of distance between unknown word ‘hatinohe’ and reference words ‘date’ and ‘hatinohe’.

DTW with Nonlinear Median filter(1)



a. Distribution of distance to "Date"



b. Distribution of distance to "Hatinohe"

Fig.9: Distributions of distance between unknown word ‘hatinohe’ and reference words ‘date’ and ‘hatinohe’.

DTW with Nonlinear Median filter(2)

- Assuming the matching distance Matrix is
  $D=[D_1,D_2,\ldots,D_M]$, $D_m=[d_{m,1},d_{m,2},\ldots, d_{m,N}]$

(1) Sorting ascendingly the distances for every reference word yields D'_m

$$D'_m = Sort(D_m) = [d'_{m,1}, d'_{m,2},\ldots, d'_{m,N}] \tag{7}$$

(2) Computing the median by the NMF.

$$a_m = Med(d'_m) = \begin{cases} d'_{m,\frac{k+1}{2}} & ,if\ k\ is\ odd, \\ \frac{1}{2}[d'_{m,\frac{k}{2}} + d'_{m,\frac{k}{2}+1}], & if\ k\ is\ even. \end{cases} \tag{8}$$

(3) In the approach we propose herein the recognized word corresponds to

$$\arg\min_{m=1:M} a_m \tag{9}$$

※The conventional DTW approaches the recognized word corresponds to

$$\arg\min_{m=1:M} d'_{m,1} \tag{10}$$

DTW and Nonlinear Median filter(3)

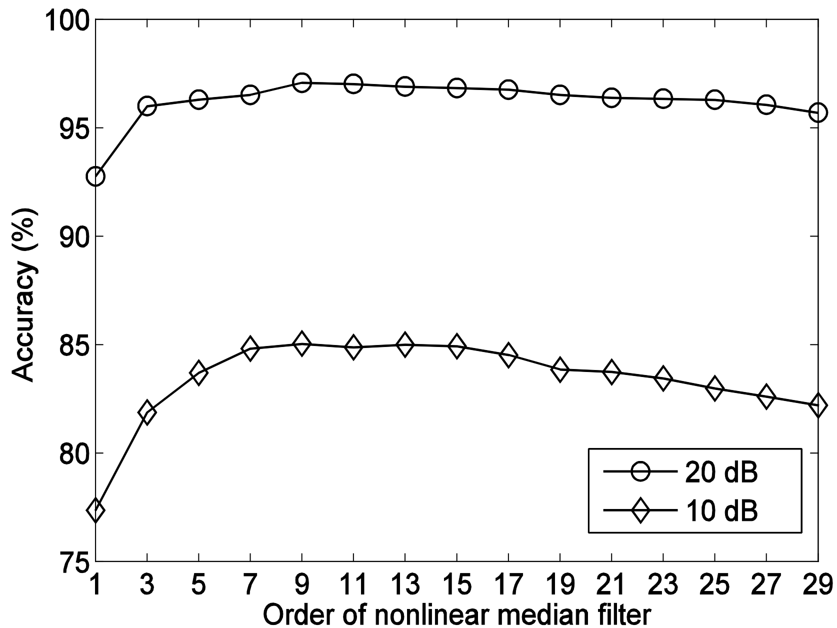


Fig.10: Accuracy of DTW with NMF vs. filter order.

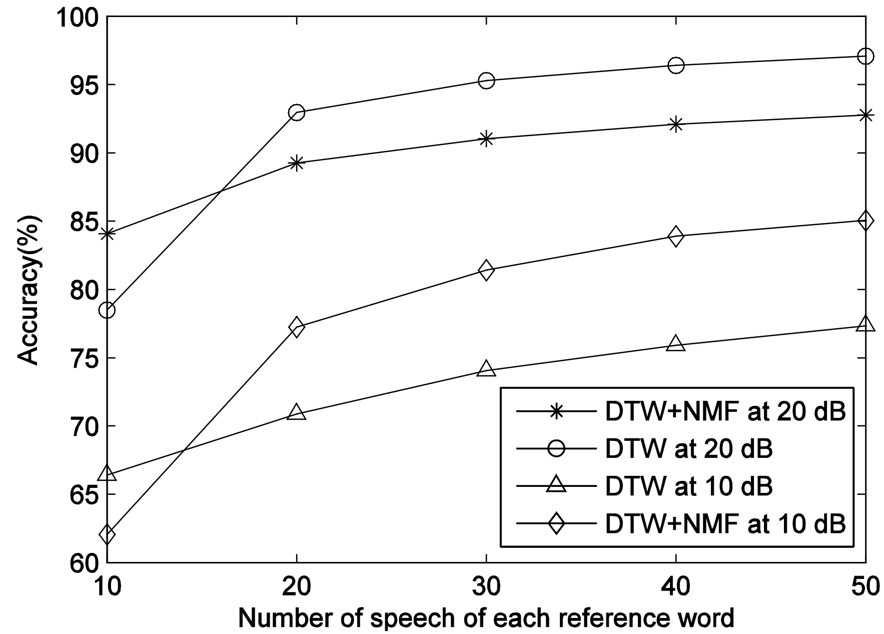
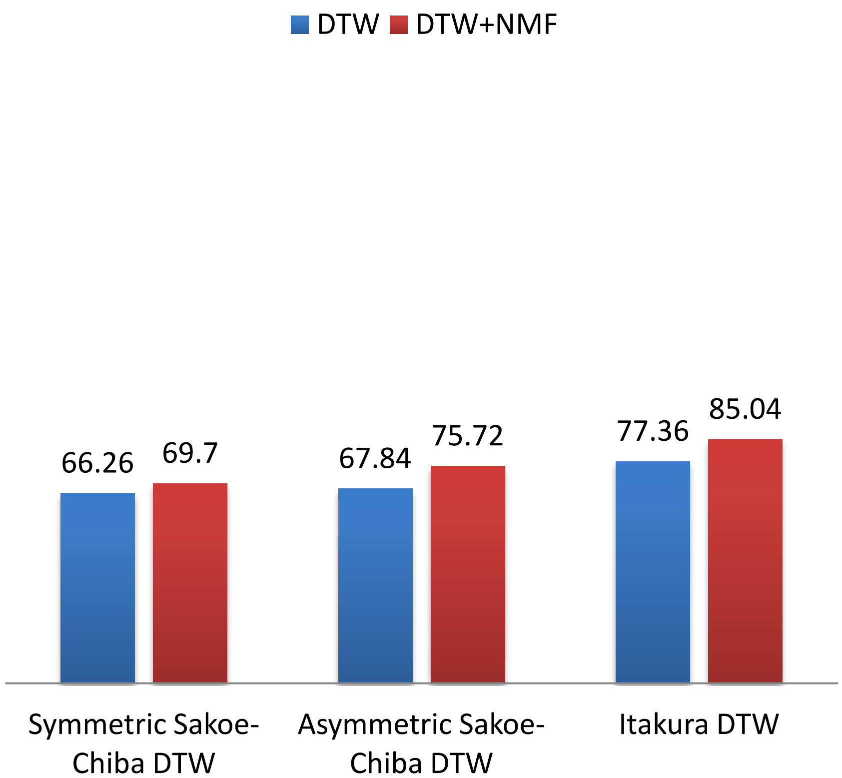
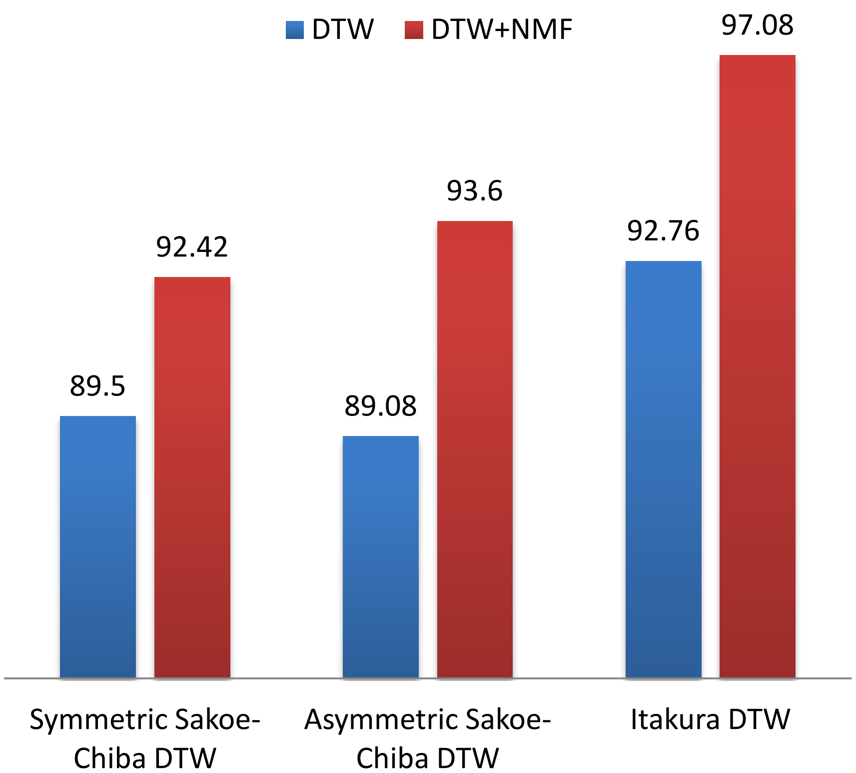


Fig.11: DTW accuracy with NMF vs. number of waveforms, for NMF order 9 and 10 dB SNR.

DTW and Nonlinear Median filter(4)



(a) SNR = 10 dB



(b) SNR = 20 dB

Fig.12: Three DTW algorithms accuracy with NMF for order 9 at 10 and 20 dB SNR (%).

Experiment

Table 2: Experiment setting and parameters

Recognition task	Isolated 100 words
Speech data	100 Japanese region names from JEIDA
Sampling	11.025 kHz, 16 bits
Window length	23.2 ms (256 samples)
Frame length	11.6 ms (128 samples)
Bandwidth of bandpass filter	1 – 16 Hz
NMF order	9
Feature vector	38-dimensional MFCC
Noise type	White noise and babble noise

Experiment

- We have considered the following cases:
  - A) CMS and DRA are applied for testing waveforms and reference waveforms
  - B) RSF and DRA are applied for testing waveforms and reference waveforms
  - C) CMS, RSF and DRA are applied for testing waveforms and reference waveforms
  - D) DRA is applied for testing data and reference waveforms
  - E) No noise reduction; testing data was recognized directly.

Result

Table 3: Recognition accuracy (%) with NMF and VAD

Case	without VAD		with VAD	
	10 dB	20 dB	10 dB	20 dB
A	67.68	86.7	78.42	93.58
B	70.54	87.38	77.36	92.76
C	70.18	87.02	77.08	92.96
D	67.34	84.29	73.38	91.76
E	14.7	65.7	19.34	72.12
clean	90.48		98.52	

Table 4: Recognition accuracy (%) of Itakutra DTW, w/o NMF

Case	DTW				DTW+NMF			
	White		Babble		White		Babble	
	10 dB	20dB	10dB	20dB	10dB	20dB	10dB	20dB
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	77.08	92.96	71.06	89.98	84.22	97.06	78.06	93.78
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
clean	98.52				98.76			

Table 5: Recognition accuracy (%) of Symmetric Sakoe-Chiba DTW, w/o NMF

Case	DTW				DTW+NMF			
	White		Babble		White		Babble	
	10 dB	20dB	10dB	20dB	10dB	20dB	10dB	20dB
A	52.28	80	42.58	72.44	55.72	84.14	45.04	75.12
B	66.26	89.5	44.32	76.5	69.7	92.42	48.62	79.74
C	66.96	89.24	56.32	84.44	70.76	92.28	59.1	86.92
D	35.8	76.92	35.3	76.08	38.2	78.26	38.1	78.32
E	7.51	56.96	6.8	54.24	7.1	60.34	11.74	63.82
clean	96.14				97.28			

Table 6: Recognition accuracy (%) of Asymmetric Sakoe-Chiba DTW, w/o NMF

Case	DTW				DTW+NMF			
	White		Babble		White		Babble	
	10 dB	20dB	10dB	20dB	10dB	20dB	10dB	20dB
A	65.02	87.46	55.4	80.16	73.7	92	63.28	85.88
B	67.84	89.08	57.62	82.36	75.72	93.6	66.34	86.7
C	68	88.84	58.06	82.52	75.86	93.9	66.48	87.2
D	49.48	74.68	39.64	64.98	61.2	83.04	49.12	72.42
E	7.54	56.96	7.72	55.6	7.1	60.34	6.94	65.9
clean	96.58				97.92			

Conclusion

- The performances of all DTW algorithms are improved by NMF.
- The accuracy of Itakura's DTW is best among all DTW algorithms and close to that of HMM.
- The VAD is necessary to all DTW algorithms.
- The method CMS, RSF and DRA are combined is best among four methods.