Robustness of Lipreading against the Variations of Rotation, Translation and Scaling

Duk-Soo Min JinYoung Kim* SeungHo Choi** KiJung Kim***
Chonnnam National University
**Dongshin University
***Kwangyang College
The Department of Electronic Engineering Chonnam National University
300, Yongbong-dong, Puk-gu, Kwangju, Korea
Tel: +82-62-530-0472, Fax: +82-62-530-0472
E-mail: dsmin@dsp.chonnam.ac.kr, *kimjin@dsp.chonnam.ac.kr

Abstract: In this study, we improve the performance of a speech recognition system of visual information depending on lip movements. This paper focuses on the robustness of the word recognition system with the rotation, transition and scaling of the lip images. The different methods of lipreading have been used to estimate the stability of recognition performance. Especially, we work out the special system of the log-polar mapping, which is called Mellin transform with quasi RTS-invariant and related approaches to machine vision. The results of word recognition are reported with HMM (Hidden Markov Model) recognition system.

Keywords: Lipreading, Log-polar mapping, RTS (Rotation, Translation, Scaling)

1. Introduction

This paper examines how the lipreading systems are robust to speech recognition against RTS variations. Our lipreading system extracts the visual features of the speech from the image of the speaker mouth.

As the processing of extracting visual information, we divided the lipreading systems into three lipreading systems. The first used the Discrete Cosine Transform for extracting, so called the DCT-lipreading [8]. And it has several advantages of simplicity, performance and best recognition among three. The others used the combination of the Fourier transform and log-polar mapping. The Fourier-Mellin transform equals to applying the operations of Fourier transform, log-polar mapping and Fourier transform to the original image [2,3]. The first Fourier transform is, up to a phase, transition invariant. The log-polar mapping provides size and rotation invariance, up to a special shift and the final Fourier transform reduces this shift to a phase [4,5]. Thus, the second approach is the FM-lipreading using the Fourier-Mellin transform. The last is the MF-lipreading, which precede a Fourier transform with the log-polar mapping to the original image.

We present a recent work on improving the performance of automated speech recognizers by using visual information, and this achieved word recognition of up to 62% in the DCT-lipreading.

How stable the lipreading system is in special variance? We seek a robustness of lipreading to image transformations such as translation, rotation and scaling. And the experiment using lipreading consists of three methods. In section 2, we define a method of lipreading to extract visual feature. Section 3 shows the degree of RTS variation for experiments on the robustness of visual speech recognition. We used only visual speech database, which is our constructions. Our database consists of Grey-level image sequences of the 22 words, spoken by 70 male. The images contain only mouth area and are digitized at 30 frames/sec, 320x240 pixels, 8 bits per pixel.

2. Lipreading Algorithms

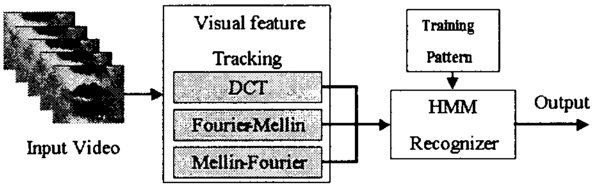


Fig 1.The visual speech recognition or lipreading

Much of the visual speech recognition system focuses on combining with audio information, assisted speech recognizer, such as audio-visual speech recognition (AVSR)[9]. We focused on the visual information. Various visual features contain the linguistic property. In general, lipreading can be grouped into lip contour-based and pixel-based ones [8].

In this paper, there are pixel-based approaches. And the visual speech feature is extracted by the pixel-based lipreading during speech. The image transform based approach is reported which obtains a compressed representation of the mouth area. We concentrate on an efficient method decreasing a lot of pixel data to be processed on an image transform approaches. To reduce the amount of pixels, we used the principal component analysis. Thus, visual features are consisted of several parameters per frame.

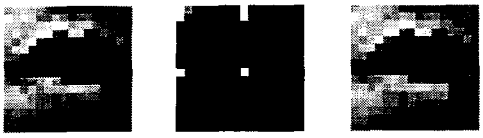
2.1 DCT Lipreading Algorithm

We first implemented the DCT to 16×16 image. To reduce the amount of the DCT coefficients, 256 vectors, we considered PCA algorithm that contains about 85 percentages of the important information among the original data. To examine the performance of the speech prediction, we used consistently a set of 15 parameters per each frame.

The definition of the two-dimensional DCT for an input image I and output image D is:

$$D(m,n)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}4I(x,y)\cos\frac{\pi}{2M}m(2x+1)\cos\frac{\pi}{2N}n(2y+1) \qquad (1)$$

where m= 0,1,...,M-1, n=0,1,...N-1.



(a) Original Image (b) DCT (c) IDCT

Fig 2. An example of the Discrete Cosine Transform

2.2 FM Lipreading Algorithm

The FM and MF lipreading are extracted by computing the Fourier power spectrum, performing a mapping from ordinary image to logarithmic-polar coordinates and computing their power spectra coefficients [1,2]. Furthermore, it has the advantage of being invariant to Mellin transformations such as any combination of rotation, scaling and translation. We seek a robustness to image transformations such as translation, rotation and scaling.

We propose an image representation invariant to 2D transformations such as rotation, shift and scaling.

First, the discrete Fourier transformation (2D-DFT) F of the Image I:

$$F(m,n)=\sum_{x=-\frac{M}{2}}^{\frac{M}{2}-1}\sum_{y=-\frac{N}{2}}^{\frac{N}{2}-1}I(x,y)e^{-2\pi i(\frac{mx}{M}+\frac{ny}{N})} \qquad (2)$$

Our goal is to achieve shift invariance, like that of the usual Fourier transform. The Fourier transform is, up to a phase, transition invariant in special domain.

We consider here the following two-dimensional log-polar transformation which is appropriate to model primate visual cortex. It is well known as the special quality of rotation and scaling invariance. The first is a coordinate transformation, from ordinary image I(x, y) to log-polar mapping L(ρ,ϕ).

Logarithmic-polar mapping is represented with the phase and power spectra ρ,ϕ:

$$Z=\rho e^{i\phi}=x+iy \qquad (3)$$

The log-polar mapping L(ρ,ϕ) between spaces can therefore be written:

$$\rho=\log_{\alpha}(\sqrt{(x^2+y^2)}) \qquad (4)$$

$$\phi=0,\frac{1}{S}2\pi,\frac{2}{S}2\pi,.........,\frac{S-1}{S}2\pi \qquad (5)$$

Where ρ = 0,1,..., M, α is an experimental constant. We use S=64 for the 64 orientation bins, and M= 64, appropriate for images of size 64×64. The log-polar mapping size is exactly the same as the original image. Thus, α=1.055645178 is a coefficient of logarithmic base.

This mapping similar to the traditional polar mapping uses logarithm which results in an exponential sampling frequency as a function of the distance from the image-centered point. In all cases the origin of the mapping is located at the center of the image.

To obtain the speech parameters, the Fourier-Mellin transforms equals to applying the operations of Fourier transform, log-polar mapping, Fourier transform to the original image. Figure 4 shows the processing in the FM lipreading. The construction process starts with the computation of the Fourier transform of the image. The power spectrum of the translated image will turn the phase spectrum. The next is the log-polar mapping with special invariance. Practically, this special transform is not particularly impressive priority of searching the pattern of visual speech parameters. The log-polar mapping provides size and rotation invariance, up to a special shift and the final Fourier transform reduces this shift to a phase.

2.3 MF Lipreading Algorithm

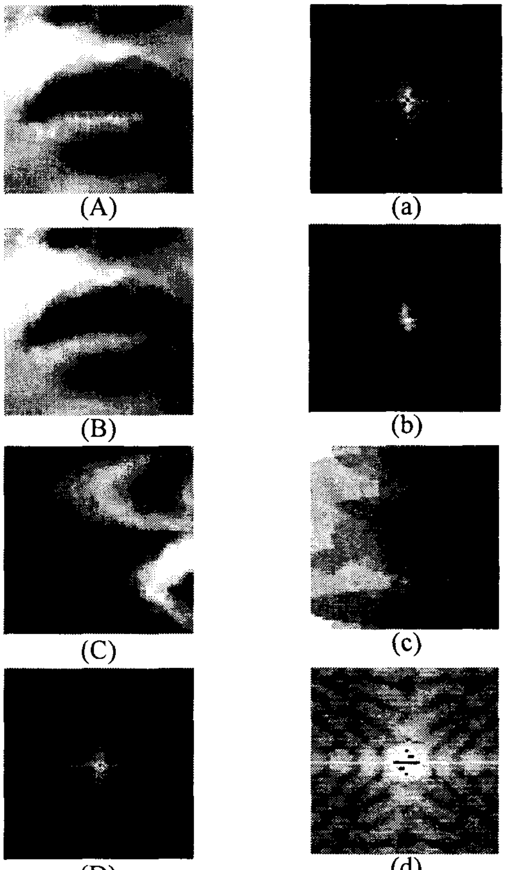


Fig 3. The illustrations of the 2D FFT and logarithmic-polar mapping with sample images. (A) ROI image through lip extraction (B) log-polar average filter of image (C) log-polar mapped image (D) Fourier power spectrum of image (a) Fourier power spectrum of image (b) log-polar average filter of power spectrum (c) log-polar mapped image of power spectrum (d) Fourier power spectrum of image

The MF lipreading, which precedes a Fourier transform with the log-polar mapping to the original image. Figure 5 shows the processing orders. In results, the MF lipreading is robust against rotation and scaling better than others.

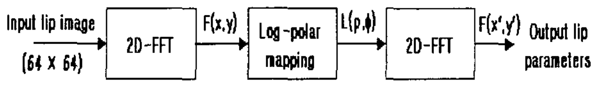


Fig 4. The processing orders of FM lipreading to obtain lip parameters

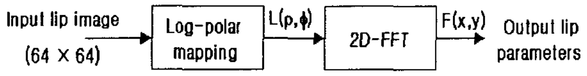


Fig 5. The processing orders of MF lipreading to obtain lip parameters

Figure 3 is the Illustration of the lipreading with log-polar mapping and 2D FFT. Left illustration is the processing of the MF lipreading and right is the processing of the FM lipreading in a sample image.

3. Related Work

The lipreading performance is influenced according an influence to database construction and feature extracting methods. Besides there are many factors, i.e. video quality, lighting variance, and so on. In this paper, we concentrate on the RTS variances of speaker's mouth area.

In order to estimate the stability of lipreading, all images from the database where made to vary constantly along the standard we suggested. Table 1 shows testing scenarios for RTS transformation by stages. Translation variance is the diagonal movements and is proportional to lip's width.

Table 1. Three parts of the visual speech database considered in this paper. The degree of variance divided into three or four levels. Test set sizes are shown as number of words multiplied by speakers (18x22).

Variance part	Degree	Test set
Rotation	5°	18×22×nfw (nfw : the number of frames in a word)
	10°	
	15°	
Translation	3%	
	6%	
	9%	
Scaling	×0.8	
	×0.9	
	×1.1	
	×1.2	

4. Experimental Results

We have used both of the visual feature extraction methods described here to build visual speech recognizers using hidden Markov models. All classification used HMM's each state, from left to right, that is associated with a one or more Gaussian densities with diagonal covariance matrix. In all cases we extensively tested HMM parameters for number of states and number of Gaussian mixtures per state.

Our lipreading system uses HMM algorithm as a means of statistical pattern matching. We use context independent, complete word, 4-8 states with 4-6 mixtures per state. HMM parameters are estimated by maximum likelihood Viterbi training.

Our database consists of Grey-level image sequences of the 22 words, spoken by 70 male. The images contain only mouth area and are digitized at 30 frames/sec, 320x240 pixels, 8 bits per pixel. In our experiment, we choose 22 Korean words utilized to browse information.

We have trained a speaker independent recognizer on 1144 sequences of visual data, and tested 396 sequences.

Figure 6 shows the recognition results of word recognition on each of the lipreading methods. There is high score at the DCT lipreading, up to 62.4%. The next is the MF lipreading, up to 51.5%. The FM lipreading is up to 44.2%. This experiment does not included RTS variance.

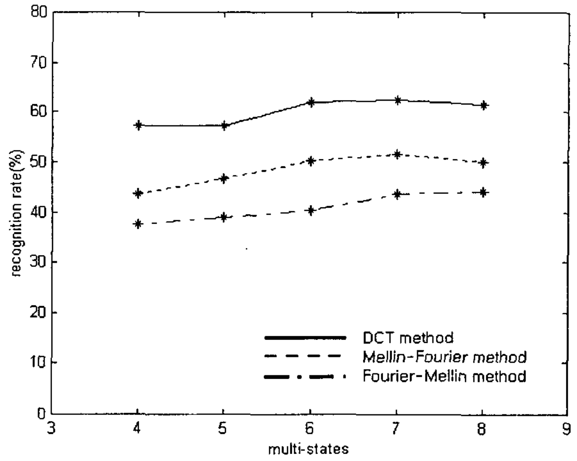


Fig 6. The recognition results on each of the lipreading methods

Figure 6,7 and 8 show the lipreading performance including the RTS variance. RTS-transformed image is included in tested set by stages, as described in Section 3. Figure 6 is the results of word recognition where rotation variance is applied. The MF lipreading performance is more stable than others against rotation. At figure 7, the FM lipreading performance is better against translation. And, at scaling, the MF lipreading is better. Following the results, the MF lipreading is robust at rotation and scaling transform, and the FM lipreading has some translation-invariant.

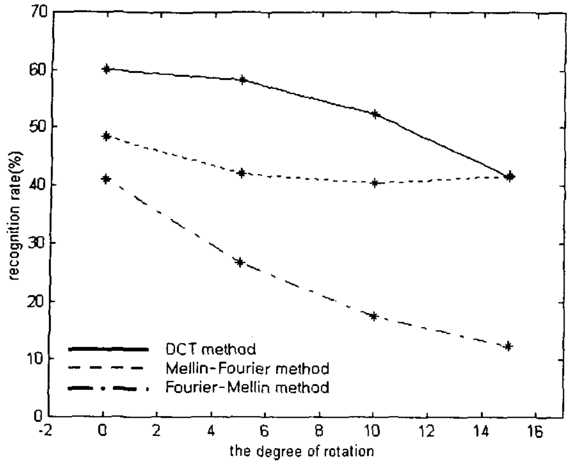


Fig 7. The recognition results with rotation variance

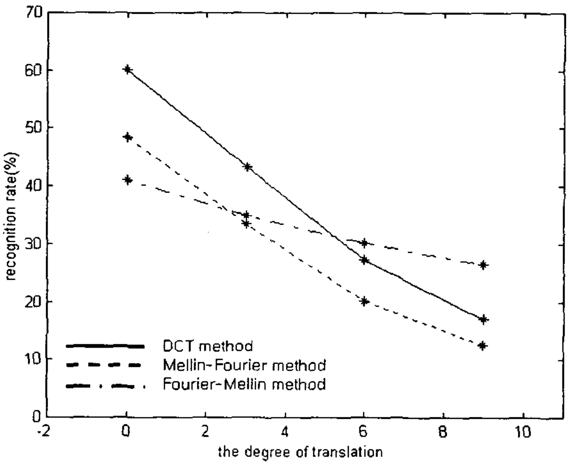


Fig 8.The recognition results with translation variance

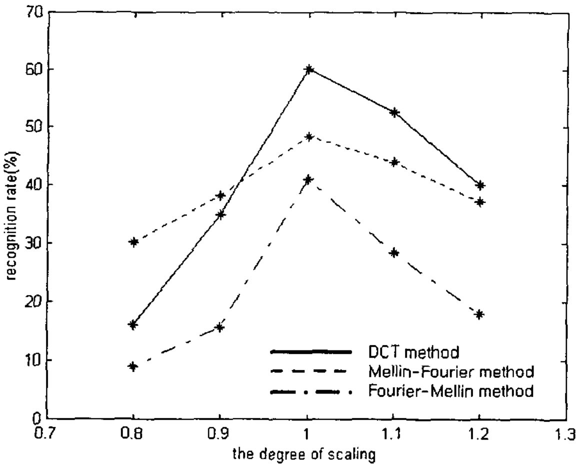


Fig 9. The recognition results with scaling variance

5. Conclusions

In this paper we presented the result of ascertaining the robustness of lipreading against the image variance such as *rotation, translation and scaling.*

We have compared three lipreading methods of visual feature analysis. This method requires the computation of the Fourier transform and DCT, performs a coordinate mapping from cartesian to log-polar mapping *and* executes the 2D image transformation. The choice of a log-polar mapping and Fourier transform is close to the characteristics of the special invariance. Thus these transform bear great significance for lipreading performance against RTS variance.

Future work will improve the highest lipreading performance of illumination variance. We will also focus on the pixel based lipreading regarding all circumstances.

References

[1] S. Startchick, R. Milaness and T. Pun,"Projective and photometric invariant representation of planar disjoint shapes". Image and Vision Computing Journal, accepted for publication, 1998.

[2] Sheng, A.Y., Lejeune, C., and Arsenault, H.H, "Frequency-Domain Fourier-Mellin Descriptors for Invariant Pattern Recognition", OptEng(27), No. 5, pp.354-357. BibRef 8800, 1988.

[3] D. Asselin and H. H. Aresenault, "Rotation and scale invariance with polar and log-polar coordinate transformations", Optics Communications, vol.104, pp. 391-404, jan. 1994.

[4] G. Engel ,D. Greve and E. Schwartz, "Space-variant active vision and visually guided robotics", ICPR pp. 487-490., 1994.

[5] Ruggero Milanese, "Invariant content-based image retrieval using the Fourier-Mellin Transform", ICAPR'98.

[6] Lièvin M. and Luthon F. "Lip features automatic extraction", Proc. Of the 5th IEEE Int. Conf. On Image Processing. Chicago. Illinois, 1998.

[7] Rajeev Sharma, Vladimir I. Pavlovic, Thomas S, Huang, "Toward Multimodal Human-Computer Interface ", Proceedings of the IEEE Vol. 86. No. 5., pp. 853-869May 1998.

[8] Gerasimos Potamianosm, Hans peter Graf, Eric Cosatto, "An Image Transform Approach for HMM based Automatic Lipreading ", Processing Of the Int. Conf. On Image Processing. pp. 173-177, 1998.

[9] T. Chen, H. P. Graf, and K. Wang, "Lip-synchronization using speech- assisted video processing", IEEE Signal Processing Lett., vol 2, pp. 57-59, 1995.