AUTOMATIC LIPREADING WITH PRINCIPLE COMPONENT ANALYSIS

Zafer Yavuz[1], Vasif Nabiyev[2] (*)

Karadeniz Technical University, Trabzon, Turkey, {[1]*zaferyavuz*,[2]*vasif*}*@ktu.edu.tr*

Abstract

Recent studies show that not only audio but also visual signs include information related with speech. This feature as supplementary information could be used to increase the performance of speech recognition. In this paper, lip movements are examined, and an automatic lipreading system is implemented. 56 different videos including Turkish vowels (a, e, ı, i, o, ö, u, ü) and digits (bir, iki, üç, dört, beş, altı, yedi, sekiz, dokuz, on) were recorded as input data set. Automatic lipreading system based on Principle Component Analysis method was realized by using input images. In the test phase, we obtained about 60% of success in condition that similar vowels (o-ö, ı-i, u-ü) are taken in to account.

1. Introduction

Speech Recognition Systems in human-computer interaction have attracted more and more researchers and become an important subject in recent years [1]. Extracting the audio information of speaker and processing this information in computers bring about complex computing processes. This complexity constitutes an important problem in speech recognition. According to the studies realized so far, information related with speech is included not only in audio signs but also in visual signs [2, 3, 4, 5]. *Sign language, facial expressions, gestures, and lip movements* are considered as visual signs. Visual information as supplementary information to the audio information increases the performance of speech recognition [1]. Assessing the lip movements is also extremely important for hearing disabled people to understand what is spoken. Evaluation of lip movements using computer is studied in this paper. This operation is called *automatic lip reading* in human-computer interaction.

Automatic lipreading problems are examined in two contexts. These are generally speech recognition and analysis of sign language [6]. Speech recognition, which is one of the most important components for human communication, is also one of the basic elements in human-computer interaction. Systems based on only audio waves are not very safe because of the fact that the signs are affected by different noises drastically. Therefore, taking in to consideration audio signals and video information (lip movements) might make speech recognition much easier. It is shown that communication using both visual and audio information is much more powerful than communication using only audio information. It is also pointed out that visual information constitutes 1/3 of the conveyed message [3, 4, 5].

Many Automatic Speech Recognition systems focused on acoustic audio signals and consequently may easily be affected by the noises in the environment. Also, while some audios may be confused in audio space, these audios can easily be differentiated from each other in visual space [2]. Therefore, accepting visual information as supplementary information in automatic speech recognition systems directed the attention on automatic lipreading systems.

2. Automatic Lipreading System

Automatic lipreading is the operation of understanding what a person says from his image without any need of audio information. This operation involves many complex computational processes. Therefore the proposed model is divided into sub modules, and every sub module is examined and realized separately.

In the study, the first frame of the image set is taken, and face in the image and lip area in the face is determined. Later on, points representing the outer contour of the lips are found in order to form the feature vector. A feature matrix is created by realizing the operations of lip

(*)This study was supported by Karadeniz Technical University Research and Development Fund for the Science, Project N 2005.112.009.01

determination and feature vector extraction for the following frames. This matrix is inserted into the database after training it with Principle Component Analysis (PCA) method. When an image is assessed for automatic lip reading, the same process is repeated for that image, and the created new feature vector is compared with the database and a decision is made on what that person says. Sub modules of the proposed model are mentioned in the following.

In this paper, in order to find the face in the image, skin color determination algorithms are employed. For that aim HSV color space is used. Firstly, vertical location of the eye is found in order to determine the lip area, and then geometric ratios for the face are exploited to find the lips [6].

A method based on dynamic tresholding is employed after the images are pre-processed to find the lip area. According to this method, pixels on the lips are strengthened by transformation of color components. After that, binary lips are obtained by applying dynamic tresholding on the strengthened lips. Pixels that belong to lips on a color image are strengthened by conducting a transformation according to $I=0.5G-0.25R-0.125B$ where the R, G and B are Red, Green and Blue values.

Feature vector is produced from the obtained binary image. The most important operation in automatic lip reading systems is the extraction of feature vector for the determined lip. The properties used for automatic lip reading are on the other hand, *bottom and upper lip location*, *left and right lip location*, *the gap between lips*, *visibility of bottom and upper teeth*, *the size of the shape of lips, the center of gravity of the shape of lips, the length and width of the shape of lips, the coordinates of 14 points representing the outer contour of lips*. The feature vector is produced by finding the mentioned properties one by one. The process of extracting the feature vector is illustrated in Figure 1a.

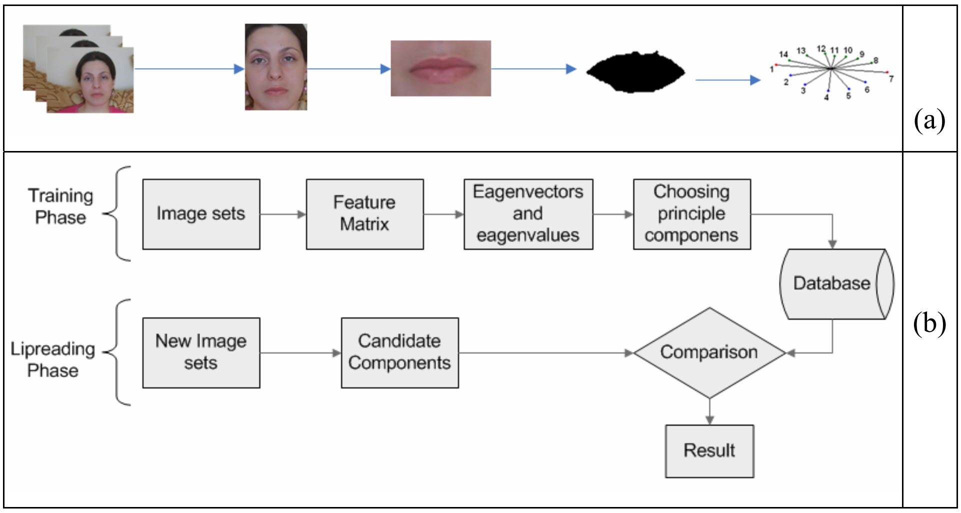

Figure 1. (a). Extracting feature vector. (b). The block diagram of the system

2.1. Lipreading by Using PCA

PCA is a useful statistical technique that could be applied in fields such as face recognition and image compression, and is a common technique for finding patterns in high dimension data. Besides, it is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences [7]. Because PCA is an unsupervised learning method, it may be used for lipreading problem as well. The working schema of PCA is shown in Figure 1b. The PCA method composed of two sub modules which are training and lipreading recognition phases.

2.1.1. Training Phase

The training phase is recording the data composed for training by unsurprised learning. The image sets are still images taken before. These image sets are normalized to k frames to provide identical structures. In this paper, the constant *k* is taken as 16. The feature vector consists of the features mentioned above. In this phase, to apply feature vector to PCA method, it should be reorganized. Each row in the PCA feature matrix corresponds to a P_{fm} feature matrix. Thus, the training matrix is composed for PCA by transforming the features matrix of each pattern to a row vector. The feature matrix for p^{th} pattern and the feature matrix used for PCA is shown in Equation (1).

$$P_{fm_i} = \left[1_f, 2_f, ..., k_f\right]^T \;, \qquad PCA_{fm} = \left[P_{fm_1}, P_{fm_2}, ..., P_{fm_l}\right]^T \qquad (1)$$

where P_{fm_i} is the feature matrix of p^{th} pattern, PCA_{fm} is the matrix used in PCA and l is the pattern number used for training. After that, we subtract the mean value of each row from each element in that row of the PCA_{fm} to construct the matrix *A* and the covariance of matrix *A* is calculated as seen in group of Equation (2).

$$\overline{X} = \frac{\sum_{i=i}^{n} X_i}{n}, \qquad A_k = X_{k_i} - \overline{X} \;, \qquad C = \frac{A^T.A}{(n-1)} \;, \qquad k = 1,2,...,l \qquad (2)$$

where *C* is the covariance matrix used in PCA. Then, the eagenvalues and eagenvectors of the covariance matrix are calculated. The eagenvalues are ordered and then corresponding eagenvectors are sorted out according to the eageanvalues. The biggest value of eagenvalues is the first element of the set. Smaller values of eagenvalues and the corresponding eagenvectors are eliminated, so that the dimension of the data set is reduced. The matrix consisting of the remaining eagenvectors is called transformation matrix *U*.

The signature is obtained by multiplying the transformation matrix *U* and corresponding row of the of the feature matrix, because each row vector in the feature matrix formed by original values is associated with a pattern [7]. The relation for this operation is given in Equation (3).

$$S_k = I_k \text{x} U \qquad (3)$$

where *I* is a row vector obtained from transformation matrix with dimension *l*, *U* is a *l*x*n* transformation matrix and *S* is resulting vector. Additionally, *l* is the number of features, *n* is the new dimension resulted by reduction done on the phase of selection of principle components.

The last phase of PCA is storing signatures into the database. All of the signatures produced for each pattern are stored into the database in the phase of signature production. However, transformation matrix *U* used in the recognition phase is stored into the database as well.

2.1.2. Lipreading Phase

In this phase, the recognition of the data set which is trained before is implemented. For this, firstly we obtain the feature vector from the new data set. This process (extract the feature vector) is similar with the training phase. Each row vector in the extracted feature vector is produced by transformation matrix U which is obtained and stored into the database before. After that the candidate signature is produced. The production of candidate signature *S'* is derived from the equation $S' = I_k'$ x *U*. After the production of the candidate signature, it is compared with all the other signatures in the database. The most similar signatures are accepted as the same after comparison.

3. Conclusions

The training and testing set used for automatic lipreading system consists of still images. Turkish vowels *(a,e,ı,i,o,ö,u,ü)* and Turkish digits *(bir<one>, iki<two>, üç<three>, dört<four>, beş<five>, altı<six>, yedi<seven>, sekiz<eight>, dokuz<nine>, on<ten>)* are used in automatic lipreading system. The system is trained on 32 patterns and tested on 24 patterns. The results obtained from vowels and digits are given in Table 1 and Table 2.

Experiments show that Turkish vowels *ı* and *i* are recognized mostly as shown in Table1. Additionally, if the vowels *(ı-i), (o-ö)* and *(u-ü)* are considered as similar ones, because they are almost the same in the visual space, the recognition ratio increases even more.

According to Table 2, *üç<three>, dört<four>, beş<five>, altı<six>, yedi<seven>* and *on<ten>* digits are recognized in rates changing between 50% and 80%. If we calculate a mean recognition ratio, we get about 59% of success. We think the basic reason of being more successful in recognizing digits than the vowels is that the digits have more syllable than vowels. However, if we assume *(iki<two>, yedi<seven>, sekiz<eight>)* and *(dört<four>, dokuz<nine>, on<ten>)* are very similar in visual space, we can cluster these digits. We have also done some experiments on different people. We played videos without any audio information telling two vowels *(o, u)* and two digits *(on<ten>, dört<four>)*, and asked people to guess about what those digits and vowels are. We informed the people about the part of video where vowels and digits are played. So, the state space is known. The decision made by human in lipreading without any audio information doesn't give any satisfactory results which are 48% for *o*, 50% for *u*, 56% for *on<ten>,* 70% for *dört<four>*.

Table 1. Results of vowels

Vowels	Recognition Results	
	First nearest	Second nearest
A	I (60%)	İ (20%)
E	E (20%)	I (80%)
İ	İ (80%)	I (20%)
I	I (80%)	İ (10%)
O	U (60%)	I (20%)
Ö	I (60%)	Ö (20%)
U	Ü (20%)	I (60%)
Ü	U (40%)	Ö (20%)

Table 2. Results of digits

Digits	Recognition Results	
	First nearest	Second nearest
Bir <one>	Yedi (67%)	Beş (33%)
İki <two>	Sekiz (50%)	Yedi (50%)
Üç <three>	Üç (50%)	Yedi (50%)
Dört <four>	Dört (50%)	Yedi (50%)
Beş <five>	Beş (67%)	Yedi (33%)
Altı <six>	Altı (50%)	Yedi (50%)
Yedi <seven>	Yedi (80%)	İki (20%)
Sekiz <eight>	Yedi (100%)	-
Dokuz <nine>	Yedi (67%)	Dokuz (33%)
On <ten>	On (80%)	Yedi (20%)

The selected patterns in human-computer comparison have lower recognition ratio. In spite of this, we achieved about 60% of success in automatic lipreading with computer. So, it can be concluded that if we take visual information addition to audio, we get more robust and successful results in human-computer interaction. Since Turkish and Azerbaijani is relative languages, the study could be applied on Azerbaijani in a future work.

Literature

[1] V.V. Nabiyev, *Artificial Intelligence: Problems-Methods-Algorithms* (in Turkish), Seckin Publishing, 2nd Press, (2005).

[2] I. Matthews, T.Cootes, J.Bangham, S.Cox, R. Harvey, *Extraction of Visual Features for Lipreading*. PA&MI, IEEE Transaction, 198-213, (2002).

[3] A.Rogozan, P.Deléglise, Visible Speech Modeling and Hybrid Hidden Markov Models / Neural Networks Based Learning for Lipreading. IEEE Computer Society, France (1998).

[4] L.Xie, X.Cai, Z.Fu, R. Zhoa, D.Jiang. *A* Robust *Hierarchical Lip Tracking Approach for Lipreading and Audio Visual Speech Recognition.* Proceedings of the 3rd ICMLC, 3620-3624, Shangai, China, (2004).

[5] S.Wamg, H.Lau, S.Leng, H.Yan. *A Real Time Automatic Lipreading System.* ISCAS'04, vol:2, p.101-104, Hong Kong, (2004).

[6] Z.Yavuz, V.V.Nabiyev. *Automatic Lipreading*, 15th SIU'07, Eskişehir, (2007).

[7] I. S. Lindsay. A Tutorial on Principal Components Analysis, USA, (2002).