

Visual Speech Recognition Through Fuzzy Set Theory

James F. Baldwin, Trevor P. Martin, Mehreen Saeed

Abstract

This paper describes an approach based on fuzzy set theory towards the various aspects of visual speech recognition. Automatic lip-reading or speech-reading is one of the growing fields of interest amongst the speech recognition community because it can greatly enhance the quality of speech recognition programs, especially in the presence of environmental noise. Visual speech data is obtained as the video of the lip movements of a speaker. In this work novel methods for representing speech data, segmenting and efficiently storing and searching mechanisms that lead to isolated word recognition are proposed. The classification methods were applied to Tulips1 database and the results obtained are slightly better than the ones obtained with techniques based on neural networks and Hidden Markov Models. Encouraged by the results of Tulips1 database, a medium sized vocabulary of around 300 words was developed to assess the recognition methods based on fuzzy set theory. Because of the ambiguity and similarity of various speech sounds a scheme was developed to select a group of words when a test word was presented to the system. The accuracy achieved in this case was 33% which is comparable to expert human lip-readers whose accuracy on nonsense words is about 30%.

Keywords

Cross Product Space Approach, Cartesian Granules, Isolated Word Recognition, Mass Assignment, Speechreading, Segmentation, Storage of Vocabulary

I. INTRODUCTION

Visual speech recognition or automatic computer lip-reading is catching the attention of many researchers working in the field of speech recognition. Current technology has made a great progress in the field of automatic acoustic speech recognition, however the quality of these systems degrades considerably in the presence of noise. Environmental noise has become one of the major obstacles in the commercial use of speech recognition techniques [11].

Study of visual and acoustic speech signals has revealed that the information regarding speech contained in visual signals is both supplementary and complementary to the information contained in audio signals, especially in the presence of noise [26]. Sounds difficult to distinguish in audio signals are easy to discriminate in visual signals and vice versa. An example of this phenomenon are the phonemes b/k which are difficult to distinguish when only their audio signals are present but easy to discriminate on the basis of their lip movements only. The contrary is true for the sounds of the phonemes p/b/m which have similar lip movements but different acoustic spectrums. Hence visual signals provide

information which is acoustically sensitive to noise. It appears that seeing complements hearing for just the sounds that need it. Some researchers e.g. [23], [24], [28] etc. have developed speech recognition programs that incorporate computer lip-reading in their audio recognition programs and their systems demonstrate a considerable improvement over systems employing acoustic signals only.

Systems for automatic computer lip-reading can be used to build aids for people with hearing difficulties. Educational packages can be built to assist teachers for the teaching and training of hard of hearing children. Visual speech recognition can also find its application in video conferencing. Data at one end can be compressed using the information of the speaker's lip movements and partial information can be transferred to the other end. Study of lip-reading is used in simulating speech and talking faces in computer graphics.

In this paper, we explore the possibility of using fuzzy set theory towards various aspects of visual speech recognition, to be incorporated into any of the systems described previously. The recognition is based on purely visual information of the lip images of a speaker and no audio information has been taken into account. Classification algorithms based on normalising time using the cross product space approach have been proposed and segmentation algorithms using fuzzy rules have been formulated. Efficient storage of the vocabulary of words, that incorporates within its structure context information has also been used.

II. SPEECHREADING

The ears and the eyes provide a dual channel that plays a key role in the perception and understanding of speech. Individuals with hearing loss rely on visual lip movements of a speaker to infer speech. A deaf person depends on his eyes for 80 to 100 percent of the received information [13]. Even humans with normal hearing use lip-reading as a supplement to bring together bits and pieces of information regarding speech, especially in a noisy environment.

The term 'speechreading' is broadly used to describe the interpretation of speech by the study of lip movements, facial cues and gestures and the recognition of patterns dictated by the common use of a language. The facial cues include the tongue, jaw and eye movements as well as facial expressions of a speaker. The patterns in a language include the knowledge

of vocabulary, grammar and the commonly used expressions in speech. Hard of hearing individuals generally construct sentences by visually recognising a few key words from a string of words. Their knowledge of the syntax and semantics of the language and also of the common and idiomatic expressions of the language helps them in the prediction of speech. Expert speechreaders making use of all these visual and literary cues can achieve almost perfect recognition and perception of speech [27].

A. Challenges in Lip-reading

Speech sounds are produced by modulations occurring in a vibrating or obstructed air stream in the vocal tract [1]. Many muscles or organs of speech such as the vocal cords or velum are inside the mouth and are not visible to the eye. Lip movements play a relatively minor part in the production of sound [13]. Therefore a speechreader, whether man or machine has only the lips, jaws and the occasionally visible tongue to guide him for inferring speech.

Another difficulty encountered by human and machine lip-readers alike is that many sounds such as those of t, d, n, l etc. do not require a prominent movement of the lips or the jaws. It is estimated that under usual viewing conditions approximately 60 percent of the speech sounds are either obscure or invisible [13]. There is confusion over half of the vowels and diphthongs and three fifths of the consonants.

Another important factor that renders speech recognition a very challenging job is that a particular lip movement is generally common to many phonemes, hence making them visually indistinguishable. An example of such phonemes is /p,b,m/ in which lips come together or /f,v/ the lower lip to upper teeth movement etc. Such sounds are grouped together into **visemes**, which are the representative units of visual speech and are the equivalent of phonemes in acoustic signals of speech. Generally speech experts put all the consonants of English language in a group of 4 to 12 speechreading movements.

A major difficulty encountered by a lip-reader in recognising speech sounds or syllables is because of the alteration of the movements of a sound by those that follow or precede it. A phoneme does not have the same appearance in every word.

One of the greater challenges in visual speech recognition tasks is to generalise speech movements over different speakers. There can be different movements that lead to the

production of the same phoneme. Different people may use different articulation patterns but may produce the same sounds [13]. For example /t/ can be produced by putting the tongue in different positions.

The modelling of speech is a tough and complicated problem. Machines have huge limitations when compared to expert human speech-readers. Human lip-readers not only study the lip movements but they also make use of the facial expressions, jaw movements or gestures as a clue to infer speech. Their knowledge of vocabulary and grammar and the common and idiomatic expressions used in the language, helps them enhance their ability to predict speech from visual movements only.

B. Past Lip-reading Systems

In this section some of the speechreading systems developed in the past are discussed. Automatic lip-reading is performed by taking the video of the lip movements of a speaker. The video is a sequence of images or frames played in succession. Generally there are 25-50 frames in one second. The approach followed towards this problem in the past is broadly classified as an image based approach or a model based approach [15]. In a model based approach explicit features of the mouth are extracted from video images and used for recognition. Generally in this approach exact measurements of the lip area are derived from the visual database. On the other hand in an image based approach the entire raw image or processed image is fed as input into a recogniser.

Both the model based approaches and the image based approaches have their advantages and shortcomings. Model based approaches are resistant to rotation or scaling of images, and also the illumination or the various lighting conditions under which the video of a speaker has been taken. However, a problem with this approach is that, since the whole image is not used in the recognition task, some information is lost in feature extraction. Moreover, this approach can lead to poor performance if the right features are not extracted from the database. It is also a difficult task to build up an accurate and robust system for accurately measuring the features from a set of visual images.

On the other hand, image based approaches have an advantage over model based approaches that they do not discard any information. The capability of analysing the changes in displacement of the various facial cues such as skin and wrinkles can be embedded in

such systems. However, a drawback with this method is that it leads to a huge dimensionality of the feature vector which may have some redundant information.

The first major work on automatic lip-reading was done by Petajan in 1984 [22]. He used an approach where contour information of the oral cavity from image sequences was used to perform speech recognition. He used linear time warping to perform template matching in this approach. Petajan's system was further extended by Goldschen [10] who developed an impressive visual only continuous speech recognition system. His system achieved an accuracy of 25% on sentences.

Finn and Montgomery [9] studied optically based speech recognition for consonants of the English language by gathering data by from a male speaker who had 12 highly reflective dots placed on his face. Mase and Pentland in [16] adopted optical flow techniques to lip-reading. Yuhas *et al* [29], [30] used an image based approach towards lip-reading and fed an entire image into a neural network for the recognition of speech sounds. Movellan [19] used a technique based on diffusion networks to perform visual speech recognition. Silsbee and Bovik [24], [25] extracted features from visual speech signals by using vector quantisation. Different mouth configurations, defined by 17 code vectors, were selected by hand. The work done by Luetttin and Thacker [14], [15] involved learning patterns of shape variability for tracking lips in gray scale video images. The extracted features were modelled by Gaussian distributions and their temporal dependencies by Hidden Markov Models.

In the past, researchers have developed optical speech recognition systems. However, these systems have only been tested by small vocabularies because of the limitations imposed by the bulk of video data. Some of the systems developed in the past that make use of a big vocabulary system containing whole words are the ones developed by [10] and [24], [25]. As opposed to most of the previous work, the work presented in this paper was not only tested on smaller databases but also on a bigger vocabulary of 310 isolated words.

C. Why Fuzzy Set Theory?

Visual speech data contains partial information regarding speech. There is ambiguity and anomaly between various sounds and phonemes. Fuzzy sets are very appropriate for modelling such data because they represent propositions that are neither entirely true nor

entirely false. When a visual speech unit, possessing several properties of belonging to different phonemes, is encountered, fuzzy set theory can provide a tool for explaining the ambiguous nature of data. Different degrees of membership values can be assigned to this speech unit indicating the degree with which it belongs to various phonemes and hence its possibility of being categorised in those classes is not ignored, but taken into account.

The use of fuzzy logic in explaining various aspects of auditory visual speech is not new. Massaro [17], while describing the psychological aspects of lip-reading, used fuzzy logic to model human perception of speech. This model matched well with children and adult subjects. A fuzzy logic model was also used by Silsbee [24] to integrate audio and video information.

The aforementioned examples show that fuzzy logic provides an appropriate method for modelling the perception of visual speech and also for the automatic integration of audio and video sources of information to recognise speech units. We would like to explore fuzzy set theory towards yet another aspect of speech analysis i.e. automatic recognition of speech. Fuzzy sets possess some very desirable properties like generalising ability, interpolation and compact representation.

Apart from providing a suitable and appropriate mechanism for representing and explaining speech units, fuzzy set theory allows the modelling of data using simple and linguistic terms such as *small*, *medium*, *large* etc. These terms are not only easy to implement but they are also comprehensible lending themselves to transparent knowledge representation. Moreover, in this work fuzzy sets have been generated from the probability distributions of data. Therefore, a generated fuzzy set is actually based on the trend and nature of data and provides a compact way of representing it on the whole.

III. FUZZY SETS AND MASS ASSIGNMENTS

Before a discussion of various issues on visual speech recognition is carried out, it is necessary to introduce the reader to the theory of mass assignments [2], [4]. The theory of mass assignments bridges the gap between probabilistic uncertainties and possibilistic uncertainties and we use mass assignments to transform probabilistic data into fuzzy sets. Formally, a mass assignment over a finite frame of discernment X is a function m ,

$m : P(X) \rightarrow [0, 1]$ where $P(X)$ is the power set of X such that:

$$\sum_{A \in P(X)} m(A) = 1; \forall A, m(A) \geq 0 \text{ and } m(\phi) \geq 0$$

A mass assignment represents a family of distributions $\{FD(x_1), \dots, FD(x_n)\}$ over X where

$$m(\{x_i\}) \leq FD(x_i) \leq \sum_{A=\{x_i\} \cup Y} m(A)$$

with the constraint

$$\sum_i FD(x_i) = 1 - m(\phi); \forall x_i \in X$$

if there is no mass assigned to the null set then this family is equivalent to a family of probability distributions over X .

The least prejudiced distribution is a special type of distribution obtained by distributing the masses equally within all elements in a particular subset. The transformation to a **least prejudiced distribution (LPD)** is therefore reversible. The LPD transformation distributes masses according to the prior; if an equally likely prior is used then the masses are distributed equally. Mass assignments are related to fuzzy sets via the voting model [3]. Formally, let $X = \{x_1, x_2, \dots, x_n\}$ be the universe and let A_1, A_2, \dots, A_n be the nested subsets of X such that $A_1 \subset A_2 \cdots \subset A_n$ where $A_i = \{x_1, x_2, \dots, x_i\}$. Let the fuzzy set f be defined as follows:

$$f = x_1/\mu_1 + x_2/\mu_2 + \cdots + x_n/\mu_n$$

where $\mu_1 = 1$ and $\mu_1 > \mu_2 > \cdots > \mu_n$

The possibility distribution induced by f is given by $\pi_f(x_i) = \mu_i$. The associated mass assignments over the nested sets $\{A_i\}$ are given by:

$$m\{A_1\} = 1 - \mu_2; \dots ; m\{A_i\} = \mu_i - \mu_{i+1}; \dots ; m\{A_n\} = \mu_n$$

IV. CROSS PRODUCT SPACE APPROACH FOR REPRESENTING TIME VARYING DATA

In this section we introduce a novel way of representing speech data in time. This method is based on the use of compound words in cartesian space and extended Fril rules derived from them.

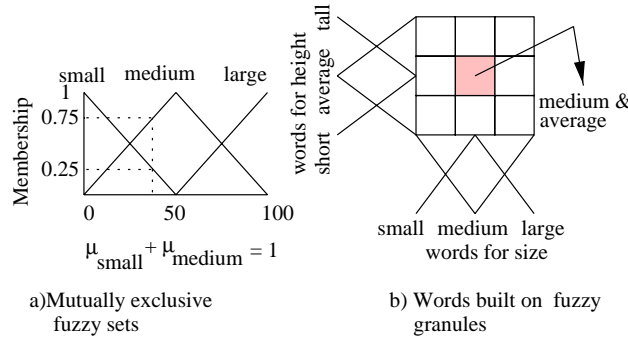


Fig. 1. Words modeled by fuzzy sets

A. Grid Built on Fuzzy Granules in Cartesian Space and Extended Fril Rules

According to Zadeh a word or a linguistic term can be represented by a fuzzy set of points representing a clump of elements drawn together by similarity [32]. Figure 1 shows 3 fuzzy sets representing the words 'small', 'medium' and 'large'. The fuzzy sets of figure 1 are termed as mutually exclusive triangular fuzzy sets. By being mutually exclusive it is meant that they partition the space over the universe in such a manner that the sum of memberships of a point in all the fuzzy sets is equal to one. Such fuzzy sets can be viewed as basis functions forming a partition of unity. Any point on the axis has a nonzero membership in at the most two fuzzy sets. Each triangular fuzzy set is thus a word (granule or a label).

Zadeh [31], [32] also introduced the concept of a word in multi dimensions. An expression of the form $A \times B$, where A and B are words, is referred to as a cartesian granule. 'x' denotes the cartesian product and the words A and B are represented by fuzzy sets which can be defined over different universes. Fuzzy sets can be constructed over cartesian granules and hence they represent a cross product of two features. In figure 1 a two dimensional grid in cartesian space with mutually exclusive fuzzy sets on its two axis has been built. Each cell models a compound word or the cross product of two linguistic labels. So for example in figure 1 the shaded cell represents the situation when the height feature is *average* and the size feature is *medium*.

In this work, a simple counting procedure based on the theory of mass assignments has been adopted for constructing the grid from example points in the data set [6]. A data

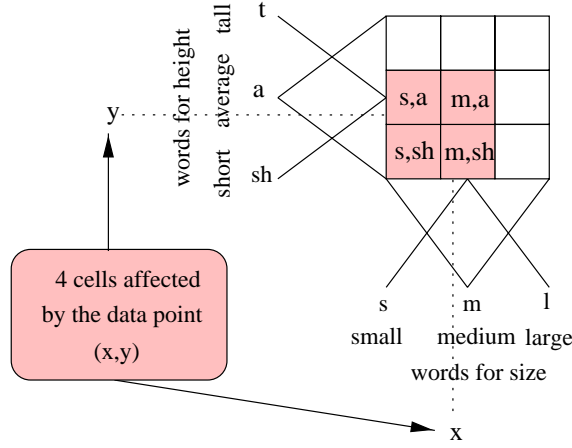


Fig. 2. Counting procedure for filling the grid

point (x, y) shown in figure 2 can be written in terms of fuzzy sets (g_x, g_y) , each fuzzy set being a value membership pair:

$$(g_x = \frac{f_s}{\mu_s} + \frac{f_m}{\mu_m} \quad , \quad g_y = \frac{f_{sh}}{\mu_{sh}} + \frac{f_a}{\mu_a})$$

where μ_i is the membership of the point in fuzzy set f_i . By making use of the theory of mass assignments and the voting model it can be easily shown that the least prejudiced distribution $lpd_x(f)$ of the fuzzy set f is equal to the membership of x in f . In this case the least prejudiced distribution represents the conditional probability of f given the point x , hence the 4 pair of values:

$$\{ lpd_x(s)lpd_y(sh) \quad ; \quad lpd_x(m)lpd_y(sh); \\ lpd_x(s)lpd_y(a) \quad ; \quad lpd_x(m)lpd_y(a) \}$$

are equivalent to:

$$\{ \mu_s \mu_{sh} \quad ; \quad \mu_m \mu_{sh} \quad ; \quad \mu_s \mu_a \quad ; \quad \mu_m \mu_a \}$$

and can be associated with their corresponding cells $\{s,sh; m,sh; s,a; m,a\}$ respectively in figure 2. Thus, for a given data tuple, the membership of feature 1 and feature 2 in their corresponding fuzzy sets is determined and the relevant cell is incremented by the product of memberships in the two fuzzy sets.

Clearly there is an advantage of using fuzzy granules to crisp sets. If crisp sets are used then the membership of only one cell is affected whereas in this case several cells

are affected depending upon their membership in various words. This phenomenon is illustrated in figure 2 where the point (x,y) affects the four cells {s,sh; m,sh; s,a; m,a}.

After filling each cell of the grid in cartesian space, each cell in the grid is divided by the total number of entries and a probability distribution θ_{ik} over the cells is obtained for a given class k . Extended Fril rules based on Bayesian theory are used to derive conditional probabilities. So the conditional probability of a $class_k$ given $cell_i$ i.e. $Pr(class_k|cell_i)$ is given by:

$$\begin{aligned} Pr(class_k|cell_i) &= \frac{Pr(cell_i|class_k)Pr(class_k)}{Pr(cell_i)} \\ &= \frac{\theta_{ik}Pr(class_k)}{\sum_j Pr(cell_i|class_j)Pr(class_j)} \\ &= \frac{\theta_{ik}Pr(class_k)}{\sum_j \theta_{ij}Pr(class_j)} \end{aligned}$$

Assuming that all classes are equally likely the above form reduces to:

$$Pr(class_k|cell_i) = \frac{\theta_{ik}}{\sum_j \theta_{ij}}$$

When an example point from the test set is encountered then a test grid is formed and the probability of each cell e_i is determined for the features in the test data point. The $support_{jk}$ for $class_j$ for a single grid which represents a single feature k is given by:

$$support_{jk} = \sum_i e_i Pr(class_j|cell_i)$$

The over all support for $class_j$ using m grids representing m different features is averaged as:

$$support_j = \frac{\sum_{i=1}^m support_{ji}}{m}$$

When classification is performed then the class with the highest support is the predicted class.

B. Generating Rules for Speech Data

In this paper a model based approach has been followed for visual speech recognition. Speech data for a word or a phoneme is acquired as a plot of feature values against time. This data is composed of sequences of different lengths. A phoneme comprising 15 frames

in one word can be composed of 20 frames in another word. Also, the duration of the utterance of consonants is much longer than that of a vowel. A method has to be devised to represent the variable sized data in time so that all sounds can be defined in a uniform way.

In this paper, a method based on the cross product space approach for representing multidimensional feature fuzzy sets is proposed. Since speech data is composed of sequences of feature values varying in time, each feature can be represented as a cross-product of time and feature values.

As can be seen from figure 3, one axis on the grid represents the feature space and the second axis represents time. The values on the time axis range from the first time frame to the last frame comprising the sound.

The number of mutually exclusive fuzzy sets placed on the time axis remains the same for every sound or word and hence they are made broader or narrower depending upon the number of sequences comprising the word. At each time step the membership of a feature in its corresponding feature fuzzy sets and time fuzzy sets is determined and a grid on the cross product space of words representing features and time is obtained. Now this grid can either be converted to a fuzzy set using the theory of mass assignments or it can be normalised with respect to various classes of sounds in an appropriate manner to derive extended Fril rules using Bayesian theorem.

It is apparent that the aforementioned scheme models the temporal characteristics of feature values in an effective and uniform manner. For example if extended Fril rules are derived for each class of sound, then this grid has very simple semantic meanings. In figure 3 it depicts the following situation:

- Probability of feature value being *small* in the *beginning* stage given class j is 0.1.
- Probability of feature value being *medium* in the *intermediate* stage given class j is 0.7.
- Probability of feature value being *large* in the *final* stage given class j is 0.5.

The framework for modelling a single sequence of sounds can be used in the generation

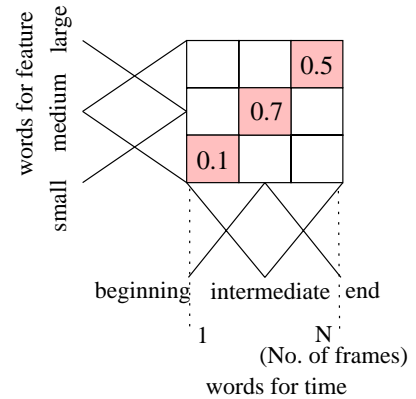


Fig. 3. Words for speech data

of a knowledge base of rules. Each viseme or a whole word can represent a class. Each class of sound is therefore represented by a number of features. When a test sequence of speech sounds is encountered, its support for each class is determined from the rule base. The class having the highest support is therefore the classified sound.

V. INITIAL FINDINGS : TULIPS1 DATABASE

A. *Initial Findings : Tulips1 Database*

To explore the performance of rules built on the cross product space of fuzzy granules towards automatic lip-reading, an experiment on a small database called the Tulips1 database was carried out. The Tulips1 database was compiled at Javier R. Movellan's laboratory at the Department of Cognitive Science, University of California, San Diego [19]. It was formed from 12 speakers, 9 male and 3 female, saying the words 'one', 'two', 'three' and 'four' twice. There were 934 gray-scale images of 100x75 pixel dimensions taken at 30 frames per second. The audio signals included in the database were not taken into account for this experiment.

For the Tulips1 Database images, the corners of the mouth and lips were marked by hand and the changes between each successive frame for six features were extracted [8]. The features used were the height and the width of the outer contours of the mouth, height and the width of the contours of the inner mouth, height of upper and lower lip (see section VI). Training was performed by generating rules from 11 speakers and leaving one out for testing. The process was repeated by including each speaker in the test set once and the results were averaged over all speakers. Since this database was prepared from words said by many speakers, the results from this experiment point to the generalising capabilities of the Fril extended rules.

Feature fuzzy sets	7	7	6
Time fuzzy sets	2	3	2
Percentage Accuracy	91.67%	91.67%	92.71%

TABLE I

VARIOUS RESULTS FOR TULIPS1 DATABASE USING EXTENDED FRIL RULES

For the Tulips1 database four extended Fril rules were generated for each word ‘one’, ‘two’, ‘three’ and ‘four’. Since rules were generated on whole words, no segmentation of the visual speech data was required. The results obtained for different parameters are shown in the table I. The first two rows indicate the number of fuzzy sets placed on the time and feature space.

Lip-reading by ...	Accuracy
Extended Fril rules	92.71%
Diffusion Networks [20]	91.7%
Hidden Markov Models [15]	90.6%
Humans without lip-reading knowledge	89.93%
Humans with lip-reading knowledge	95.49%

TABLE II

TABLE OF COMPARISON FOR TULIPS1 DATABASE

The results obtained from extended Fril rules are comparable and even slightly better than the results obtained by various other methods applied on the Tulips1 database. In [20] Movellan and Mineiro achieved an accuracy of 91.7% by training diffusion networks which are a stochastic version of recurrent neural networks. Luettin and Thacker [15] attained an accuracy of 90.6% on this database by training Hidden Markov Models on the 5 most discriminating features representing shape and intensity and also their delta parameters. When presented with images from the same database, humans with no lip-reading knowledge achieved an average of 89.93%, while hearing impaired people with knowledge of lip-reading obtained 95.49% [19] accuracy. The comparison is shown in table II. The tables I and II clearly illustrate the effectiveness of the extended Fril rules. The rules are speaker independent and hence robust enough to handle different speakers with different ethnic origins.

VI. DATABASE OF UNIVERSITY OF BRISTOL

Encouraged by the results of the Tulips1 database, a visual speech database was developed at the University of Bristol. The lip movement of a phoneme varies depending upon the phonemes that follow or precede it. Therefore, when training the system on various sounds, it is important to provide it with a sample of such sounds that occur in different contexts within a word. Keeping this point in view, a medium sized vocabulary of words was developed with a well balanced phonetic content [21]. There were a total of 302 distinct words in the database.

To form the visual database, the words in the vocabulary were said by a female speaker who does not have a strong accent in English. The video was taken in a well lit room. The camera was focused only on the mouth of the speaker. Video frame rate was 25 frames per second. The length of the video sequences for each word ranged from 11 to 37 frames. There were around 6000 coloured images, 250x160 pixels in size, occupying about 720M bytes of disk space.

A. Features Used

The features used for isolated word recognition are the following and illustrated in figure 4:

1. Width of lips
2. Height of lips
3. Width of inner mouth
4. Height of inner mouth
5. Height of lower lip
6. Height of upper lip
7. Ratio of width of lips to height of lips
8. Ratio of width of inner mouth to height of inner mouth

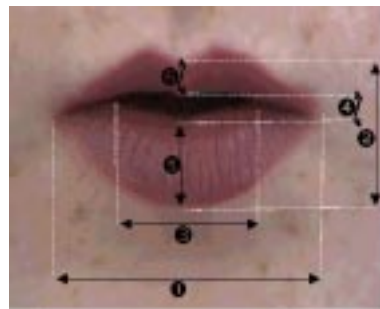


Fig. 4. Features Extracted From Video Sequences

Additional features which are not always visible, were also taken into account.

	Data Set 1	Data Set 2	Data Set 3	Data Set 4	Average
Consonants	61.82%	63.72%	63.28%	61.39%	62.54%
Vowels	78.5%	77.72%	81.22%	71.96%	77.3%
Total	67.46%	68.54%	69.35%	65.03%	67.58%

TABLE III

RESULTS FROM CLASSIFICATION OF PHONEMES INTO VISEMES FOR THE TEST SET

- Height of tongue below upper lip
- Height of tongue above lower lip
- Height of tongue between upper and lower teeth
- Height of upper teeth
- Height of lower teeth

The extraction of these features was based on classifying each pixel in an image as one of the lip, teeth or skin classes using a rule base of red green or blue colour value. See [7] for a detailed discussion of this method. After classification of individual pixel values the above features were extracted using heuristic searching algorithms as explained in [21]. Since the objective of this paper is to illustrate the speech recognition side of this application, the details of feature extraction will not be presented here.

After successful feature extraction the words in the vocabulary were segmented manually on phoneme boundaries and fuzzy clustering algorithms were employed to group similar phonemes into viseme classes [21]. The viseme grouping is shown in appendix A.

After the formation of consonants, vowels and diphthongs viseme groups, the performance of classification of phonemes into their corresponding viseme groups has to be explored. For this purpose the entire database of words was segmented manually on viseme boundaries. The dataset of 302 words was split differently into 4 datasets, each set having 151 distinct words [21]. Training was performed on 151 words and 151 unseen words were used in the test set. The results obtained from these 4 sets for the test set are summarised in table III [21].

VII. SEGMENTATION OF SPEECH DATA

For isolated word recognition, it is necessary to have an automatic scheme for determining the boundary of phonemes within a word. Three methods are described for automatic segmentation in this section.

A. Segmentation by Picking out Points of Maximum Change

The simplest method of segmenting visual speech can be derived by considering points of maximum change in the sequences of feature values comprising a word. The assumption behind this method is that a big change in feature values occurs in the transition from one phoneme to another.

To illustrate this method, figure 5 shows the difference values obtained for the words ‘poke’ and ‘feet’. The phoneme boundaries have been indicated by hand. The boundaries for the word ‘poke’ lie at areas where the changes in feature values have acquired high values. On the other hand, the word ‘feet’ presents a difficult case for segmentation when only the differences are considered. The transition from an /f/ phoneme to an /i/ is quite clear from the graph but the changing values between /i/ and /t/ cannot be detected by this algorithm.

B. Generating Rules for Segmentation

In this section a method is described to generate rules for segmenting visual speech data [21]. In section IV a model was described to represent the variable length sequences of data in time. It can be seen from figure 3 that the cartesian grid built on fuzzy granules represents the feature values at different points in time. This grid can be interpreted as representing the patterns of feature values at different stages in time, the number of stages

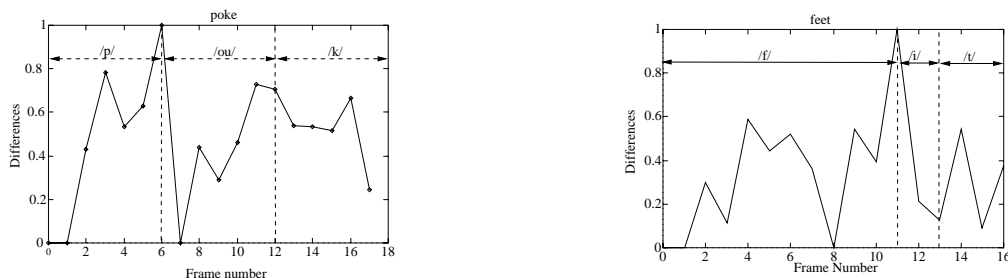


Fig. 5. Normalised differences between successive frames for ‘poke’ and ‘feet’

being equal to the number of time fuzzy sets used on the time axis. Assuming that the feature values follow certain patterns at different stages, these time fuzzy sets can be used for the generation of rules for determining where the boundary of a certain class begins and where it ends. Each time fuzzy set can be considered a class, so that each rule is of the form:

For a speech Sound v_i

<i>Rule 1</i>	<i>Rule 2</i>	<i>Rule M</i>
The time step is TF_1 given	The time step is TF_2 given	The time step is TF_M given
feature 1 is f_{11}	feature 1 is f_{21}	feature 1 is f_{M1}
feature 2 is f_{12}	feature 2 is f_{22}	feature 2 is f_{M2}
...
feature n is f_{1n}	feature n is f_{2n}	feature n is f_{Mn}

f_{ij} is the i^{th} feature fuzzy set for the j^{th} time step

The segmentation boundaries from the test set are determined as follows. When a test sequence ranging from 1 to t time frames is encountered, then for class C_j , a support for all the time fuzzy sets is obtained for each time frame. Suppose, at the i^{th} time frame the support for the M time fuzzy sets is:

$$s_{i1}, s_{i2}, \dots, s_{iM}$$

These supports can be interpreted as representing the probability that a sequence value belongs to a certain time fuzzy set, therefore, they can be normalised to add to one and are given by:

$$s'_{i1}, s'_{i2}, \dots, s'_{iM} \quad \sum_{j=1}^M s'_{ij} = 1$$

Therefore, the support for the i^{th} time frame being in the M^{th} or the last stage in time is s'_{iM} . This support is evaluated for all the sequence values between the maximum and minimum frames comprising a viseme. The sequence value with the highest support marks the boundary for class C_j .

C. Combining Rules with Points of Maximum Change

The method of generating fuzzy rules for segmenting speech data is quite intuitive because it involves learning phoneme boundaries from a training set. On the other hand if it is assumed that the transition from one phoneme to another involves a change in the trend of feature values then the method described in section VII-A should be taken into account. In this section it is proposed to combine the two methods of segmentation as described below.

1. When a test sequence is encountered, the feature values are normalised and the sum of differences between the successive frames is obtained. Take the frame number t_1 in this sequence (ranging from 1 to N) whose differences are greater than the threshold. This subsequence, say S_1 (varying from 1 to t_1), is then converted to its corresponding cross product fuzzy set representation and classified amongst all possible classes as class C_1 .
2. The next step is to consider the whole sequence S and find the best possible boundary for C_1 using the rule base as described in the previous section VII-B. This gives the sequence S_2 . Now convert this sequence to its corresponding fuzzy set representation and classify this sequence accordingly as class C_2 . If $C_1 = C_2$ then the boundary for the class has been found and this particular sub sequence is class C_1 .
3. If C_1 is not equal to C_2 then go back to step one and take the next sequence value at which the differences amongst the successive frames are greater than the threshold.

The algorithm described above not only takes into account points of maximum change but it is also coherent with the training database. It has been shown to give better results, as seen in section IX.

Before the results obtained from the aforementioned segmentation algorithms are discussed, a scheme for representing the words in the vocabulary is explained.

VIII. REPRESENTATION OF A VOCABULARY OF WORDS

After the classification and segmentation routines have been formulated, a scheme has to be devised for storing the vocabulary of words in the database. For this purpose all the words in the test vocabulary must be represented in an efficient way, which enables the system to perform a quick search for the most likely uttered word. One possibility is to

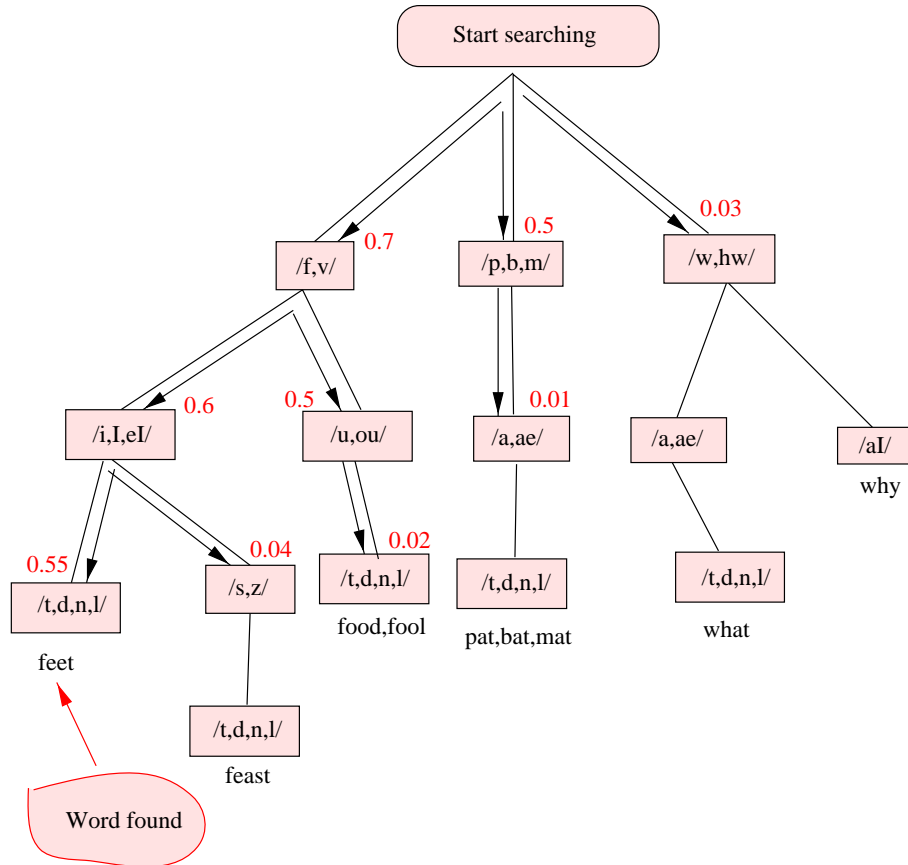


Fig. 6. Representation of the Vocabulary of Words

take a test sequence and compare the probability of its occurrence with all the utterances in the test set (e.g. [10]). This might be a good scheme for small vocabularies but for larger test sets it would be slow and inefficient. In this section the representation of words as a tree like structure is proposed, as shown in figure 6.

The tree consists of branches representing a viseme group. The terminal nodes or leaf nodes point to a possible set of words that are found by taking a particular path. The words that are common to one node have the same characteristic lip movement.

The tree like structure enables a quick search for the possible likely word. When a test sequence is encountered, the system takes the first branch and finds the first likely boundary in the sequence for the relevant viseme class by the segmentation algorithms described in section VII. Then using the corresponding boundary the system determines a support for that branch. The branches with low supports are abandoned and the ones

with high supports are followed until the terminal node is reached. At the terminal node all the supports are averaged to find an overall support for the word. Since the phoneme at the beginning of a word has a more prominent lip movement than at the end, therefore a weighted average of the supports at the various branches should be taken into account. For this work, a heuristic measure was used to assign the weights to each branch. The weights were assigned as $w_i = 100 - 10i$, where i represents the level or the depth of a branch. Hence the overall support for a word or a set of words at a terminal node is:

$$s_{word} = \frac{\sum_i w_i s_i}{\sum_i w_i}$$

An example of the search mechanism is shown in figure 6 where the system follows only the branches with high supports. Later in this paper we discuss how criteria can be established for following a branch or leaving it out.

The structure for the representation of the vocabulary of words not only enables a quick and efficient searching for a likely set of words but it has other advantages also. It automatically embeds the context information of the occurrence of a viseme with respect to its neighbours in a word. For example the occurrence of a /pbm/ viseme group is highly unlikely after an /fv/ group. Therefore when a /pbm/ group is found the system only goes into the relevant branches which reduces the search space by a considerable amount and takes the context information into account.

IX. RESULTS FOR ISOLATED WORD RECOGNITION

In this section the results obtained for isolated word recognition using the segmentation algorithms described in section VII and the classification of words using the tree structure illustrated in section VIII are presented. The database was split into a training and test set as explained in section VI. A summary of the results obtained on the training and test set is shown in table IV and V. The various rows of the table show the results for the Peak, Rules, Combined and Manual method.

A. Discussion of Results

The results summarised in tables IV and V illustrate the performance of various methods in visual speech recognition. When only a single word is selected out of an entire vocabu-

Method	Set 1	Set 2	Set 3	Set 4	Average
Peak	12.58%	13.25%	15.89%	11.92%	13.41%
Rules	29.14%	20.53%	23.18%	27.15%	25%
Combined	35.10%	23.84%	29.80%	31.79%	30.13%
Manual	43.05%	37.09%	40.40%	38.41%	39.74%

TABLE IV
RESULTS FOR THE TRAINING SET

Method	Set 1	Set 2	Set 3	Set 4	Average
Peak	10.60%	11.92%	9.93%	16.56%	12.25%
Rules	14.57%	18.54%	17.22%	17.88%	17.05%
Combined	24.50%	22.52%	19.21%	17.88%	21.03%
Manual	28.48%	35.10%	28.48%	33.77%	31.46%

TABLE V
VARIOUS RESULTS ON THE TRAINING AND TEST SET

lary of words then it can be seen that the ‘Combined’ method of segmentation gives the best results with an average of 21% accuracy. When the test sequence is segmented by only looking at the points of maximum change within data i.e. the ‘Peak’ method, the system performance is quite poor. The classification accuracy improves considerably by generating rules for segmentation. The generation of rules gives a deeper insight into the nature of data and hence they are more robust in deciding phoneme/viseme boundaries.

X. WORD SELECTION

The results described in the previous section depict a case where only one word was classified out of a whole vocabulary of test words. However, there are words in the vocabulary which have a very similar lip movement and hence there is a high support associated with them when classification is performed. These words should not be discarded and the system should come up with a selection of words having a high likelihood of occurrence.

To investigate which support should be used as a threshold, the following scheme was adopted. When going down the viseme tree only the branches b' with supports greater than the threshold support s_t were selected. Hence all words reached by going along branches b' are the selected words. Suppose the i^{th} test sequence generates a selection of words $w_{i1}, w_{i2}, \dots, w_{in}$ with supports $s_{i1}, s_{i2}, \dots, s_{in}$ respectively. The supports are normalised to add to one and are given by $s'_{i1}, s'_{i2}, \dots, s'_{in}$. These supports represent a probability distribution across the selection of words. If the test word presented to the system is amongst this selection of words then the normalised support s'_{is} associated with it is added to a count c . Hence after all the test words were presented to the system for a certain threshold support s_t the total count for the correctly classified words is:

$$c = \sum_j s'_{js}$$

Adopting the aforementioned method, the count for the correct words at the terminal node of the viseme tree depends not only on the support associated with a word but also on the number of words in the selected group. If the test word is contained in the selection of words and the group of words is large then the normalised support associated with it would be low, whereas if the group of words is small then the support associated with it would be high.

The experiment of classification was repeated with different threshold supports and total count associated with each of these threshold supports was recorded. By increasing the threshold, the count increases till a certain value, and then starts decreasing with a further increase in the value of the threshold. The accuracy obtained for the highest count is the accuracy for the selection of words and the corresponding value of threshold support should be taken when going down the tree. The phenomenon is illustrated in figure 7 for various data sets. The results obtained are shown in the figure 7. The dotted lines in these figures indicate the count obtained for different values of threshold for manual segmentation and the solid lines indicate the threshold support for automatic segmentation.

Table VI summarises the results obtained on the various datasets when manual segmentation and automatic segmentation were performed. In case of automatic segmentation using fuzzy rules, the threshold support for all the 4 sets lies around 0.057. The average count for these sets is 10.29 and the average accuracy obtained at this threshold value is

	Manual Segmentation		
Data Set	Threshold	Count	Accuracy
1	0.056	23.60	54.97%
2	0.057	21.50	53.64%
3	0.056	22.03	56.29%
4	0.058	22.12	49.0%
Average	0.057	23.06	53.48%
	Automatic Segmentation		
Data Set	Threshold	Count	Accuracy
1	0.055	11.34	35.10%
2	0.055	11.04	40.41%
3	0.06	9.91	25.83%
4	0.057	8.86	31.79%
Average	0.057	10.29	33.28%

TABLE VI

RESULTS OF DIFFERENT DATASETS FOR THE BEST THRESHOLD

33.28%.

XI. SUMMARY AND DISCUSSION

In this paper various methods involved in automatic recognition of isolated words in visual speech have been discussed. Novel methods for representing speech data and its classification and segmentation have been illustrated. When selecting only one word out of a selection of words, the accuracy of the system was 21%. The reader should be reminded again that a high accuracy in visual speech should not be expected because visual data carries only partial information regarding speech. Expert human lip-readers achieve an estimated accuracy of 30% on nonsense words [10] and if an average lip-reader is presented with a series of syllables with consonants followed by a vowel, his accuracy would be about 25% [12]. For his thesis, Goldschen achieved an accuracy of 25% on a test set of 150 sentences. However, the task of recognising isolated words is more difficult

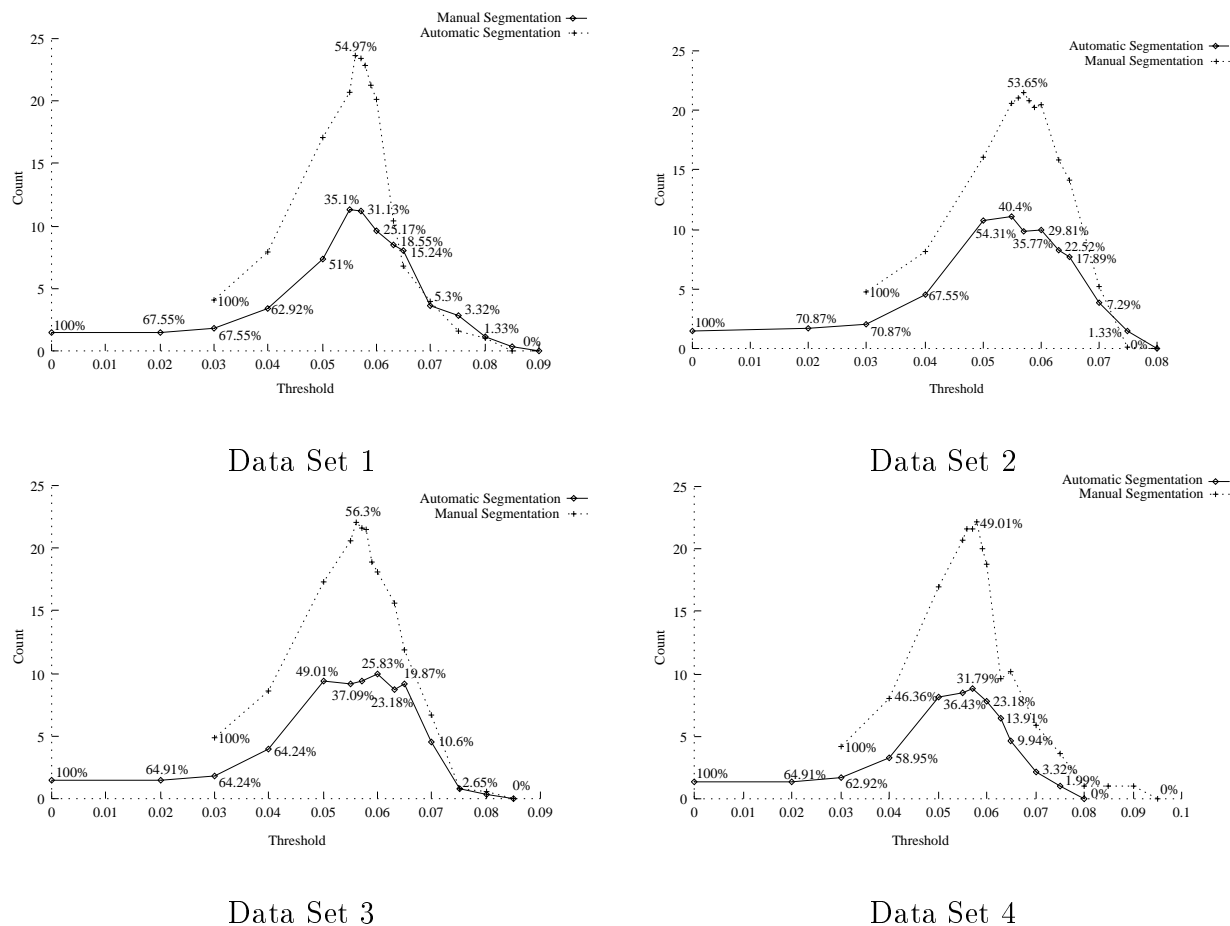


Fig. 7. Plot of Count Against Threshold for Data Set 1,2,3 and 4

since it involves discriminating between many visually similar words. The system might come up with a slightly higher support for a word that is visually similar to the actual test word, leading to a misclassification. Apart from Goldschen, Silsbee and Bovik [25] achieved an average accuracy below 20% for 2 speakers on a visual only subsystem for a 500 word vocabulary discrimination task. In comparison to these systems, the performance of extended Fril rules and fuzzy rules is 21% on a set of 151 test words, which shows the feasibility of the application of fuzzy set theory to visual speech recognition.

When selecting a group of words, the accuracy of the system was 33%. The accuracy achieved by manual segmentation is higher than the accuracy attained through automatic segmentation. The results from automatic segmentation are poorer because of the difficulty in distinguishing between the boundaries present between a consonant and a vowel at the end of a word. Generally this transition involves a very subtle change in feature values

which is hard to detect. There is a need for improvement of segmentation algorithms which can detect such changes. Interestingly, both automatic and manual segmentation, however, agree on the optimum value of threshold support i.e. 0.057. Hence the support value 0.057 can be taken as the threshold value when going down the tree to classify words.

APPENDIX

I. VISEME GROUPING EMPLOYED

Table VII shows the viseme grouping used in this work. The phonemes symbols conform to the International Phonetic Alphabet. The '*' symbol is suffixed for a phoneme when it is at the beginning of the word e.g. r* is the /r/ in run. The '&' symbol is suffixed for phonemes when they are in the middle or at the end of a word e.g. r& is the /r/ in blur.

Consonant Grouping		Vowel Grouping
f v	l* h	i I eI ε æ ɜ aI
w hw	l&	Λ a
p b m	t& d& s& z&	e
r*	θ	u ɔ U OU O ɒ a
r&	ð	ɜ
t*	f ɜ tʃ ɟ	
d* s* z* n* n& k g	j	
ŋg ɣk	q	

TABLE VII
VISEME GROUPING

REFERENCES

- [1] Patricia Ashby. *Speech Sounds*. T J Press (Padstow) Ltd, Padstow, Cornwall, UK, 1995.
- [2] James F. Baldwin. A theory of mass assignments for artificial intelligence. *Fuzzy Logic and Fuzzy Control*, editors Dimiter Driankov, Peter W.Eklund, Anca L.Ralescu, IJCAI'91 Workshops on fuzzy logic and fuzzy control, in *Lecture notes in artificial intelligence*, pages: 22-34, Springer-Verlag, Sydney, Australia, August 1991.
- [3] James F. Baldwin. Combining evidences for evidential reasoning. *International Journal of Intelligent Systems*, 6(6), pages: 569-616, 1991.

- [4] James F. Baldwin. Fuzzy and probabilistic uncertainties. *Encyclopedia of AI*, editor: S.A. Shapiro, pages: 528-537, John Wiley (2nd ed.), 1992.
- [5] James F. Baldwin, Trevor P. Martin, Bruce W. Pilsworth. *FRIL-Fuzzy and Evidential Reasoning in Artificial Intelligence*. Research Studies Press (Wiley Inc.), 1995.
- [6] James F. Baldwin. Logic programming with uncertainty and computing with words. *Logic Programming and Soft Computing*, editors: Trevor P.Martin and F. Arcelli Fontana, pages: 19-48, RSP/Wiley, 1998.
- [7] James F. Baldwin, Simon J. Case, Trevor P. Martin. Machine interpretation of facial expressions. *BT Technology Journal*, 16(3), pages: 156-164, 1998.
- [8] James F. Baldwin, Trevor P. Martin, Mehreen Saeed. Automatic Computer Lip-reading Using Fuzzy Set Theory. *Proceedings of Auditory Visual Speech Processing AVSP'99*, pages: 92-96, Santa Cruz, California, U.S.A, August 7-9, 1999.
- [9] Kathleen E. Finn and Alen A. Montgomery. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8(3), pages: 159-164, 1988.
- [10] Alan J. Goldschen. *Continuous Automatic speech recognition by lipreading*. PhD thesis, George Washington University, Washington, D. C., 1993.
- [11] Yifan Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, Vol. 16, pages: 261-291, 1995.
- [12] Mary E. Henegar and R.Orin Cornett. *Cued Speech Handbook for Parents*. Cued speech program, Gallaudet College, Kendall Green, Washington, D.C., 1971.
- [13] Janet Jeffers and Margaret Barley. *Speechreading (Lipreading)*. Charles C Thomas Publisher, Springfield, Illinois, U.S.A., 1971.
- [14] Juergen Luettin and Neil A. Thacker, Steve W.Beet. Speechreading using shape and intensity information. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, Vol. 1, pages: 58-61, USA, February 1996.
- [15] Juergen Luettin and Neil A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2), pages:163-178, February 1997.
- [16] Kenji Mase and Alex Pentland. Lip reading: automatic visual recognition of spoken words. *Proceedings Image Understanding and Machine Vision*, Optical Society of America, June 12-14, 1989.
- [17] Dominic W.Massaró. Speech perception by ear and eye. *Hearing By Eye: The Psychology of Lip-reading*, editors: Barbara Dodd and Ruth Campbell, pages: 3-51, Lawrence Erlbaum Associates Ltd., London, 1987.
- [18] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, Vol. 264, pages: 746-748, December 23/30, 1976.
- [19] Javier R. Movellan. Visual speech recognition with stochastic networks. *Advances in Neural Information Processing Systems*, editors: G. Tesauro, D. Toruetzky and T. Leen (eds.), Vol. 7, MIT Press, Cambridge, 1995.
- [20] Javier R. Movellan and Paul Mineiro. A diffusion network approach to visual speech recognition. *Proceedings of Auditory Visual Speech Processing AVSP'99*, pages: 92-96, Santa Cruz, California, U.S.A, August 7-9, 1999.
- [21] Mehreen Saeed. Soft AI Methods and Visual Speech Recognition. PhD Dissertation, Department of Engineering Mathematics, University of Bristol, UK, 1999.
- [22] Eric Petajan. Automatic lipreading to enhance speech recognition. PhD dissertation, University of Illinois at Urbana-Champaign, 1984.

- [23] Eric Petajan, Bradford Bischoff, David Bodoff and Michael N. Brooke. An improved automatic lipreading system to enhance speech recognition. *ACM SIGCHI-88*, pages: 19-25, 1988.
- [24] Peter L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, University of Texas, 1993.
- [25] Peter L. Silsbee and Alan C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5), pages: 337-351, September, 1996.
- [26] Quentin Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. *Hearing By Eye: The Psychology of Lip-reading*, editors: Barbara Dodd and Ruth Campbell, pages: 3-51, Lawrence Erlbaum Associates Ltd., London, 1987.
- [27] Quentin Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London*, Series B, 335, pages: 71-78, 1992.
- [28] Gregory J. Wolff, K.Venkatesh Prasad, David G. Stork, Marcus Hennecke. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. *Advances in Neural Information Processing Systems*, Vol. 6, Editors Jack D.Cowan, Gerald Tesauero, Joshua Alspector, pages: 1027-1034, Morgan Kaufmann publishers INC, San Francisco, 1994.
- [29] Ben P. Yuhas, Moise H. Goldstein and Terrence J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, pages: 65-71, November 1989.
- [30] Ben P. Yuhas, Moise H. Goldstein, Terrence J. Sejnowski and Robert E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10), pages: 1658-1668, October 1990.
- [31] Lofti A. Zadeh. Soft computing and fuzzy logic. *IEEE Software*, Vol. 11, pages: 48-56, November, 1994.
- [32] Lofti A. Zadeh. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2), pages: 103-111, May, 1996.