6

Automatic Visual Speech Recognition

Alin Chiţu[1] and Léon J.M. Rothkrantz[1,2]
[1]*Delft University of Technology*
[2]*Netherlands Defence Academy*
The Netherlands

1. Introduction

Lip reading was thought for many years to be specific to hearing impaired persons. Therefore, it was considered that lip reading is one possible solution to an abnormal situation. Even the name of the domain suggests that lip reading was considered to be a rather artificial way of communication because it associates lip reading with the written language which is a relatively new cultural phenomenon and is not an evolutionary inherent ability. Extensive lip reading research was primarily done in order to improve the teaching methodology for hearing impaired persons to increase their chances for integration in the society. Later on, the research done in human perception and more exactly in speech perception proved that lip reading is actively employed in different degrees by all humans irrespective to their hearing capacity. The most well know study in this respect was performed by Harry McGurk and John MacDonald in 1976. In their experiment the two researchers were trying to understand the perception of speech by children. Their finding, now called the McGurk effect, published in Nature (Mcgurk & Macdonald, 1976), was that if a person is presented a video sequence with a certain utterance (i.e. in their experiments utterance 'ga'), but in the same time the acoustics present a different utterance (i.e. in their experiments the sound 'ba'), in a large majority of cases the person will perceive a third utterance (i.e. in this case 'da'). Subsequent experiments showed that this is true as well for longer utterances and that is not a particularity of the visual and aural senses but also true for other perception functions. Therefore, lip reading is part of our multi-sensory speech perception process and could be better named visual speech recognition. Being an evolutionary acquired capacity, same as speech perception, some scientists consider the lip reading's neural mechanism the one that enables humans to achieve high literacy skills with relative easiness (van Atteveldt, 2006).

Another source of confusion is the “lip” word, because it implies that the lips are the only part of the speaker face that transmit information about what is being said. The teeth, the tongue and the cavity were shown to be of great importance for lip reading by humans (Williams et al., 1998). Also other face elements were shown to be important during face to face communication; however, their exact influence is not completely elucidated. During experiments in which a gaze tracker was used to track the speaker's areas of attention during communication it was found that the human lip readers focus on four major areas: the mouth, the eyes and the centre of the face depending on the task and the noise level (Buchan et al., 2007). In normal situations the listener scans the mouth and the other areas

relatively equal periods of times. However, when the background noise increases, the centre of the face becomes the central point of attention. Most probably the peripheral vision becomes extremely active in these situations. When the task was set to the inference of the emotional load of the interlocutor, the listener's gaze started to be shifted towards the eyes since they convey more emotional related information. It is well accepted that the human lip readers make great use of the context in which the interaction takes place. This can be one of the reasons the human listener scans the entire face during the interaction. In (Hilder et al., 2009) the authors found that when a human lip reader was presented with appearance information, compared with only mouth shapes, his performance increased considerably from 42.9% to 71.6%.

We should realise that during face to face interaction a human engages in a complex process which involves various channels of information corresponding to our senses. In this way the speaker builds up the context using both verbal and non-verbal cues such as body gesture, facial expressions, prosody, and other physiological manifestations. Other information about the settings in which the communication takes place is used as well as the knowledge accumulated in time through experience. A human is a multi-modality, multi-sensory, multi-media fusing machine.

The rest of the chapter is organized as follows: section 2 presents relevant research works in the area of lip reading. Section 3 presents the aspects related to building an automated lip reader. Section 4 details the characteristics of the facial model used during the visual analysis of the lip reading process. The next sections illustrate the results and discuss the conclusions of the algorithms presented in the chapter.

2. State of the art in lip reading

It is about three decades since automatic lip reading domain emerged in the scientific community. However, only starting from the 90s, and more sustained in the second half of the 90s, the subject started to become viable. Even today it still lags the speech recognition by some decades. Until some years ago the most impeding factor was the computational power of the computers. Nowadays it is the difficulty in finding the most suitable visual features that capture the information related with what is being spoken. Also it is the hard problem of accurately detecting and tracking the facial elements that convey speech related information. The automatic and robust detection and tracking of the face elements is still not entirely achieved by the current technology. As in other similar visual pattern recognition applications, the two monsters “illumination variations” and “occlusions” are still alive and menacing. A special case of occlusion is in this case generated by the posture of the speaker.

Therefore, any study concerning lip reading deals with the overwhelming task of manually or in the best case semi-automatically processing the data corpus. The data corpora for lip reading are still very small due to partially the storage and bandwidth limitations and other recording related settings, but much more limiting due to the overwhelming task of processing and preparing the data for experiments. Because of these issues, each data corpus is created for a stated recognition task. The lip reading experiments to this date are limited to isolated or connected random words, isolated or connected digits, isolated or connected letters. Some of the reported performance is listed below. However, it is very important to keep in mind that, because the data corpus used influences in great respect the performance

of the lip reader, a comparison among the experiments is not always possible. When the corpora are about the same, then the comparison of the different feature types and feature extraction techniques becomes feasible. It can still give an impression on the state of the art in lip reading.

The task of isolated letters was among the first analysed by Petajan et al. (Petajan et al., 1988) back in 1998. The authors report the correct recognition close to 90%. However, based on the AVletters data corpus, Matthews et al. (Matthews et al., 1996) reports only a 50% recognition rate. Li et al. (Li et al., 1995) reports a perfect recognition 100% on the same task, but two years later in (Li et al., 1997) only 90% recognition. The second most popular task is digit recognition either in isolation or as connected strings. Based on the TULIPS1 data corpus, which only contains the first four digits, Luettin et al. (Luettin et al., 1996) and Luettin and Thacker (Luettin & Thacker, 1997) reported 83.3% and 88.5% recognition rates, respectively. Arsic and Thiran (Arsic & Thiran, 2006) report on the same data corpus 81.25% and 89.6% depending on the feature extraction method. Other experiments with the digit recognition task are: Potamianos et al. (Potamianos et al., 1998) reported 95.7%, Dupont and Luettin (Dupont & Luettin, 2000) reported 59.7%, Wojdel reported in his thesis (Wojdel, 2003) 91.1% correct recognition and 81.1% accuracy, Patamianos et al. (Potamianos et al., 2004) reported 63% and Perez et al. (Prez et al., 2005) 47%. Lucey and Potamianos (Lucey & Potamianos, 2006) reported 74.6% recognition rate for the isolated digits task. Potamianos et al. (Potamianos1998a) report 64.5% recognition rate for the connected letter task. For the isolated word task Nefian et al. (Nefian et al., 2002) report 66.9%, Zhang et al. (Zhang et al., 2002) report 42%, Kumar et al. (Kumar et al., 2007) report 42.3%. We can conclude that there is still a large variation in the performances obtained, and there is still no convergence visible since the newer studies do not necessarily show an increase in accuracy. This is, to our opinion, clearly a sign of the immaturity of the lip reading domain. Also, as can be observed in the listing above, there are yet no results of experiments with continuous speech. Patamianos et al. (Potamianos et al., 2004) report an extremely low result on the continuous speech task, namely 12%. The lip reading domain is still young and there are many limiting factors that need to be conquered. Therefore, the experiments in lip reading are still dealing with relatively easy tasks. However, the promising results in these tasks give us hopes that larger experiments are possible. As the domain becomes more popular, the number of data corpora will increase and with a better cooperation among scientists it will be possible to better compare the achievements. However, there are objective factors which limit the performance of the lip readers. Nevertheless, as shown in many studies, lip reading can be successfully used in conjunction with speech for an enhanced speech recognition system.

3. Building an automatic lip reader: Overview

Building a lip reader is in many ways similar to building any automatic system which performs an autonomous role in its environment. The first decision needed to be made before starting the construction of the system is with respect to the role of the system and with respect to the environment where the system will be deployed. After establishing, in pattern recognition jargon, the recognition task, building the system consists of four separate stages: data acquisition, data parametrization, model training and model testing. Figure 1 describes the general process of building a lip reader. These activities are

performed in cycles, the larger the cycle the less frequent its corresponding process is performed.

The data acquisition process should ensure that the resulting corpus correctly describes the distribution of the possible states of the modelled process. The importance of the data parametrization is twofold; it should extract only the relevant information from the data and it should reduce the dimensionality of the feature space, therefore increasing the tractability of the problem. Training and testing are dependent on the mathematical models chosen for inference. These range from plain heuristics to complex probabilistic graphical models. The training process should solve two problems: identify the structure of the models such as the number of parameters and their relation, and compute the values of the models' parameters.

Training and testing is usually performed in a cycle which will fine tune the structure of the models and the values of the weights in the model. However, the data parametrization step is the one that is most of the time investigated, since there are many ways to extract suitable information for the process under study. Choosing the right parametrization is not straightforward and usually a trial and error sequence of experiments is started.

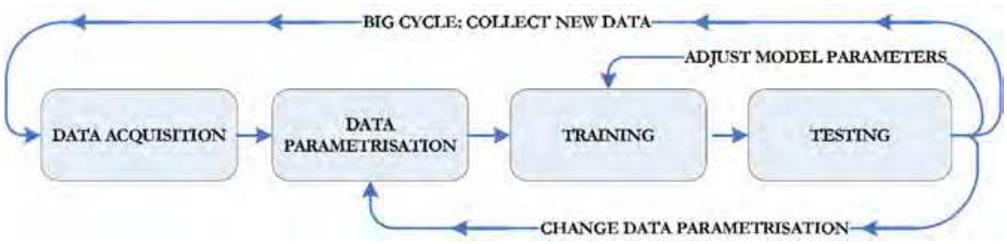


Fig. 1. The activity sequence for building a lip reader

A lip reader and in general a speech recogniser is built for a particular target language. The recognition task, namely the size of the vocabulary and the type of utterances accepted, are paramount for the entire design of the system. For instance if for a small vocabulary (i.e. a few tens of words) one model can be used to recognise one entire word, for larger vocabularies it is more suitable to build sub-word models, i.e., to directly recognise sub-words and build the words and sentences using dictionaries and grammar networks.

So far, the most successful approach for speech recognition, and therefore also applied to lip reading, is the Bayesian approach. In the Bayesian approach, the recognition problem can be formulated as follows: given a set of possible words and an observation sequence $O=(O_1,O_2,\ldots,O_n)$ the solution of the recognition problem is the word that maximizes the probability $P(W|O)$. Based on the Bayesian rule we can write: $P(W|O)=\frac{P(O|W)P(W)}{P(O)}$, where $P(O|W)$ is the likelihood of the observation given the word W and P (W), usually called the language model, represents the probability of the word W. The problem can be thus rewritten as: $\hat{W}=argmax_W\left(P(O|W)P(W)\right)$, where W is the recognized word. In the above equation the denominator $P(O)$ has been deleted since it does not influence the solution. Therefore, the recognition problem is reduced to building a language model P(W) and a word model $P(O|W)$ for each legal word.

3.1 On building a data corpus for lip reading: A comparison

In order to evaluate the results of different solutions to a certain problem, the data corpora used should be shared between researchers or otherwise there should exist a set of guidelines for building a corpus that all datasets should comply with. In the case when a data corpus is build with the intention to be made public, a greater level of reusability is required. In all cases, the first and probably the most important step in building a data corpus is to carefully state the targeted applications of the system that will be trained using the dataset. Some of the most cited data corpora for lip reading are: TULIPS1 (Movellan, 1995), AVletters (Matthews et al., 1996), AVOZES (Goecke & Millar, 2004), CUAVE (Patterson et al., 2002), DAVID (Chibelushi et al., 1996), ViaVoice (Neti et al., 2000), DUTAVSC (Wojdel et al., 2002), AVICAR (Lee et al., 2004), AT&T (Potamianos et al., 1997), CMU (Zhang et al., 2002), XM2VTSDB (Messer et al., 1999), M2VTS (Pigeon & Vandendorpe, 1997) and LIUM-AVS (Daubias & Deleglise, 2003). With the exception of M2VTS which is in French, XM2VTSDB which is in four languages and DUTAVSC which is in Dutch the rest are only in English (Table 1). Since the target language for our research was Dutch, we had only one option, namely the DUTAVSC (Delft University of Technology Audio-Visual Speech Corpus). For reasons that will be explained in the next paragraphs, we decided to build our own data corpus. This corpus was build as an extension to the DUTAVSC and is called NDUTAVSC (Chitu & Rothkrantz, 2009) which stands for "New Delft University of Technology Audio-Visual Speech Corpus".

Some aspects related to the data set preparation are as follows:

- The complexity of audio data recording is much smaller than of the video recordings. Therefore, all datasets store the audio signal with sufficiently high accuracy, namely using a sample rate of 22 kHz to 48 kHz and a sample size of 16 bits. Therefore, the quality of the audio data is not subject to storage accuracy but from the perspective of recording environment. There are two approaches to the recordings environment: specific and neutral. In the first case the database is built with a very narrow application domain in mind such as speech recognition in the car. In this case the recording environment matches the conditions of the environment where the system will be deployed. This approach can guarantee that the particularities of the target environment are closely matched. The downside of this approach is that the resulting corpus is too much dedicated to the problem domain and suffers from over training, and offers little generalization. In the second approach the dataset can be recorded in controlled, noise free environment. The advantage of this approach is the possibility to adapt the corpus to a specific environment in a post process. Therefore, a data corpus of this kind can be used for virtually any number of applications. The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data.
- In the case of video data recording there is a larger number of important factors that control the success of the resulting data corpus. Hence, not only the environment, but also the equipment used for recording and other settings is actively influencing the final result. In the case of the environment the classification made for audio holds for video as well. The environment where the recordings are made is important since it can determine the illumination of the scene, and the background of the speakers. In the case of a controlled environment the speakers background is usually monochrome so that by

using a "colour keying" technique the speaker can be placed in different locations inducing in this way some degree of visual noise. However, the illumination conditions of different environment are not as easily applied to the clean recordings, since the 3D information is not available anymore. In controlled environments the light is reflected by special panels which cast the light uniformly, reducing the artefacts on the speaker's faces.

- The equipment used for recording plays a major role, because the resolution and the sample rate is still a heavy burden. Hence, the resolution of the recordings ranges from 100x75 pixels in Tulips1 and 80x60 pixels in AVletters datasets to 720x576 pixels in AVOZES and CUAVE datasets. The same improvement in quality is also observed in colour fidelity.
- The frame rate of the existing data corpora is conforming to one of the colour encoding systems used in broadcast television systems. Therefore, the video is recorded at 24Hz, 25Hz, 29.97Hz of 30Hz depending on the place in the world where the recordings are made. The data corpus used for the current research was recorded at 100Hz.
- The Region Of Interest (ROI) is important as well. For lip reading only the lower half of the face is important. However, in case context information is required, a larger area might be needed. Most of the datasets show, however, a passport like image of the speaker. In our opinion, at least for increasing the performance of the parametrization process a smaller ROI is more advantageous. Of course a ROI that is too narrow adds high constraints on the performances of the video camera used and it might be argued that this is not the case in real life where the resulting system will be used. Recording only the mouth area as is done in the Tulips1 data set is a tough goal to achieve in an uncontrolled environment. However, by using a face detection algorithm combined with a face tracking algorithm we could automatically focus and zoom in on the face of the speaker. A small ROI facilitates acquiring a much greater detail of the area of interest, in our case the mouth area, while keeping the resolution and, therefore, the bandwidth needs in manageable limits.

Figure 2 shows some examples from six available data corpora. The differences among the examples in this figure are clearly visible, with the exception of the DUTAVSC corpus, all other corpora reserve a small number of pixels for the mouth area. Table 2 gives the sizes of the mouth bounding box in all six samples. This low level of detail makes the detection and tracking of the lips much more difficult. Any parametrization that considers a description of the shape of the mouth will be heavily influenced by image degradation. In the paper

Fig. 2. The resolution of the ROI in some data corpora available for lip reading

(Potamianos et al., 1998) the authors report that the degradation of the video signal by the image compression algorithm by the addition of white noise does not influence the lip reading performance unless the Signal to Noise Ratio(SNR) falls under some threshold: 50% and 15%, respectively. These findings are reported when the features used are a linear transformation of the intensities in the images, namely discrete wavelet transform.

Corpus	Language	Sessions	Respondents	Audio Quality	Video Quality	Language Quality	Stated purpose
TULIPS1	English	1	7male, 5female	11.1kHz, 8bits controlled audio	100x75, 8bit, 30fps mouth region	first 4 digits in English	small vocabulary isolated words recognition
AVletters	English	1	5male, 5female	22kHz, 16bits controlled audio	80x60, 8buts, 25fps mouth region	the English alphabet	spelling English alphabet
AVOZES	English	1	10male, 10female	48kHz, 16bits controlled audio	720x480, 24bits, 29.97fps entire face, stereo view	digits from '0' to '9' continuous speech application driven utterances	continuous speech recognition for Australian English
CUAVE	English	1	19male, 17female	44kHz, 16bits controlled audio	720x480, 24bits 29.970fps passport view	7,000 utterances connected and isolated digits	continuous speech recognition
Vid-TIMIT	English	3	24male, 19female	32kHz, 16bits controlled audio	512x384, 24bits, 25fps upper body	TIMIT corpus 10 sentences per person	automatic lipreading, face recognition
DAVID	English	12	132male, 126female (in 4 groups)	--	entire face, upper body, profile view multi corpora: controlled and degraded background, highlighted lips	vowel – consonants alternation, English digits	speech or person recognition
IBM LVCSR*	English	1	290 Unknown gender	22kHz, 16bits --	--	connected digits isolated words	audio-visual speech recognition
AVICAR	English	5	50male, 50female	48kHz, 16bits, 8channels 5 levels of noise car specific	4 cameras from different angles, passport view car environment	isolated digits, isolated letters, connected digits, TIMIT sentences	speech recognition in a car environment
DUTAVSC	Dutch	10-14	7male, 1female	48kHz, 16bits, controlled audio	384x288, 24bits, 25fps lower face view	spelling, connected digits, application driven utterances, POLYPHONE corpus**	audio-visual speech recognition, lipreading

* Not available to the public
** Data corpus for Dutch. Recordings are made over phone lines. More details can be found in (Damhuis et al. 1994)

Table 1. Characteristics of data corpora.

Data corpus	Width	Height
AVOZES	122	24
CUAVE	75	34
M2VTS	46	28
TULIPS1	76	37
VidTIMIT	53	25
DUTAVSC	225	133

Table 2. Resolution of the mouth area in six known corpora for lip reading.

3.1.1 Language quality

By its nature lip reading requires, irrespective of the other qualities, that the data corpus has a good coverage of the language and task vocabulary. Therefore, in the case of a word based recognizer all the words in the vocabulary need to be present in the corpus. In the case of a sub-word recognizer every sub-word item needs to be present in the corpus in all existing contexts. Therefore, the co-articulatory effects appear with a reasonable frequency. However, due to the amount of work necessary and the storage and bandwidth required

most of the data corpora only consider small recognition tasks and small language corpus. Most frequently the data corpora focus on the digits and letters of the language considered. These are recorded either isolated, or in short sequences, or as in DUTAVSC in spelling of words. Some corpora even only consider nonsense combinations of vowels(V) and consonants(C) (e.g. DAVID considers VCVCVC sequences, AVOZES repetitions of "ba" and "eo" constructions, AT&T CVC sequences). The continuous speech case is only considered in AVOZES which contains only 3 phonetically balanced sentences, in AVICAR which contains ten sentences from the TIMIT (Garofolo, 1988) speech data corpus, XM2VTSDB and M2VTS which contains one random sentence and DUTAVSC which contains 80 phonetically rich sentences. The DUTAVSC is by far the most rich data corpus. The NDUTAVSC corpus which was built as an extension of DUTAVSC contains more than 2000 unique rich sentences. However, none of the existing corpora can match the language coverage offered by the data corpora used for speech recognition which can easily have a vocabulary of 100k words (e.g. the Polyphone corpus (Boogaart et al., 1994) contains more than one million words recorded and a vocabulary of 150k words).

3.2 Feature vectors definition

There are many approaches to data parametrization, but with respect to the feature vectors definition they all fit in three broad classes: texture based features, geometric based features, and combination of texture and geometric features. A good overview of most of the feature extraction methods can be found in (Potamianos et al., 2004). In the first class the feature vectors are composed of pixels' intensities values or a transformation of them in some smaller feature space. The main function of the projection is to reduce the dimensionality of the feature space while preserving as much as possible the most relevant speech related information. Principal Component Analysis (PCA) is one of the first choices, and therefore very popular, and was used in many studies e.g. (Bregler et al., 1993); (Bregler & Konig, 1994); (Duchnowski et al., 1994); (Li et al., 1995); (Tomlinson et al., 1996); (Chiou & Hwang, 1997); (Gray et al., 1997); (Li et al., 1997); (Luettin & Thacker, 1997); (Potamianos et al., 1998); (Dupont & Luettin, 2000); (Hong et al., 2006). The feature definition is based on the notion of eigenfaces or eigenlips which represent the eigenvectors of the training sets. An alternative to PCA, very common as well, is Discrete Cosine Transform (DCT) such as in (Duchnowski et al., 1995); (Prez et al., 2005); (Hong et al., 2006); (Lucey & Potamianos, 2006). Linear Discriminant Analysis (LDA), Maximum Likelihood Data Rotation (MLLT), Discrete Wavelet Transform, Discrete Walsh Transform (Potamianos et al., 1998) are other methods that fit in this class and were used for lip reading. Virtually, any other method, usually borrowed from the data compression domain, which results in a lower dimensionality of the feature vectors can be applied for data parametrization in the lip reading domain. Local Binary Patterns (LBP) is just another technique, borrowed from the texture segmentation domain, and shows promising results for lip reading as well (Morn & Pinto-Elas, 2007); (Zhao et al., 2007); (Kricke et al., 2008). LBP was developed by Timo Ojala and Matti Pietikainen and presented in (Ojala & Pietikainen, 1997). A special place in this class is taken by the feature vectors that are based on Optical Flow Analysis (OFA) (Mase & Pentland, 1991); (Martin, 1995); (Gray et al., 1997); (Fleet et al., 2000); (Iwano et al., 2001); (Tamura et al., 2002); (Furui, 2003); (Yoshinaga et al., 2003); (Yoshinaga et al., 2004); (Tamura et al., 2004); (Chitu et al., 2007); (Chitu & Rothkrantz, 2009). The optical flow is defined as "the apparent velocity field in an image". This definition closely matches the affirmation of

Bregler and Konig in their 1994 paper (Bregler & Konig, 1994): "The real information in lipreading lies in the temporal change of lip positions, rather than the absolute lip shape". The OFA can be used as well as a measure of the overall movement and be employed for onset/offset detection. The main advantage of this approach is that it can be easily automated, since it requires only the definition of the Region Of Interest (ROI). The ROI can be considered the bounding box of the face or the bounding box of the mouth, thus requiring some object detection and tracking algorithm. A good example is the face detection algorithm developed by Viola and Jones in (Viola & Jones, 2001). The main disadvantage of this type of features is that the a-priory information about lip reading is not inherently used in the process of feature extraction. Therefore, there is minimum control over the information contained in the resulting feature vectors, on whether this information is relevant for lip reading or not. The exceptions can be the OPA and LBP where the analysis is usually performed in carefully chosen regions around the mouth. We defined the set of features based on OFA and analyzed the performance of the lip reading system trained on our data corpus. The features from the second class share the belief that in order to accurately capture the most relevant features, with respect to lip reading, a careful description of the contour of the speaker's mouth is needed. The feature extraction proceeds in two steps; first a number of key points are detected and based on these points the mouth contour is recovered, and second the feature vectors are defined based on the shape of the mouth. The detection of the key points is performed based on colour segmentation techniques that identify pixels that are on the lips. Thereafter, the contour of the lips is usually extracted by imposing a lip model to the detected points. These methods are using the so called "smart snakes" (Lievin et al., 1999); (Luettin & Thacker, 1997); (Salazar et al., 2007), or as called in (Eveno et al., 2004) "jumping snakes", or later on Active Shape Models (ASM) (Luettin et al., 1996); (Prez et al., 2005); (Morn & Pinto-Elas, 2007) or Active Contour Models (ACM). Any other parametric model can be used here. The lips' contour is usually detected as a result of an iterative process which searches to minimise the error between the real contour and the approximation of the contour the parametric model allows for. The actual feature vectors are defined in the second step. The feature vectors fall into two categories here: model based features and mouth high level features. In the first category the feature vectors contain directly the parameters of the models used for describing the mouth contour. In the second category the feature vectors contain measurable quantities, which are meaningful to humans. The most used high level features are mouth height, mouth width, contour perimeter, aperture height, aperture width, aperture area, mouth area, aperture angle and other relations among these (e.g. the ratio between the width and the height) (Chitu & Rothkrantz, 2009); (Goecke et al., 2000a, 2000b); (Kumar et al., 2007); (Matthews et al., 2002); (Yoshinaga et al., 2004).

In our research we used Statistical Lip Geometry Estimation (LGE) which is a feature extraction method introduced by Wojdel and Rothkrantz (Wojdel & Rothkrantz, 2000). This method is special because it is a model free approach for describing the shape of the lips. It strongly depends, however, on the performance of the image segmentation technique used to detect the pixels which belong to the lips. The third class consists of feature vectors that contain both geometric and texture features. The features from each category are usually concatenated in a larger feature vector. For instance (Dupont & Luettin, 2000) and (Luettin et al., 1996) combine ASM with PCA features and (Chiou & Hwang, 1997) combines snake features with PCA. It was shown that the tongue, teeth and cavity have great influence on

lip reading (Williams et al., 1998), therefore, the addition of these appearance related elements has significant influence on the performance of lip reading (Chitu et al., 2007). A special example is the so called Active Appearance Models (AAM) (Cootes et al., 1998) which combines the ASM method with texture based information to accurately detect the shape of the mouth or the face. The searching algorithm is iteratively adjusting the shape such that to minimise the error between the generated shape and the real shape. The core of AAM is PCA which is applied three times, on the shape space, on the texture space and on the combined space of shape and texture. The AAM based features can either consist of AAM model parameters in which case we have a combined geometric and texture feature vector, or of high level features computed based on the shape generated in which case we have a geometric feature vector. The lip reading results based on high level feature vectors which are computed starting from the lips' shape generated based on AAM are given in this chapter.

3.3 Lip reading primitives

This section introduces the visemes which are the lip reading counterparts of the phonemes.

3.3.1 Phonemes

In any spoken language a phoneme is the smallest segmental unit of sound which generates a meaningful contrast between utterances. Thus a phoneme is a group of slightly different sounds which are all perceived to have the same function by speakers of the language or dialect in question. An example of a phoneme is the group of /p/ sounds in the words pit spin and tip. Even though these /p/ sounds are formed differently and are slightly different sounds they belong to the same phoneme in English because for an English speaker interchanging the sounds will not change the meaning of the word, however strange the word will sound. The phones, or sounds, that make up a phoneme are called allophones. A speech recognizer can be built at word level or at sub-word level. While for a small vocabulary recognition task a word level system might be preferred, for large vocabulary, continuous speech task systems the phonemes are used as building blocks. Therefore, each phoneme in the target language corresponds to a recognition model in the speech recognizer.

In the Dutch language, approximately 40 distinguishable phonemes are defined. However, there can be slight differences among different phoneme and phoneme sets as a consequence of the target dialect and definition of accepted words. In the present research we used the phoneme set defined in (Damhuis et al., 1994). One problem is generated, for instance, by the neologisms. These words are divided in two classes: the ones that are already established into the language (e.g. the words of French origin) and have a stable pronunciation but which contain phonemes that are still under-represented in the language and a second class of very new words (e.g. the International English words from various technical and economical background) which bring a set of new phonemes that have no correspondence in Dutch. Table 3 shows the phonemes of the Dutch language as used in the Polyphone corpus. The phonemes are given in International Phonetic Alphabet (IPA), Speech Assessment Methods Phonetic Alphabet (SAMPA), and HTK notations, respectively.

3.3.2 From phonemes to visemes

Even though the definition of the concept of phoneme crosses the boundary of the auditory realm, and therefore is not bound to any sensory modality, the term "viseme" is used as the counter part of phoneme in the visual modality. The term was introduced by Fisher in (Fisher, 1968).

The visemes have a similar definition with the phonemes, namely, a viseme is a set of indistinguishable phonemes; indistinguishable phonemes from the point of view of the visual information available and not as in the phonemes case from the point of view of their meaning. There are two direct consequences of this definition. Firstly, there is no exact method of deciding the number and composition of the viseme classes; this is actually done either by a theoretical discussion of auditory-visual lip reading of phonemes or by modelling the human ability of recognizing the phonemes in the absence of the auditory stimulus, therefore, by modelling the degree of confusion of phonemes in the visual modality. Secondly, since there is no one-to-one mapping between the phonetic transcription of an utterance and the corresponding visual transcription, the separability of utterances in the visual modality decreases, which decreases the theoretical performance of a lip reader. The dependence of the visemes on the phonemes can be thought of as one reason why a new term was needed.

Unlike for English, to date there is only a limited number of publications which deal with the definition of visemes in Dutch; this is an almost complete list of them: (Breeuwer, 1985), (Corthals , 1984), (Eggermont, 1964), (van Son et al., 1994), (Visser et al., 1999) and (Beun, 1996). The papers (van Son et al., 1994) and (Beun, 1996) cited in (Wojdel, 2003), are the only examples, at least to the author's knowledge, where the classification of the viseme sets is done by elicitation of the human confusion matrices of phonemes. The authors of (van Son et al., 1994) found in their experiments that the Dutch lip readers are only able to recognize four consonantal and four vocalic visemes.

3.3.3 Modelling the visemes using HMM

As a sub-word based speech recogniser, the building blocks of our lip reader are the visemes of the Dutch language. Therefore, one HMM corresponds to one viseme. To the set of visemes are added two special models, namely sp for "short pause" and sil for "silence". The sp model is used for recognition of the short pause between words, while sil is used for the silence moments before and after the utterance. Depending on the recognition task, some visemes do not appear at all in the expected utterances and are, therefore, excluded from the study. This is the case for the digit and letter tasks.

The set of visemes which appear in the digit recognition task are listed in Table 4 and the set of visemes which appear in the letter recognition task are listed in Table 5. The visemes "at" and "a" are only present in the digit set, while the visemes "aa" and "pbm" are only present in the letter set.

The topology of the models used for modelling the visemes, usually used for phoneme-based speech recognition as well, is a 3-state left-right with no skips as shown in Figure 3. For implementation reasons, HTK requires that the models start and end with a non emitting node that facilitate the generation of recognition networks. A recognition network

consists of a string of linked models which are used during recognition by matching to the input utterance.

	Symbol			Example Word		
	IPA	SAMPA	HTK	Orthography	Transcription	Translation
1	p	p	p	pak	p a k	package
2	b	b	b	bak	b a k	container
3	t	t	t	tak	t a k	branch
4	d	d	d	dak	d a k	roof
5	k	k	k	kat	k a t	cat
6	g	g	gg	goal	gg oo l	goal(sports)
7	f	f	f	fel	f e l	fierce
8	v	v	v	vel	v e l	sheet
9	s	s	s	sein	s ei n	signal
10	z	z	z	zijn	z ei n	to be
11	x	x	x	acht	a x t	eight
12	ɣ	G	g	negen	n ee g at	nine
13	ɦ	h	h	hand	h a n t	hand
14	ʒ	Z	zj	bagage	b a g aa zj at	luggage
15	ʃ	S	sh	sjaal	sh aa l	scarf
16	m	m	m	met	m e t	with
17	n	n	n	nek	n e k	neck
18	ŋ	N	nn	bang	b a nn	scared
19	l	l	l	land	l a n t	country
20	r	R	r	rand	r a n t	edge
21	ʋ	w	w	wit	w i t	white
22	j	j	j	ja	j a	yes

Table 3. Polyphone's Dutch phoneme set: consonants.

	Viseme		Viseme
1	gkx	8	ei
2	oyu	9	sz
3	l	10	eeh
4	iee	11	at
5	td	12	a
6	fvw	13	sil
7	ie	14	sp

Table 4. The viseme set in HTK working notation for the digit recognition task.

	Viseme		Viseme
1	aa	9	h
2	pbm	10	ie
3	iee	11	ei
4	sz	12	l
5	td	13	oyu
6	eeh	14	sil
7	fvw	15	sp
8	gkx		

Table 5. The viseme set in HTK working notation for the letter recognition task.

In Figure 3 the numbers on the arcs represent the initial transition probabilities, set before training. Under the emitting states there is a generic drawing of the distribution of the feature vectors which is approximated by a mixture of Gaussian distributions. The modelling of the two silence models are introduced in the next section.

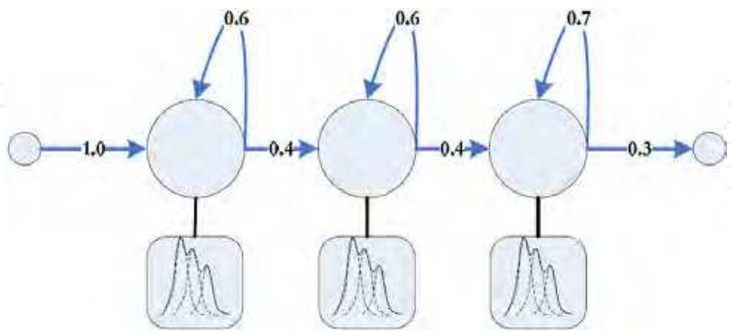


Fig. 3. The models used for modelling the visemes. The topology is 5-State Left-Right with three emitting states. The arcs are annotated with transition probabilities

3.3.4 Silence and pause models

It is not possible to build a continuous speech recognizer without including a model for silence. However, there are two types of silence, the ones between the words and the ones that appear in the beginning of the utterance and at the end of the utterance. The silence model that covers the entering and exit time of the utterances can be modelled using the same topology as for viseme models (i.e. 3-state left-right topology). However, in order to make the model more robust by allowing the states to absorb more non verbal mouth movement, the silence model is modified so that a backwards transition from state 4 to state 2 is accepted. The model for short pause is build starting from the model for [sil]. The short pause model is a so called tee-model and has a single emitting state which is tied to the central state of the [sp] model. This means that the central state of the [sil] model and the emitting state of the [sp] model share the same Gaussian mixture and therefore are trained using the same data. Parameter tying is very often used in speech recognition for the cases when there is not sufficient data for training models for similar entities. The topology used for the two silence models is shown in Figure 4. The silence models defined above are the same as the ones used for speech recognition. However, there is a big difference between the concept of silence in speech recognition and the concept of silence in lip reading. Consequently, the noise can have a more robust definition. For instance, in the case of visual speech the speaker can move his mouth for non verbal reasons (e.g. to moisture his lips, or to exteriorise the emotional status by showing a facial expression). The noise sources are more diverse for lip reading. Even though the silence model has an extra backward arc which should, in principle, also accommodate for noise in the training data, we found out in our experiments that the silence model defined in this way did not perform at the same level as in the case of speech. As we will see later in the results sections, sometimes the insertion rate was unexpectedly large. This can also be due to poorly trained silence models.

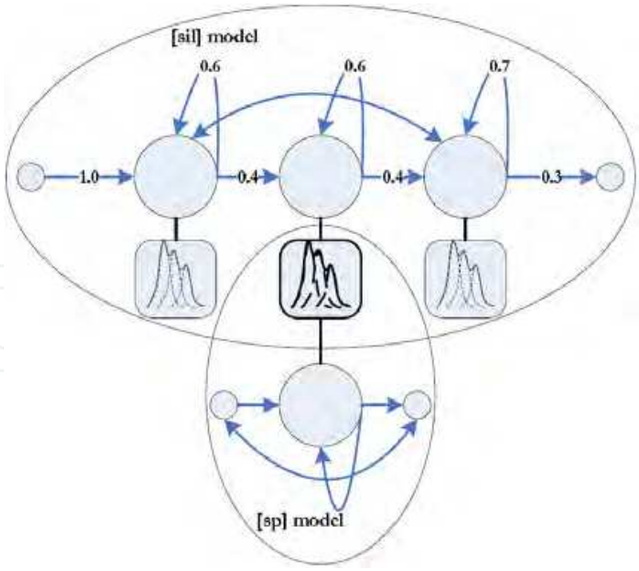


Fig. 4. The models used for modelling the silence

3.3.5 Modelling the low level context using Tri-visemes

In order to model the context at the level of the visemes, each viseme is considered in all the possible contexts. Only a one step context is considered, namely for each viseme only the left and the right possible visemes are considered, therefore, the name of the new entity is tri-viseme. The notation for tri-visemes is lf-vis+rt, where "vis" is the viseme in question, "lf" is the left context and "rt" is right context. For instance the word nul with the viseme transcription gkx oyu l will generate the following tri-visemes: gkx+oyu, gkx-oyu+l and oyu-l. The context of each viseme can be build at word level, also called word internal, or at the level of utterance called word external. In the first case, for finding all possible contexts of a viseme, only the words in the vocabulary are considered, while in the second case also the possible combinations of words can build the context. It should be noted that sometimes bi-visemes (i.e. viseme context containing only the left or the right viseme) are also generated. For each tri-viseme, a new model will be build which makes the number of models explode, making the data requirements for training a tri-viseme based recognizer many times larger. The major problem with the tri-visemes is that some contexts can appear only once (or a very small number of times) in the training data, or can even be absent from the training data, as in the case of trans-word boundary contexts. To solve this problem the parameter tying technique is used. The clustering of possible similar contexts can be made either by a data-driven approach, or by the use of decision trees. Even after the parameter tying, there can still be tri-viseme models which are undertrained.

3.4 Gaussian mixtures

The HMM approach considers that each of the emitting states in the model will be described by a continuous density distribution. This distribution is approximated in HTK by a mixture of Gaussian distributions. Building of the models in HTK starts by using only one Gaussian distribution. In the refining step the number of Gaussian mixtures is increased iteratively by 1 or 2 units until the optimum number of components is obtained. By monitoring the

performance change, the optimum number of mixtures can be found. During our experiments we iteratively increased the number of mixtures by one until a maximum of 32 mixtures. The "magic" number 32 was found sufficiently big to cover the optimum number of mixtures in all the experiments.

4. Facial model for lip reading

The AAM algorithm iteratively searches for the best fit of a model defined by a set of landmarks and the image being processed. Based on a-priori knowledge about the shape of the object, the set of landmarks is defined such that it optimally describes the object. In our case we required that the points selected describe the shape of the mouth in detail, especially capturing the speech related aspects. Therefore, the final model should exactly segment the lips in all moments during speech. After experimenting with different models and analysing the results, followed by long discussions, we decided to use a model composed of 29 points, distributed around the mouth, chin and nose. This model is shown in Figure 5.

For training a model, a number of two to four hundred images was manually processed. In order to obtain reliable results the images were selected such that they cover all the variance in the data. This was achieved in an iterative process. We first started with a random selection of a few tens of images which were used to build a first model. This model was used for processing until the performance of the model decreased below some visually assessed threshold. The images that were badly processed were added in the training set and a new model was obtained. This process continued until the performance of the model stabilized. In the end we trained a number of models for each speaker in the dataset. For speakers that recorded multiple sessions we trained one model per session.

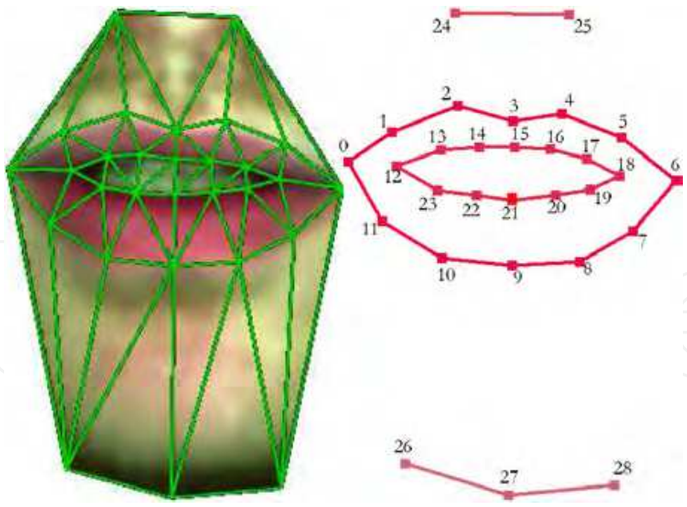


Fig. 5. The AAM model

Even though the process is fairly automated, this was an extremely laborious work, since the corpus contains more than 4.3 million frames, and was split among various people. Each assistant was asked to train a model and supervise the processing of the rest of the frames. Splitting the data among different people makes it more difficult to guaranty the uniformity over the entire corpus of the end result. Therefore, to assure uniformity of the processing we

used a strict definition of the landmarks. We defined as well constraints that acted on pairs of landmarks. The rest of this section gives the definition used for the landmarks. Before going to the next paragraphs, we should introduce some anatomical elements on which the definition of the landmarks depends.

4.1 AAM results on the training data

The AAM process is very fast and very accurate given that a good training set was selected. We combined the AAM searching scheme with the Viola&Jones mouth detection algorithm, which made the selection of a very good location for the initial guess possible. This has speeded up the search process to real time performance. The mouth detection was used only in the first few frames of the recording. In the subsequent frames the initial guess used was the result of the processing in the previous frame. This approach was very successful both in speeding up the search scheme and improving the accuracy of the detection. Figure 6 shows the first six most important components in PCA terminology. The mean shape and texture model is shown on the centre row. The top row shows for each mode the resulting object after an adjustment by two standard deviations is applied to on the corresponding mode. The bottom row shows the result when the adjustment is negative. The first two modes seem to have more control over the vertical and horizontal movement of the mouth, while mode four seems to control the presence of the tongue. However, there is no strict separation between the information controlled by each mode, at least not easily discernable by visual inspection. This model was trained on a set of 440 images, selected in an iterative process. All three models (i.e. appearance, shape and combined models) were truncated at 95% level. Based on the 95% level truncation, the final combined model had 38 parameters, while the shape model had 11 parameters and the texture model had 120 parameters. The first six modes in the combined model cover 78.65%. However, in the case of the shape models the first two modes already cover 82.53% of the total variation, while the first six cover 91.83% of the variation.

Fig. 6. Combined shape and appearance statistical model. The images show from left to right the first six most important components in PCA terminology. These modes account for 78.65% of the total variation. Centre row: Mean shape and appearance. Top row: Mean shape and appearance +2σ. Bottom row: Mean shape and appearance -2σ

4.2 Defining the feature vectors

The first approach towards lip reading and other similar problems was to use as visual features directly the AAM parameters. The other approach is to use the final results of the method, namely the co-ordinates of the landmarks as assigned by the algorithm for the current image. In our research we adopted this latter approach. Based on the position of the landmarks we defined seven high level geometric features. The features are computed as the Euclidean distances and areas between the certain key points that describe the shape of the mouth, namely mouth height and width, mouth aperture width and height, mouth area, aperture area and the nose to chin distance. The features are graphically described in Figure 7.

4.3 Visual validation of the feature vectors

Figure 8 shows the plots of the feature vectors computed for a random recording of the letter F having the viseme transcription [eeh fvw]. In this case the onset and offset moments of the utterance are clearly visible around the frame 75 and the frame 200 of the video recording, respectively. The onset of the viseme [eeh] is around the frame 80, while the onset of the viseme [fvw] is seen around frame 160. The actual shape of the mouth can be seen in the images shown below the graphs, which are extracted from the video sequence.

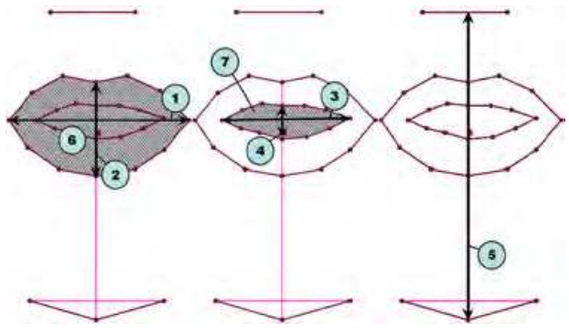


Fig. 7. The high level geometric features: 1) Outer lip width, 2) Outer lip height, 3) Inner lip width, 4) Inner lip height; 5) Chin to nose distance, 6) Outer lip area, 7) Inner lip area

Figure 9 shows the plots of the feature vectors for seven letters of the alphabet and the digit < 8 > ([a gkx td]). We see that the variability of the features is very high which makes them suitable for the recognition task at hand. We can also remark that, for instance, even though the viseme [aa] is present in the transcription of all letters, A([aa]), H([h aa]) and K([gkx aa]) we can clearly see that there is a slight difference between them with respect to the duration in each instance. This is best visible in the curve showing the height of the mouth, which shows that the duration of the viseme is shorter in the utterance of the letter K and H than in the case of the letter A.

An interesting result was obtained when visually inspecting the curves described by the feature vectors for all the visemes. By simple visual inspection we found that we could easily distinguish between some of the visemes, which proved that the feature set captures much of the speech related information. Table 6 summarises our findings in this respect. For a simple recognition task such as for instance the recognition of isolated visemes, or even the recognition of isolated digits, based on this table we could use a static classifier such as Support Vector Machines (SVM) (Ganapathiraju, 2002). However, for these types of

classifiers the features need to be global features because they cannot handle time series. Therefore, the generalisation to longer and of variable length utterances is not possible.

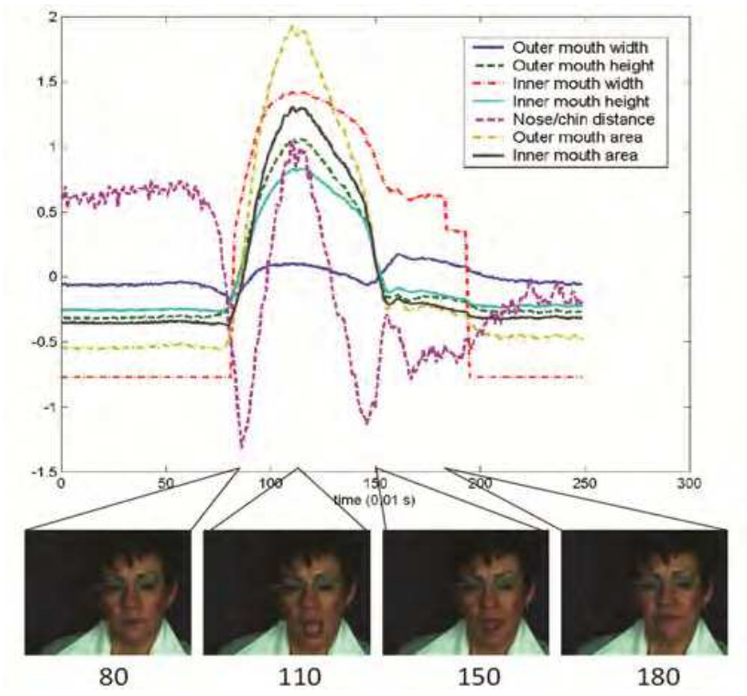


Fig. 8. The seven features plotted for one recording for the letter F transcribed using the visemes: eeh and fvw

	aa	h	gkx	a	oyu	ie	ei	iee	td	sz	eeh	l	pbm	fvw	at
Outer width	-		+	-	-	+	-+	+	+	+	+-	+-		+-	
Inner width	+		+	+		+	+	+			+		-		
Nose/chin dist	-	+	+	-	-	-	-	+-	+		+-	+-			
Height/area	+	+	+	+	-	+	+	+			+				

Table 6. Feature patterns per viseme: +) peak -) valley -+) increase +-) decrease.

4.4 AAM as ROI detection algorithm

It is worth mentioning that AAM can be used as well as a preprocess for defining a more accurate ROI. Therefore, the ROI defined using a mouth detection algorithm is further improved using the AAM. A more accurate ROI makes the data parametrization process more robust, because the background is better removed and, therefore, there is less noise in the input data.

5. Lip reading results

The method presented in this section produces for each frame in the corpus a vector with seven entries: mouth width, mouth height, aperture width, aperture height, mouth area, aperture area and the distance between the nose and the chin. We trained and tuned a lip reader based on the HMMs approach for each recognition task. In a similar approach, we

considered both the case with simple static features (i.e. seven geometric features) and the case when the feature space was enriched with dynamic information consisting of deltas and accelerations (i.e. making 21-dimensional vectors). We trained systems based on mono-visemes as well as context aware tri-viseme systems. We used a Gaussian mixture arrangement to better describe the feature space and we performed a 10-fold validation in order to increase the confidence in the observed results. The best results obtained were WRR 90.32% with word accuracy 84.27% for the CD recognition task. In this case, 75% of the sequences was recognized correctly. Figure 10 shows the plot of the performance of the best recognizer as a function of the number of Gaussian mixtures used.

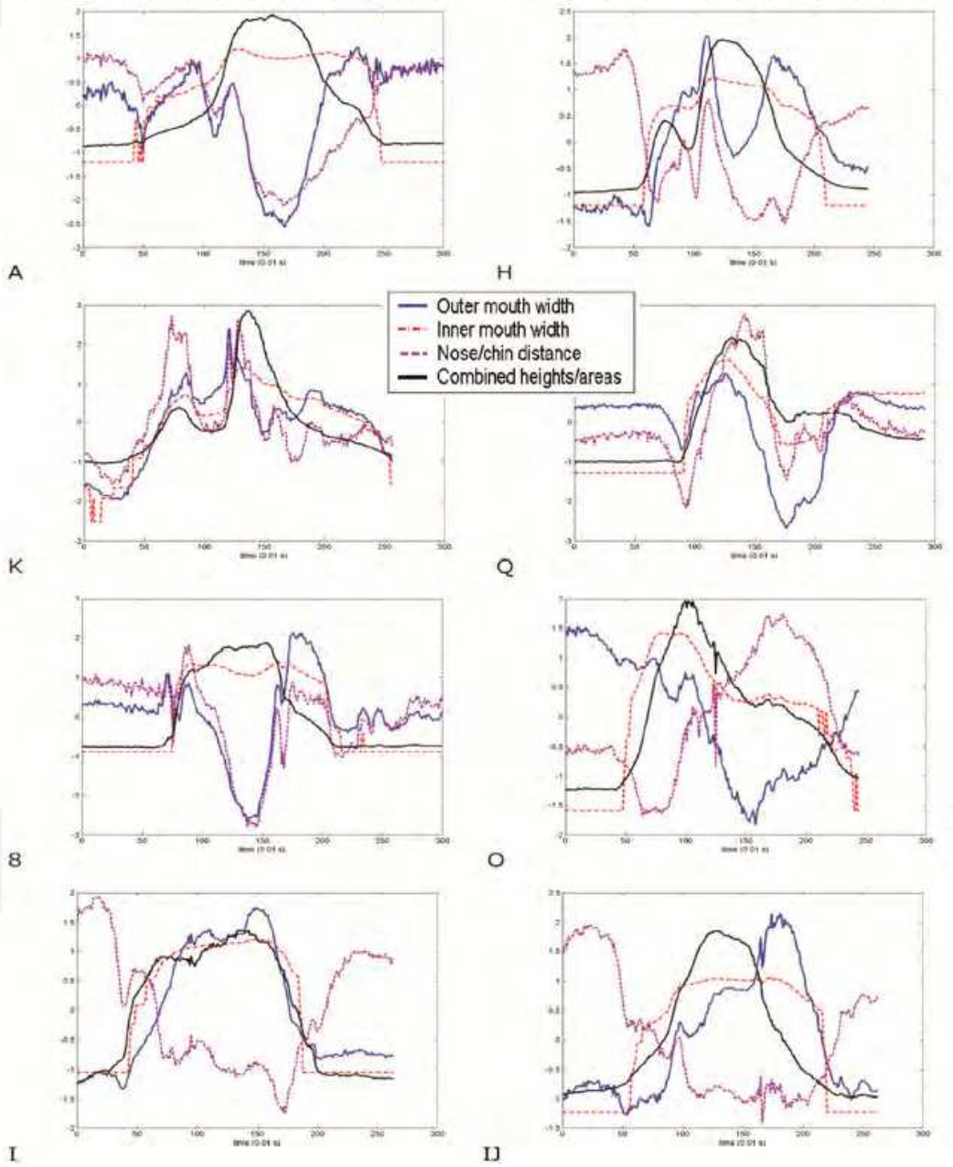


Fig. 9. Feature values plotted for the letters A ([aa]), H ([h aa]), K ([gkx aa]) and Q ([gkx oyu]), I ([ie]), O ([oyu]), IJ ([ei]) and 8 ([a gkx td]). The vectors are scaled using the time variance and centred around their mean

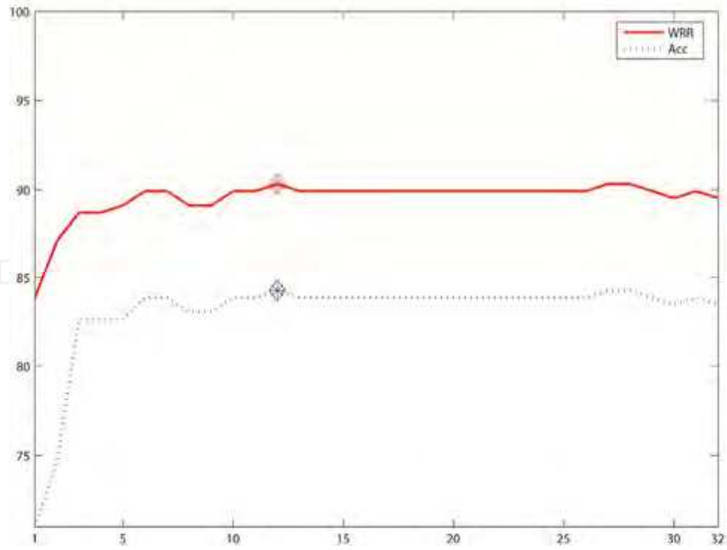


Fig. 10. The WRR and Acc results for CD recognition task as a function of the number of mixtures. The X axis gives the number of mixtures and the Y axis shows the results obtained. The feature vectors consisted of geometric features computed based on the AAM shape corroborated with their corresponding deltas and accelerations. The HMM models consisted of intra-word tri-visemes

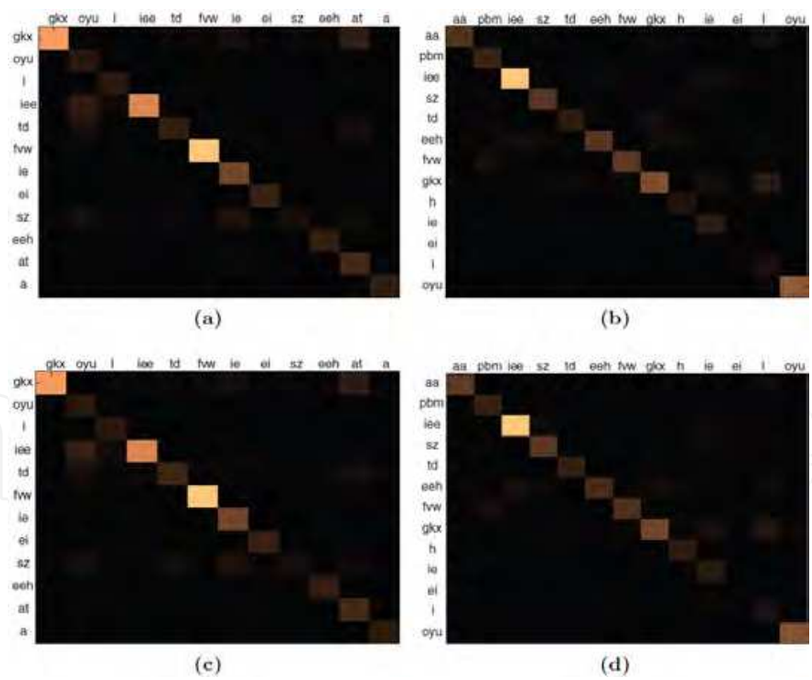


Fig. 11. The confusion matrices obtained by the best systems in the CD and CL tasks at the viseme level, respectively. a) the confusion matrix for CD task in the best case. b) the confusion matrix for CL task in the best case. c) the mean, over the mixture number, confusion matrix for the CD task. d) the mean, over the mixture number, confusion matrix for the CL task

For the GU recognition task we observed a 56% WRR. Using an N-Best approach with five most probable outcomes did not improve the result, which suggests the system is fairly robust. The 10-fold validation showed an 80.27% mean WRR with a 6% standard deviation, the minimum performance being 74.80% WRR. This shows some instability, however, the minimum is still a very good result. We also tested the results of the recognition at viseme level (i.e. before using the language model to build the corresponding words). This is useful for analysing the degree of confusion between different visemes. Figure 11 shows the confusion matrix for the best case. The mean confusion matrices computed over the mixture number is also displayed. We can remark in these figures that the degree of confusion is relatively small. However, the confusion is greater for visemes defined by larger phoneme sets. This is the case especially for the visemes [oyu] and [gkx] which are very often a source of confusion.

6. Conclusion

We introduced in this chapter an AAM based approach for lip reading. The AAM method is in our opinion a valuable tool for lip reading, both as a data parametrization method but also as a ROI detection technique. The method can be very robust and has a good generalization for unseen faces, however, the training process can be very long for satisfactory results to be obtained. Nevertheless, the shape obtained from the search scheme can be used as a starting point for testing other feature types, since it can always function as background elimination stencil. Based on the shape computed using the AAM searching scheme, we defined a set of high level geometric features. Based on these features we built different lip readers with very good results. These results validate the findings reported in the literature which showed that the width and the height of the mouth largely capture the content of the spoken utterance (Wojdel, 2003). This also justifies why a simple mouth model for lips synchronization based only on varying the mouth opening synchronous with the sound output is so convincing. We did not include in the feature vectors used in this chapter any information that describes the presence of the teeth, tongue or other elements of the mouth. This information was shown in the literature but also in our other experiments to be very important for lip reading. We expect that this is the case in the current settings as well. However, we did not include this information here because we wanted to have a clear understanding of the factors that influence the observed results.

7. References

Arsic, I. & Thiran, J.-P. (2006). *Mutual information eigenlips for audiovisual speech recognition*, In 14th European Signal Processing Conference (EU-SIPCO)

Atteveldt, N. van. (2006). *Speech meets script fMRI studies on the integration of letters and speech sounds*. Ph.D. thesis, Universiteit Maastricht

Beun, D. (1996). *Viseme syllable sets*, Master's thesis, Institute of Phonetic Sciences, University of Amsterdam

Boogaart, T.; Bos, L. & Bouer, L. (1994). *Use of the dutch polyphone corpus for application development*. In 2nd IEEE Workshop on Iterative Voice Technology for Telecomunication Applications. September

Breeuwer, M. (1985). *Speechreading Suplimented With Auditory Information*, Ph.D. thesis, Free University of Amsterdam

Bregler, C.; Hild, H. ; Manke, S. & Waibel, A. (1993). *Improving connected letter recognition by lipreading*. In IEEE International Conference on Acoustics Speech and Signal Processing, vol. 1. Institute of Electrical Engineers Inc (IEE)

Bregler, C. & Konig, Y. (1994). *Eigenlips for robust speech recognition*. In Acoustics, Speech, and Signal Processing, ICASSP-94 IEEE International Conference on

Buchan, J. N.; Pare, M. & Munhall, K. G. (2007). *Spatial statistics of gaze fixations during dynamic face processing*, Social Neuroscience, vol. 2(1), pp.1-13

Chibelushi, C. ; Gandon, S. ; Mason,J. ; Deravi, F. & Johnston, R. (1996). *Design issues for a digital audio-visual integrated database*, In Integrated Audio-Visual Processing for Recognition, Synthesis and Communication (Digest No: 1996/213), IEE Colloquium on

Chiou, G. I. & Hwang, J. N. (1997). *Lipreading from color video*, IEEE Transactions on Image Processing, vol. 6(8),pp. 1192-1195

Chitu, A. G. ; Rothkrantz, L.J.M. ; Wiggers, P. & Wojdel, J.C. (2007). *Comparison between different feature extraction techniques for audio-visual speech recognition*, In Journal on Multimodal User Interfaces, vol. 1,no. 1, pages 7-20, Springer, March

Chitu, A. G. & Rothkrantz, L. J. M. (2009). *The New Delft University of Technology Data Corpus for Audio-Visual Speech Recognition*. In Euromedia'2009, pp. 63-69. April

Chitu, A. G. ; Rothkrantz, L.J.M. (2009). *Visual Speech recognition- Automatic System for Lip Reading of Dutch*, In Journal on Information Technologies and Control, vol. year vii, no. 3, pages 2{9, Simolini-94, Sofia, Bulgaria

Cootes, T.; Edwards, G. & Taylor; C. (1998). *Active appearance models*, In H. Burkhardt and B. Neumann, editors, Proc. European Conference on Computer Vision 1998, vol. 2, pp. 484-498. Springer

Corthals , P. (1984). *Een eenvoudige visementaxonomie voor spraakafzien [a simple viseme taxonomy for lipreading]*, In Tijdscrijf Log en Audio, vol. 14, pp. 126-134

Daubias, P. & Deleglise, P. (2003). *The lium-avs database: a corpus to test lip segmentation and speechreading systems in natural conditions*, In Eighth European Conference on Speech Communication and Technology

Duchnowski, P.; Meier, U. & Waibel, A. (1994). *See me, hear me: Integrating automatic speech recognition and lip-reading*. Reading, vol. 1(1)

Duchnowski, P.; Hunke, M.; Büsching, D.; Meier, U. & Waibel, A. (1995). *Toward movement-invariant automatic lip-reading and speech recognition,* In International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95, vol. 1, pp. 109-112

Dupont, S. & Luettin, J. (2000). *Audio-visual speech modeling for continuous speech recognition*, In IEEE Transactions On Multimedia, vol. 2. September

Eggermont, J. P. M. (1964). *Taalverwerving bij een Groep Dove Kinderen [Language Acquisition in a Group of Deaf Children]*

Eveno, N.; Caplier, A. & Coulon, P.-Y. (2004). *Automatic and accurate lip tracking*, In IEEE Transactions on Circuits and Systems for Video technology, vol. 15, pp. 706-715. May

Fisher, C. G. (1968). *Confusions among visually perceived consonants*, Journal of Speech, Language and Hearing Research, vol. 11(4), p. 796

Fleet, D. J.; Black, M. J.; Yacoob, Y. & Jepson, A. D. (2000). *Design and Use of Linear Models for Image Motion Analysis*, International Journal of Computer Vision, vol. 36(3), pp. 171-193

Furui, S. (2003). *Robust Methods in Automatic Speech Recognition and Understanding*, In EUROSPEECH 2003 - Geneva

Ganapathiraju, A. (2002). *Support vector machines for speech recognition*, Ph.D. thesis, Mississippi State University, Mississippi State, MS, USA, 2002. Major Professor-Picone, Joseph

Goecke, R.; Tran, Q. N.; Millar, J. B.; Zelinsky, A. & Robert-Ribes, J. (2000). *Validation of an automatic lip-tracking algorithm and design of a database for audio-video speech processing*, In Proc. 8th Australian Int. Conf. on Speech Science and Technology SST2000, pp. 92-97

Goecke, R.; Millar, J. B.; Zelinsky, A. & Robert-Ribes, J. (2000). *Automatic extraction of lip feature points*, In Proc. of the Australian Conference on Robotics and Automation ACRA2000, pp. 31-36

Goecke, R. & Millar, J. (2004). *The audio-video australian english speech data corpus avoze,* In Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004, vol. III, pp. 2525-2528. Jeju, Korea, October

Garofolo, J. (1988) *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database*, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA

Gray, M. S.; Movellan, J. R. & Sejnowski T. J. (1997). *Dynamic features for visual speechreading: A systematic comparison*, Advances in Neural Information Processing Systems, vol. 9, pp. 751-757

Hilder, S.; Harvey, R. & Theobald, B. J. (2009). *Comparison of human and machine-based lip-reading*, In B. J. Theobald and R. W. Harvey, editors, AVSP 2009, pp. 86-89. Norwich, September

Hong, X.; Yao, H.; Wan,Y. & Chen, R. (2006). *A PCA based visual DCT feature extraction method for lip-reading*, pp. 321-326

Iwano, K.; Tamura, S. & Furui, S. (2001). *Bimodal Speech Recognition Using Lip Movement Measured By Optical-Flow analysis*, In HSC2001

Kricke, R.; Gernoth, T. & Grigat, R.-R. (2008). *Local binary patterns for lip motion analysis*. In Image Processing 2008, 15th IEEE International Conference on, pp. 1472-1475

Kumar, K.; Chen, T. & Stern, R. M. (2007). *Profile view lip reading*, In Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP, vol. 4, pp. 429-432

Lee, B.; Hasegawa-Johnson, M.; Goudeseune, C.; Kamdar, S.; Borys, S.; Liu, M. & Huang, T. (2004). *Avicar: Audio-visual speech corpus in a car environment*, In INTERSPEECH2004-ICSLP. Jeju Island, Korea, October

Damhuis, M.; Boogaart, T.; Veld, C.; Versteijlen, M.; Schelvis, W.; Bos, L. & Boves, L. (1994). *Creation and analysis of the dutch polyphone corpus*, In Third International Conference on Spoken Language Processing. ISCA

Li, N.; Dettmer, S. & Shah, M. (1995). *Lipreading using eigen sequences*, In Proc. International Workshop on Automatic Face- and Gesture-Recognition, pp.30-34. Zurich, Switzerland

Li, N.; Dettmer, S. & Shah, M. (1997). *Visually recognizing speech using eigensequences*, Motion-based recognition, vol. 1, pp. 345-371

Lievin, M.; Delmas, P.; Coulon, P. Y.; Luthon, F. & Fristot, V. (1999). *Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme*, In IEEE Conference

on Multimedia, Computing and Systems, ICMCS99, vol. 1, pp. 691-696. Fiorenza, Italy, June

Lucey, P. & Potamianos, G. (2006). *Lipreading using profile versus frontal views*, In IEEE Multimedia Signal Processing Workshop, pp. 24-28

Luettin, J.; Thacker, N. A. & Beet, S. W. (1996). *Statistical lip modelling for visual speech recognition*, In Proceedings of the 8th European Signal Processing Conference (EUSIPCO96)

Luettin, J. & Thacker, N. A. (1997). *Speechreading using probabilistic models*, Computer Vision and Image Understanding, vol. 65(2), pp. 163-178

Martin, A. (1995). *Lipreading by optical flow correlation*, Technical report, Compute Science Department University of Central Florida

Mase, K. & Pentland, A. (1991). *Automatic lipreading by optical-flow analysis*, In Systems and Computers in Japan, vol. 22, pp. 67-76

Matthews, I. A.; Bangham, J. & Cox, S. J. (1996). *Audiovisual speech recognition using multiscale nonlinear image decomposition*, In Fourth International Conference on Spoken Language Processing

Matthews, I., Cootes, T. F.; Bangham, J. A.; Cox, S. & Harvey, R. (2002). *Extraction of visual features for lipreading*, In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 198-213

Mcgurk, H. & Macdonald, J. (1976). *Hearing lips and seeing voices*, Nature, vol. 264, pp. 746-748, December

Messer, K.; Matas, J.; Kittler, J.; Luettin, J. & Maitre, G. (1999). *XM2VTSDB: The Extended M2VTS Database,* In Audio- and Video-based Biometric Person Authentication, AVBPA'99, pp. 72-77. Washington, D.C., March

Morn, L. E. L & Pinto-Elas, R. (2007). *Lips shape extraction via active shape model and local binary pattern*. MICAI 2007: Advances in Artificial Intelligence, vol. 4827, pp. 779-788

Movellan, J. R. (1995). *Visual Speech Recognition with Stochastic Networks*, In Advances in Neural Information Processing Systems, vol. 7. MIT Press, Cambridge

Nefian, A. V.; Liang, L.; Pi, X.; Liu, X. & Murphy, K. (2002). *Dynamic bayesian networks for audio-visual speech recognition*, EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1274-1288

Neti, C.; Potamianos, G.; Luettin, J.; Matthews, I.; Glotin, H.; Vergyri, D.; Sison, J.; Mashari, A. & Zhou, J.(2000). *Audio-visual speech recognition*, In Final Workshop 2000 Report, vol. 764

Ojala, T. & Pietikainen, M. (1997). *Unsupervised texture segmentation using feature distributions*, Image Analysis and Processing, vol. 1310, pp. 311-318

Patterson, E.; Gurbuz, S. ; Tufekci, Z. & Gowdy, J. (2002). *CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research*, In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing

Petajan, E., Bischoff, B. & Bodoff, D. (1988). *An improved automatic lipreading system to enhance speech recognition,* In CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 19-25. ACM Press, New York, NY, USA

Pigeon, S. & Vandendorpe, L. (1997). *The M2VTS multimodal face database(release 1.00)*, Lecture Notes in Computer Science, vol. 1206, pp. 403-410

Potamianos, G.; Cosatto, E.; Graf, H. & Roe, D. (1997). *Speaker independent audio-visual database for bimodal ASR*, In Proc. Europ. Tut. Work. Audio-Visual Speech Proc., Rhodes

Potamianos, G.; Graf, H. P. & Cosatto, E. (1998). *An image transform approach for hmm based automatic lipreading*, In Proc. IEEE International Conference on Image Processing, vol. 1

Potamianos, G.; Neti, C.; Luettin, J. & Matthews, I. (2004). *Audio-visual automatic speech recognition: An overview*, Issues in Visual and Audio-Visual Speech Processing

Prez, J. F. G.; Frangi, A. F.; Solano, E. L. & Lukas, K. (2005). *Lip reading for robust speech recognition on embedded devices*, In Int. Conf. Acoustics, Speech and Signal Processing, vol. I, pp. 473-476

Salazar, A.; Hernandez, J. & Prieto, F. (2007). *Automatic quantitative mouth shape analysis*, Lecture Notes in Computer Science, vol. 4673, pp. 416-421

Son van, N.; Huiskamp, T. M. I.; Bosman, A. J. & Smoorenburg, G. F. (1994). *Viseme classifications of Dutch consonants and vowels*, The Journal of the Acoustical Society of America, vol. 96

Tamura, S.; Iwano, K. & Furui, S. (2002). *A robust multi-modal speech recognition method using optical-flow analysis*, In Extended summary of IDS02, pp. 2-4. Kloster Irsee, Germany, June

Tamura, S.; Iwano, K. & Furui, S. (2004). *Multi-modal speech recognition using optical-flow analysis for lip images*, Journal VLSI Signal Process Systems, vol. 36(2-3), pp. 117-124

Tomlinson, M. J.; Russell, M. J. & Brooke, N. M. (1996). *Integrating audio and visual information to provide highly robust speech recognition*, In IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2

Viola, P. & Jones, M. (2001). *Robust Real-time Object Detection*, In Second International Workshop On Statistical And Computational Theories Of Vision Modelling, Learning, Computing, And Sampling. Vancouver, Canada, July

Visser, M.; Poel, M. & Nijholt, A. (1999). *Classifying visemes for automatic Lipreading*, Lecture notes in computer science, pp. 349-352

Williams, J. J.; Rutledge, J. C. & Katsaggelos, A. K. (1998). *Frame rate and viseme analysis for multimedia applications to assist speechreading*. Journal of VLSI Signal Processing, vol. 20, pp. 7-23

Wojdel, J. C. & Rothkrantz, L. J. M. (2000). *Visually based speech onset/offset detection*, In Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia 2000), pp. 156-160. Antwerp, Belgium

Wojdel, J.; Wiggers, P. & Rothkrantz, L.J.M. (2002). *An audio-visual corpus for multimodal speech recognition in dutch language*, In ICSLP, Conference Proceedings of

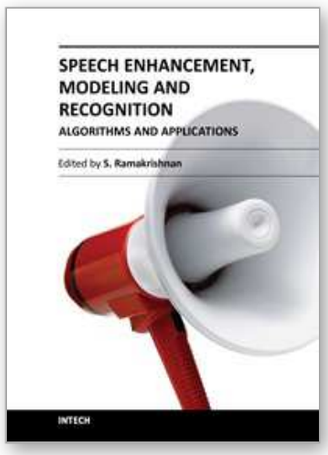
Wojdel, J. C. (2003). *Automatic Lipreading in the Dutch Language*, Ph.D. thesis, Delft University of Technology, November

Yoshinaga, T.; Tamura, S.; Iwano, K. & Furui, S. (2003). *Audio-Visual Speech Recognition Using Lip Movement Extracted from Side-Face Images*, In AVSP2003, pp. 117-120. September

Yoshinaga, T.; Tamura, S.; Iwano, K. & Furui, S. (2004). *Audio-visual speech recognition using new lip features extracted from side-face images*

Zhang, X.; Broun, C. C.; Mersereau, R. M. & Clements, M. A. (2002). *Automatic speechreading with applications to human-computer interfaces,* EURASIP Journal Appl Signal Process, vol. 2002(1), pp. 1228-1247

Zhao, G., Pietikäinen, M. & Hadid, A. (2007). *Local spatiotemporal descriptors for visual recognition of spoken phrases*, In Proceedings of the international workshop on Human-centered multimedia, pp. 66-75. ACM



Speech Enhancement, Modeling and Recognition- Algorithms and Applications

Edited by Dr. S Ramakrishnan

ISBN 978-953-51-0291-5
Hard cover, 138 pages
Publisher InTech
Published online 14, March, 2012
Published in print edition March, 2012

This book on Speech Processing consists of seven chapters written by eminent researchers from Italy, Canada, India, Tunisia, Finland and The Netherlands. The chapters covers important fields in speech processing such as speech enhancement, noise cancellation, multi resolution spectral analysis, voice conversion, speech recognition and emotion recognition from speech. The chapters contain both survey and original research materials in addition to applications. This book will be useful to graduate students, researchers and practicing engineers working in speech processing.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Alin Chițu and Léon J.M. Rothkrantz (2012). Automatic Visual Speech Recognition, Speech Enhancement, Modeling and Recognition- Algorithms and Applications, Dr. S Ramakrishnan (Ed.), ISBN: 978-953-51-0291-5, InTech, Available from: http://www.intechopen.com/books/speech-enhancement-modeling-and-recognition-algorithms-and-applications/towards-robust-visual-speech-recognition

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen