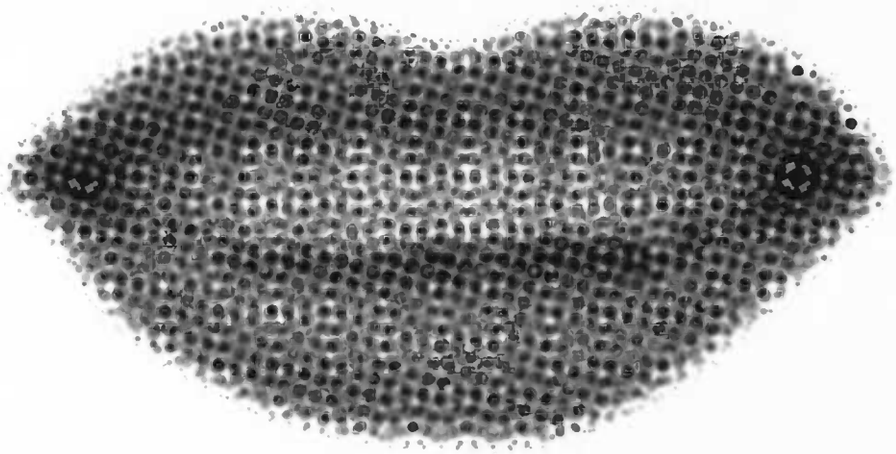
# Hue-based Automatic Lipreading



Peter Duifhuis
studentnr. 0968838
26 Augustus, 2003

Supervisor:
Esther Wiersinga-Post

Artificial Intelligence
Rijksuniversiteit Groningen

# Hue-based Automatic Lipreading

Peter Duifhuis
studentnr. 0968838
August 26, 2003

Supervisor:
Esther Wiersinga-Post

Artificial Intelligence
Rijksuniversiteit Groningen

# Contents


**1 Introduction** 5
- 1.1 What a piece of work is man! . . . 5
- 1.2 Talking with machines . . . 6

**2 Theoretical Background** 8
- 2.1 Automatic speech recognition . . . 8
  - 2.1.1 Automatic speech recognition at work . . . 8
  - 2.1.2 The state of the art . . . 10
  - 2.1.3 The problems . . . 11
- 2.2 Human speech reading . . . 12
  - 2.2.1 The McGurk effect . . . 13
  - 2.2.2 Confusing phonemes and visemes . . . 13
  - 2.2.3 What people look at when reading speech . . . 14
- 2.3 Audio-visual speech recognition . . . 16
  - 2.3.1 Audio-visual speech recognition at work . . . 16
  - 2.3.2 State of the art . . . 18
  - 2.3.3 Problems . . . 18

**3 HSB-based Automatic Lip detection** 20
- 3.1 Requirements . . . 20
- 3.2 Method . . . 21
  - 3.2.1 Subjects . . . 21
  - 3.2.2 How to begin . . . 21
  - 3.2.3 A simple parabolic filter . . . 22
  - 3.2.4 Locating the region of interest . . . 23
  - 3.2.5 Measuring higher level features . . . 23
  - 3.2.6 Test procedure . . . 24

**4 Evaluation** 26
- 4.1 Method of evaluation . . . 26
  - 4.1.1 Common errors . . . 26
  - 4.1.2 Quantitative results . . . 29
  - 4.1.3 Judging whether a speaker is silent or not . . . 32

**5 Conclusion** 34
- 5.1 Recommendations . . . 35

**6 Acknowledgments** 37

**A International Phonetic Alphabet** 41

**B The subjects** 43

**C Comparing filters** 45
C.1 Overview . . . . . 45
C.2 Parabolic filtering with hue, saturation and brightness . . . . . 45
C.3 Red/green threshold . . . . . 46
C.4 Red/green colour burn . . . . . 46

**D Hue-based Automatic Lip-detection: Images** 48

**E Results: traces** 59

**F Results: some examples of phonemes and the corresponding shape of the mouth** 71

**G Software** 74

# Abstract

While they might not even notice it, humans use their eyes when they are understanding speech. Especially when the quality of the sound deteriorates, the visual counterpart can contribute considerably to the intelligibility of speech.

Artificial speech recognizers have great difficulty with discerning speech from varying background noise. We can learn from humans that incorporating visual information in the recognition process, can be a fruitful approach to this problem. The field of artificial audio-visual speech recognition is indeed a popular and growing one, with still a lot of territories to explore.

An overview of audio-visual speech recognition today is given, as well as an investigation into where visual speech processing can really contribute to speech recognition. Three different methods are discerned, namely:

- Detecting whether there is a speaker at all.
- Knowing when someone is speaking or silent.
- Distinguishing similar sounding phonemes.

A system was created with the purpose of exploring the problems and possibilities of audio-visual speech recognition in 'real-life' situations, without the help of artificial circumstances to facilitate recognition. This system estimates a set of features that can be used for distinguishing similar phonemes, and for estimating whether a speaker is silent or not. Although it has not been implemented, the system could very well be expanded to detect whether there is a speaker at all.

It was found that detecting the whereabouts of a mouth in a video frame, with the precondition that the image contains a face at a certain distance, can be done in a simple en computationally cheap manner. This method is based primarily on the selection of pixels with a certain hue, and to a lesser degree saturation and brightness. The extraction of features such as the region of interest, the height and width of the outer contour and the height of the inner contour of the mouth, renders varying results. Some subjects give very good results, whereas others give poor results.

The main problems lie in articulation and the differences between speakers. In continuous speech, visemes are heavily influenced by surrounding visemes, and therefore it is hard to discern them accurately. Furthermore, due to the differences between speakers it is hard to create a single system that works well for all subjects. Speakers articulate differently[1] and although lips have a similar hue, the distribution of colour of the faces differs as well.

With regard to the methods of improving auditory speech recognition, the discrimination between phonemes will most probably be very difficult with this system. Although the system can predict reliably whether a mouth is opened or closed, other viseme-related features such as 'rounded' or 'spread', are hard to categorize. Next to unclear articulation, this is because in continuous speech visemes are heavily influenced by surrounding visemes. It is estimated that detecting whether a speaker is silent or speaking can only be done in situations where the speaker closes his mouth for a longer period of time.


[1] As could be expected, see [20].

To conclude, a crude method has been implemented that can be used for further research. Not only can the lip detection be refined, this system also begs the development of a module that classifies the estimated features. Aside from speech recognition, the method for detecting areas of a certain colour may prove successful in a lot more applications.

# Chapter 1

# Introduction

## 1.1 What a piece of work is man!

*...In apprehension, how like a god!*[1]

When humans are having a conversation, the most important part of understanding one another is *hearing* what the other has to say. The actual hearing is a complicated process. Waves travel through the air to reach your ear, the eardrum sends vibrations to the tiny ossicles, which beat on the cochlea. The cochlea transforms vibrations into electrical signals and sends them to your brain. The brain somehow transforms these pulses into intelligible speech.

Transforming these signals into something you can understand is a very complicated process, but your brain handles it in a very smart way. For example, if the auditory input stream is disrupted, the brain is very good at filling in the blank spots. Say, you're having a discussion with a friend in the kitchen, and the loud *ping!* from the microwave oven causes you to miss a word your friend is saying. You probably would not even notice it! Or if someone says: *"Your parents were very mice"*, you'd be inclined to hear: *"Your parents were very nice"*, simply because that would be a lot more likely thing to say[2]. The process of understanding speech is very robust. The mind can make mistakes and may take a wrong guess when filling in the blank spots. You've probably experienced the confusing situation where you misheard someone.

Next to 'guessing the gaps', another way of improving perception is looking at the speakers face. Facial expressions can greatly help you in guessing what message the speaker is trying to convey, and looking at the movements of the mouth may help to distinguish between confusing sounds. The sounds for /*m*/[3] and /*n*/ for example, are very similar, but if you see someone saying either *mice* or *nice* with an open mouth when pronouncing the first sounds, your eyes will tell you you could not have heard *mice*. We see that hearing includes more than sound alone.

[1] The title and this quote are both from William Shakespeare: *Hamlet, Prince of Denmark*, 1601

[2] Maybe you even had to reread the examples, because you did not notice the word *mice* the first time. The same thing occurs when you hear someone speak.

[3] Throughout this thesis phonetic spelling will be enclosed in /'s. The SAMPA alphabet [28] is used. See also appendix A.

## 1.2 Talking with machines

We are able to understand each other in a quiet office, but also in a crowded discotheque or across the street. We can even read one's lips mouthing *I love you* behind a window. Humans can speak with each other in variable and unpredictable situations. But when we try to engineer a machine that can understand speech as good as a human can, the results are disappointing.

For at least half a century scientists and engineers have been trying to make machines that can interpret speech. Estimates that a truly robust automatic speech recognition system is about 5 to 10 years away are regularly reiterated [34]. At the moment, some applications that do automatic speech recognition are created and sold, but these systems typically operate under limited conditions. Examples are telephone services and dictation programs. Their use is not very widespread, because for the telephone services, it only works if the customer is allowed to say a very limited set of words[4]. Dictation programs require a lot of training per user, which people usually do not consider worth the effort.

Figure 1.1: HAL reading lips in 2001: A Space Odyssey

The question is whether the scene from Stanley Kubrick's 1968 masterpiece movie *2001: A Space Odyssey*, where the man-made computer HAL 9000 'overhears' a conversation of the unfortunate crew by reading their lips, is just a fantasy, or that it could happen as soon as 2010?

Artificial Intelligence is a field of research that focuses on the way natural, cognitive systems solve problems. Humans have been speaking with each other for millennia and the most complex techniques have emerged throughout evolution. Consequently, it may be fruitful to let the human be an inspiration for developing an artificial speech recognizer. One of these techniques for better understanding is looking at the speaker. Thus, one of the things we can copy is making use of available visual information. That will be the central question of this thesis:

> How can automatic speech perception based on sound alone, improve with automatic lipreading?

[4] e.g. only numbers

This thesis is an exploration of the field of audio-visual speech recognition. What is the state of the art of speech recognition today, where lie the problems, and where can visual input contribute? When and how do humans make use of visual information when they are interpreting speech, and what can the field of automatic speech recognition gain by incorporating similar methods? Furthermore, an attempt is made to create a system that can extract meaningful features from video recorded speakers. The focus here lies on discovering the problems and possibilities when employing a speech reading system in 'real-life', where one has but limited control over the input.

# Chapter 2

# Theoretical Background

## 2.1 Automatic speech recognition

In the 1950's researchers at Lincoln Labs first started work on automatic speech recognition (ASR) [13]. Several developments in the fields of digital signal processing and pattern recognition, such as the Viterbi algorithm and Hidden Markov Models, had a great influence on the way most automatic speech recognition systems operate today. This section is an overview of research and accomplishments in the field of automatic speech recognition until today.

### 2.1.1 Automatic speech recognition at work

Nowadays, most ASR systems are based on finding the best match between a sequence of observations and a sequence of possible utterances [2, 35]. The set of possible utterances makes up the **corpus** of the recognition system. Entries in the corpus can vary from a limited set of words (e.g. *"zero"*, *"one"*, *"two"*, ..., *"nine"*) to entries for each phoneme[1] in a language.

The process of speech recognition can be divided into three stages. The **first stage** is the transforming of sound into a waveform. The sound, which includes speech, is recorded by a microphone and transformed into a digital signal. In the **second stage** the waveform is digitally filtered, and relevant features are extracted. Periodically a feature vector is calculated that represents an instant in time. The values of such a vector can for example relate to the magnitude, or the change in magnitude, of a range of frequencies in the original signal. Because one feature vector represents an instant in time, a stream of speech is represented by a sequence of feature vectors.

In the **third stage** the feature vectors are used to estimate the most likely utterance from the corpus. The central rule for computing the likelihood of an utterance, given a sequence of feature vectors, is Bayes' decision rule:

$$P(w|y) = \frac{P(y|w)P(w)}{P(y)} \tag{2.1}$$

Here, $y$ is the observation of a set of feature vectors, $w$ is an utterance from the corpus. $P(y)$ is the probability of observing $y$, $P(w)$ the probability of observing

[1] Phonemes will be discussed in greater detail in section 2.2.2.

$w$, and $P(y|w)$ the probability of observing the feature vectors $y$ given utterance $w$. $P(w)$ and $P(y)$ are relatively easy to estimate. If one has a large dataset of which one has all sounds and know the utterances that are made, for example the sound consists of the audio track from the movie *2001, A Space Odyssey*, and one has the script with a transcription of all spoken text, one can count all occurrences of $y$ in the audio track, and count all occurrences of $w$ in the script. Say the corpus has entries for syllables, then $y$ could for example be the sequence of feature vectors from the average sound for */pod/*, and $w$ could be the syllable *"pod"* taken from the script. Representing $P(y|w)$ is the most difficult task. The common approach is to use Hidden Markov Models (HMMs).

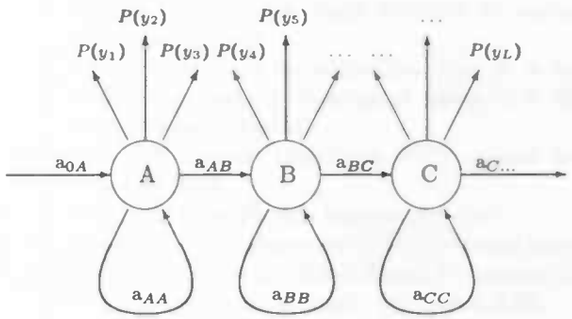



Figure 2.1: A three state Hidden Markov Model. A speech recognition system utilizes a large set of such models for each of the utterances (e.g., words).

For each entry in the corpus a Hidden Markov Model is trained. The HMM in figure 2.1 consists of three states (A, B and C). These states could represent for example the onset, nucleus and coda[2] of a syllable. The weights $a_{AA}$, $a_{AB}$, ..., define the probabilities of going from one state to another. For a long */o:/* nucleus, the recurrent $a_{BB}$ would be higher than for a short */o/*. The probabilities $P(y_1)$, $P(y_2)$, ..., $P(y_L)$, define the chance of observing feature vector $y_1$, $y_2$, ..., $y_L$, in the corresponding state. The Viterbi algorithm[3] is used for estimating whether the utterance this HMM represents matches the observed sequence of feature vectors. Together the values of $a_{AA}$, $a_{AB}$, ..., and $P(y_1)$, $P(y_2)$, ..., define $P(y|w)$ according to equation 2.2.

$$P(y|w) = a_{0\pi_1} \prod_{i=1}^{L} P(y_i) a_{\pi_i \pi_{i+1}} \tag{2.2}$$

Here $\pi_i$ is the state that corresponds to $P(y_i)$. Now, when an unknown sequence of observations is being processed, the HMM that predicts the highest chance of observing this sequence is the HMM modeled for the utterance that the system

[2]The onset, nucleus and coda of a syllable correspond to the start, middle and end of the syllable

[3]The Viterbi algorithm is an algorithm that computes the likelihood of a sequence of states with a given observation.

will recognize. Defining the corpus and training the HMMs are the challenges of this type of speech recognition. Training all HMMs is a vast and time-consuming task.

That sums up the three stages for the recognition process. As there are a lot of variations on this process of speech recognition, this is only a global indication of what is the machinery in the average speech recognition system.

### 2.1.2 The state of the art

| | |
|---|---|
| 1876 | Graham Bell invents the telephone. Attempts to reduce the necessary bandwidth give a research in speech synthesis and perception a boost. |
| 1950's | Lincoln Laboratories begin research in automatic speech recognition. |
| 1967 | Viterbi introduces an algorithm, that is a way of finding the most likely series of states in a Hidden Markov Model (HMM) |
| ca. 1967 | The Fast Fourier Transform (FFT) makes its entrance in ASR. |
| 1970's | Hidden Markov Models become popular. |
| 1980's | Artificial Neural Networks (ANN) become popular. |
| 1980's | The introduction of Digital Signal Processing (DSP) chips on the markets greatly facilitates ASR. |
| 1986 | IBM exhibits the Tangora system, a user specific isolated word recognition program. |
| 1992 | AT&T deploys an automated telephone service. |
| 1995 | Apple introduces dictation systems for fluent speech. |

Table 2.1: A brief history of automatic speech recognition, adapted from [13].

Speech recognition is an active field of research in Artificial Intelligence. Teams of researchers throughout the world are working on robust recognition systems, and to measure progress several tests were devised. The Natural Institute of Standards and Technology (NIST) creates such tests, an example of which is the Hub 4 Broadcast News evaluation [24]. In total 30.8 hours of news have to be recognized, within a time that is less than ten times the length of the original signal. The news material is especially challenging because of the wide variety in sound. The background noise varies greatly, and there are intervals with no speech at all. Sometimes the speech is compressed to a small bandwidth, and the speakers themselves have different accents, intonations, etcetera. The results on this test vary around a word error rate[4] of 14% to 20% [25, 23, 7]. Another test created by NIST is the Hub 5 Conversational Telephone evaluation. The dataset consists of natural telephone conversations. The lowest word error rates roughly vary from 20% to 30% for different sets [8].

Although these results seem pessimistic, various automatic speech recognition systems are used in 'real-life'. Telephone services exist, which can take

[4] Word error rate (WER) is a common measure for ASR systems. A word error rate of 20% means that 20% of the words were incorrectly recognized

input from multiple users. Typically, these systems assume a very rigid dialog structure, because the more a systems knows about the context, the better it can estimate what is said. Dictation programs have been around since 1995, but still rely on a lot of speaker-dependent training. An example are Microsoft's palmtop computers that can be primarily speech controlled [18]. Furthermore, a lot of cell phones are equipped with voice dialing: saying a name will make it call the person it is trained for. We see that numerous companies are trying very hard to improve and sell speech recognition software, but at this moment there are very few successful implementations, other than systems that only operate under strict conditions.

### 2.1.3 The problems

Why is it so hard to create a proper recognition system, that works well under different conditions? Humans easily recognize the familiar sound of a human voice, and it is even easy to discern different voices. Humans are able to localize a speaker, because they can estimate where a sound comes from. If humans see a speaker, they can match the sounds he makes to the movement of his lips. When the speaker is silent, they know no speech is uttered. In other words, humans have little difficulty with selecting that part of the environmental sound, which is speech.

Computers, on the other hand have great difficulty with discerning speech from background noise. If the speech is uttered in a situation where the background sounds are predictable, most systems render reasonable results. But in a complex auditory environment, the results degrade catastrophically. The computer has to know when speech is uttered, and when a speaker is silent, otherwise background noise will be interpreted as speech. Furthermore, different speakers have different voices, different intonation, different accents, which makes the set of intelligible utterances a whole lot wider. These problems all boil down to the **signal-in-noise paradox** [2]:

> Selecting the desired signal can be done if the noise is known. But only once the desired signal is selected properly, one knows what part of the signal is noise.

The limitations of Hidden Markov Models may be an obstacle in the way of solving this problem. Since the likelihood of observing certain utterances in a state depends on the current state alone and not the states before, HMMs assume that no correlation exists between input observations over time. In the case of a noisy versus clean observation, one HMM will not be able to make use of knowledge about the noise. In order to compensate for this, different models have to be made for a noisy and a clean utterance. In contrast, if a human is listening to someone in a noisy situation, he will expect noisy speech to follow.

As stated above, humans seem to solve this paradox all at once; selecting speech from all of the incoming sounds from the environment hardly seems to make comprehending speech more difficult. This thesis focuses on one of the mechanisms the human uses, namely looking at the speakers face to improve recognition.

## 2.2 Human speech reading

Just by looking, we can tell where a sound is coming from, and whether a speaker is silent or speaking. When hearing deteriorates, people rely more and more on lipreading. For the deaf, lipreading can be the most important way to perceive speech. Sumby and Pollack (1954) did research where intelligibility of speech increases around 80% at different signal to noise ratios (SNRs)[5], if the auditory information is accompanied by visual information (when the subjects could see the speaker clearly)[6] [30]. This ratio means that the visual information compensates up to 80% of the intelligibility loss caused by the noise poluting the audio signal. Interesting to note is that this ratio does not change much for different signal to noise ratios. For a signal to noise ratio of 0 dB the auditory information alone is enough for near perfect recognition. In another experiment by Risberg and Lubker from 1978 (described in [29]), subjects saw a speakers face, and heard a low-pass filtered auditory signal of the speaker. With the visual information alone, the subjects recognized 1% correctly, and with auditory input alone a mere 6%, but if the visual and auditory information was presented, 45% of the words was recognized correctly. The characteristic increase of intelligibility due to additional visual information is depicted in figure 2.2.

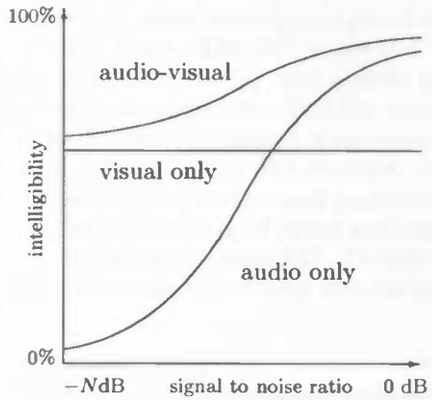



Figure 2.2: Characteristic intelligibility scores at an increasing signal to noise ratio, for audio only, visual only and audio-visual speech recognition by humans. Here $N$ could be a value around 20 or 30. The exact scores depend on the nature of the different stimuli, the size of the set of words which have to be recognized, etcetera.

[5] A signal to noise ratio of +1 dB implies that the speech signal was 1 decibel louder than the background noise.

[6] The intelligibility also depends on the context wherein the auditory stimulus is presented; in the case of Sumby and Pollack words had to be recognized, out of different sets with sizes varying from 8 to 32 words. Signal to noise ratios from -30dB to +∞dB (no noise) were investigated.

### 2.2.1 The McGurk effect

A striking example of how the mind uses visual input when understanding speech is what is called the McGurk effect. This effect was first described by McGurk and MacDonald in 1976 [17], and shows how vision can greatly influence speech perception. In an experiment, the researchers showed their subject the face of a woman uttering sounds like */gaga/*, dubbed with the sound of the woman saying */baba/*. The subjects then had to repeat the sound they heard. With closed eyes, thus ignoring the visual stimulus, the recognition rate was very high. But when hearing and seeing both stimuli, the larger part of the subjects reported hearing */dada/*! Even though the subjects could correctly reproduce the stimulus by hearing alone, if they also used what they saw, they *heard* something different. So not only can visual information aid in speech perception, visual input can influence speech perception even in cases where the auditory information is clearly audible.

The confusion can be explained when looking at different aspects of the sounds */d/*, */g/* and */b/*. The aspects to consider here are the *place* and *manner* of articulation[7]. In this case the place of the visual input is different from the place of the auditory input. The subject sees an opened mouth when the woman pronounces the velar */g/* but actually hears the bilabial */b/*. The subject seems to combine visually perceived place, namely an opened mouth, with auditory perceived manner, namely plosive. The */d/* is the best fit. It is a better fit than */g/*, because it sounds more like */b/* and visually the */g/* and */d/* look very much alike. However, the simplification that the visually perceived place is combined with the auditory perceived manner is sometimes too straightforward [31], it is often some 'in between' form. For example, the subject sometimes reports hearing both the acoustically and visually perceived place (such as */bga/*. An overview of some of the combinations of visual and auditory consonants and the way they were perceived is given in table 2.2[8]. The effect is most pronounced where a bilabial auditory utterance is combined with the lip movements of a non labial utterance [15].

| Visual | Audio | Perceived |
|---|---|---|
| */ga/* | */ba/* | */da/* 64.0%, */ga/* 27.0%, */ba/* 9.0% |
| */ba/* | */ga/* | */ga/* 83.0%, */bga/* 17.0% |
| */ka/* | */pa/* | */pa/* 70.0%, */ta/* 10%, */ka/* 10%, */tha/* 10% |
| */pa/* | */ka/* | */ka/* 82.0%, */pa/* 9.0%, */pka/* 9.0% |

Table 2.2: The McGurk effect: some examples of what subjects reported hearing with different auditory and visual stimuli, from [17].

### 2.2.2 Confusing phonemes and visemes

In section 1.1 we saw that the sounds */m/* and */n/* are very easily confused, but if we are looking at the speaker, it is easy to discern them. The shortest meaningful speech sound is called a **phoneme**. Phonemes can thus be discerned

[7] See appendix A

[8] For a more complete overview, including a quantitative measure of responses see [17, 15].

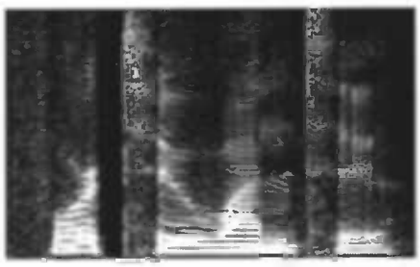


Figure 2.3: A **phoneme** is audible. A spectral sound pattern gives insight into what frequencies contribute the most to a specific phoneme. The horizontal axis is the time scale, and the vertical axis is the frequency scale.

Figure 2.4: Shape and movements of the mouth are aspects of a viseme.

by the (change in) phase and amplitude of certain frequencies. The visual aspects of a phoneme is referred to as the **viseme**. Aspects of a viseme are the shape and movements of the mouth, but also the chin, cheek and eyes may contribute to the visual intelligibility of a viseme. See figure 2.3 and 2.4 for a visual explanation. So we can say that the phonemes for */m/* and */n/* are very easily confused but the visemes are not. We have seen with the McGurk effect, that for example the phonemes for */b/* and */d/* sound very similar. A comparison of confusion among visemes and phonemes gives insight into where visual information can contribute the most.

Miller and Nicely [19] investigated the auditory confusion among consonants, and Walden et al. [32] did research into confusion among consonants that were visually presented. Summerfield [31] compares the results of these two experiments. In the experiment where the auditory confusion among consonants was measured, different signal to noise ratios were used. The results in figure 2.5 show that confusion among different categories increases with a lower signal to noise ratio, as one expects. For example, when the background noise is 6 dB louder than the speech signal, confusion within the three groups voiceless, nasal or voiced, is so high that the different members of these groups can not be discerned. So the */m/* and the */n/* (both nasal) can not be discerned, but the */m/* and the */g/* (nasal and voiced) can. Next to the auditory confusion, the measurements of visual confusions are shown in figure 2.6. On the latter diagram, the vertical axis indicates which visemes are sooner confused, so the two visemes that are the hardest to tell apart are */th/* and */dh/*. The 75% line indicates that on 75% of the presentations of the visual stimuli the consonants were confused with consonants from the same cluster (e.g. */g/* would be confused with */k/* or itself, but not with */w/*). Comparing these two diagrams gives insight into which consonants are likely to be mixed up with each other if one hears them, but can be easily distinguished if one sees the speaker.

### 2.2.3 What people look at when reading speech

In order to be able to simulate human speech reading on a machine, a choice has to be made in what features of the face will be used to extract information

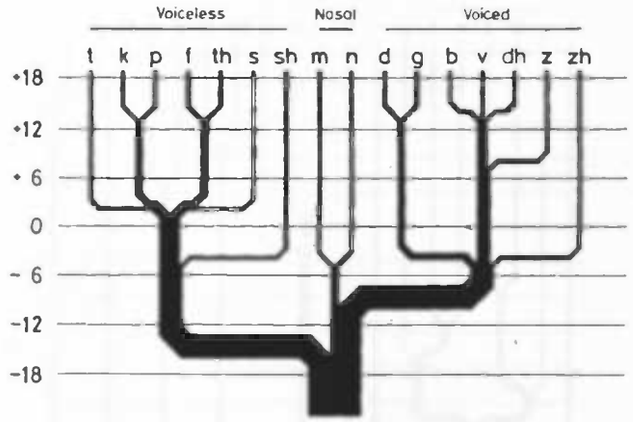



Figure 2.5: Auditory confusion among consonants, taken from [31]

from. We have seen that the place of articulation is important, specifically opened versus closed is highly visible. Rounded versus spread is highly visible as well, but front versus back is hardly visible at all [27]. Benoît et al. [3] investigated different models and their contribution to intelligibility of speech in different signal to noise ratios. These models were each controlled by (a subset of) the following six parameters: internal lip height and width, protrusion of the upper and lower lip, lip contact and jaw rotation. It was found that the most complete facial model gave the best intelligibility scores. This model consisted of an animated face controlled by all six parameters. The two other models were a jaw/skull model, controlled by the same parameters, but skin other than the lips was left out, and a model which consisted only of the lips and was controlled by all parameters except for the jaw rotation. For comparison, at a SNR of -18 dB, the intelligibility score of the facial model is around 40%, the skull model scores around 30% and the lips only model scores around 25%. The contribution to intelligibility of the image of a natural face (the eyes were covered), and just natural lips was investigated as well. Interesting to note is that the natural face gives by far the best intelligibility scores of all, for lower signal to noise ratios more than 1.5 times the score of the animated face model (at a signal to noise ratio of -18 dB, the audio + natural face scores around 65% percent intelligibility), but that the natural lips alone did not score that much better than the lip model. For exact scores see [3]. The conclusion is made that even though the lips give the most important cues for speech reading, the more information about the face is given, the better humans can speech read. For example, cues given by the teeth, chin, cheeks and tongue are important as well.

To conclude the section on human speech reading, three different situations where humans use visual information when understanding speech are discerned:

1. Judging whether a sound is coming from a speaker or if it is background noise, by looking at the speaker.

2. Checking whether a speaker is silent or speaking.

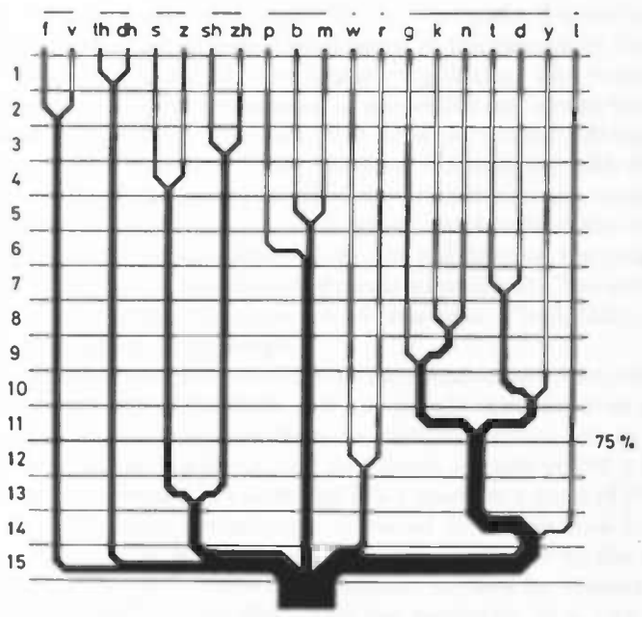



Figure 2.6: Visual confusion among consonants, taken from [31]

3. Making better estimates of what was uttered, by combining the sounds and the movement of a speakers face.

So when you are having a conversation in the crowded discotheque, or across a noisy street, you'll also use what you see, rather than only what you hear.

## 2.3 Audio-visual speech recognition

We have seen the core problems of automatic speech recognition, and we have taken a look at lipreading as a human way of solving the problems that are similar to the difficulties in ASR. Both are good reasons for taking a look at the possibilities of audio-visual speech recognition. The field of audio-visual speech recognition (AVSR) is a relatively new one, it started becoming popular in the 1980s. The focus of the field lies on the classification of possible visemes.

### 2.3.1 Audio-visual speech recognition at work

Although there is a wide variation of methods of automatic speech reading, a global description of the audio-visual speech recognition process will be given here. Since the process of auditory speech recognition is outlined in section 2.1.1, the focus here lies on automatic lipreading and the integration of visual and acoustical information.

The process can be divided in three stages. In the first stage the movements of the speakers face are captured on video. Dependant on the requirements of

the system different types of cameras can be employed. Aspects to consider are the frame rate of the camera, the distance of the camera to the face, the amount of pixels in a frame and the lighting conditions. The duration of the average phoneme is generally estimated at about 100 ms, so the frame rate of the camera should be at the very least 20 frames per second[9]. A frame rate of 25 Hz is common for a normal 'cheap' camera, but material with frame rates of 50 Hz or 60 Hz is used as well [12]. The distance of the camera to the face is important to consider, because the closer the face is to the camera, the greater effect outward pronunciation of the lips will have on the image, due to perspective. Furthermore, the amount of pixels is important. The more pixels in an image, the more precise distances can be measured. The lighting conditions greatly affect the colour of an image.

An example of visual data used for audio-visual speech recognition is the IBM ViaVoice audio-visual database. It is a large collection of movies of different speakers. Movies in this database have an interlaced frame rate of 30 Hz, so 60 frames per second are available, and each frame consists of $704 \times 480$ pixels. The movies are compressed in via the MPEG-4 mode at a ratio of 50:1 [22].

In the **second stage** information is extracted from individual frames. Dependant on whether the face is at a fixed position relative to the camera or not, first the face has to be located. A common method for speaker detection makes use of the hue[10] of the skin. This has proven to be a very successful cue, since the hue of the facial skin is very similar across people, even across different races [12]. Furthermore, cues such as edges and the intensity of pixels relative to their environment are used to locate facial features such as the eyes, nostrils or mouth. Once the face is located the relevant features of the face have to be extracted. Choosing which features should be extracted is a very important decision to make, since it defines the different states of the face or lips which the system can discern. No real consensus exist about which features exactly describe the linguistic states of the face correctly. Low and high level approaches can be discerned here [12, 22]. In the low level approach the region of the image containing the lips is considered as a whole as a region of interest (ROI). Usually, machine learning, like training neural networks [12] and Principal Component Analysis (a statistical method for extracting non correlated components) [22], are applied when categorizing pixel based regions of interest.

The second approach uses a priori knowledge about the face and lips to measure higher level features. The lower level knowledge about the region of interest can be used to locate features, as well as cues such as colour [33] of the lips or shadows from the oral cavity [11]. Sometimes, artificial cues such a blue lipstick or other markers on the face are used to facilitate accurate detection [1]. Examples of higher level lip detection are estimating the (inner or outer) width and height of the lips, the size of the area between the lips, the size of this area plus the lips, or the change, or the change in change of these parameters [1, 11, 22]. A more sophisticated approach is the matching of an inner and outer contour to the lip region to possible lip shapes [16, 22, 26]. Sometimes 3D models of the lips are used [12]. Petajan was the first to measure radial vectors representing the distance of the lips from the center of the mouth at different angles [22, 29, 33]. Another common method to obtain the outer contour of the

[9] In order to be able to measure frequencies of 10 Hz, a sampling rate of 20 Hz is required.
[10] For an explanation on hue, saturation and brightness see section 3.2, footnote 2.

mouth uses snakes, a method for finding the contour which fits a region the best [5]. The information thus extracted (be it as radial vectors, snakes, etcetera) is used for the following stage.

In the **third stage** estimates of possible utterances are made. Detecting if the speaker is silent or not can happen at this stage as well, if silence is considered a special type of utterance. Just as in auditory speech recognition, the predicting of the most likely utterance is done with Hidden Markov Models. Before using the visual features to estimate possible utterances, a decision has to be made about when to combine the auditory and visual information. When and how to integrate the auditory and visual stream is still very much a subject of research.

With *early integration* the auditory and visual feature vector are concatenated to form one feature vector, which is then processed by the HMM [1, 10]. *Late integration* can be considered a special case of early integration [12] and should improve an early integration system. In this case the auditory and the visual stream are categorized before the information is fused. The two separate categorizers both produce a list of probabilities for all utterances in the corpus. Combining these lists renders the best estimate. A simple way of combining these list is by taking the cross product of the probabilities of both channels of all utterances and selecting the highest candidate. More sophisticated methods of selection provide better results, however. By estimating the degree of certainty of each output channel, a choice can be made to what channel should have (more) influence on selecting the best candidate [1, 10]. Ideally, both channels should be recoded into articulatory features and then combined into an articulatory categorizer, but there is not enough knowledge about articulatory dynamics to do so, according to [10]. Although late integration should improve an early integration system, a vast amount of probabilities have to be calculated, which makes it practical only for a small set of possible utterances. For example, late integration is used when categorizing phonemes, and a word or sentence recognizer is built on top of that [12]. Another argument against late integration is that evidence, such as the McGurk effect, suggests that for humans audio-visual integration happens prior to phonetic categorization [10, 31].

### 2.3.2 State of the art

Although many reports are made about the successful use of visual information to improve recognition [1, 12, 21, 22, 29], it is hard to give an estimate of where we are now in the field of AVSR. Central tests to measure progress, with such a widespread participation as exist in the field of auditory speech recognition, do not exist today. Hennecke et al. in 1996 [12] compare the field of audio-visual speech recognition with auditory speech recognition in the 1940s and 1950s. A lot is to be learned from auditory speech recognition however, so development could go a lot faster.

### 2.3.3 Problems

A central issue in the field is feature extraction: what features should be extracted, and how. Especially with regard to the similarities and differences between speakers, the choice of appropriate features can make or break an AVSR system designed for speaker independence. The movement of the head alters

the measured features. So when a the head is turned sideways, the lips will be observed differently. If a face is closer to the camera, the change in perspective could be confused with a change in protrusion of the lips. When the head is tilted, or rolled, the height and width of the mouth rotate as well [26]. Furthermore, the fusion of visual and auditory data is subject to research. Next to the question of when the two data streams should be fused, the synchronization can be a problem, since the two streams are not always in sync, depending on the equipment used. Another type of problem is that AVSR lacks the tradition of tests such as the Hub 4 Broadcast News evaluation, and does not have a lot of accessible large labeled databases.

# Chapter 3

# HSB-based Automatic Lip detection

The majority of speech reading systems nowadays are developed to operate under conditions that are adapted to facilitate recognition. For example, blue lipstick is used, a very limited vocabulary is used or the subjects are asked to articulate clearly. However, in order for an automatic lipreading module to improve an auditory recognizer that operates in everyday circumstances, this module will have to cope with 'natural' circumstances as well. This was the motivation to explore the problems that occur when trying to process data that is obtained in a 'real-life' situation. A system for automatic feature extraction was created, in order to gain first-hand insight into automatic lipreading. At a low level the smallest rectangle containing the lips is extracted, and with the use of this region higher level features are extracted as well. The requirements are formulated as followed:

## 3.1 Requirements

**Simplicity.** The system has to be able to run on a normal home computer and produce results in real time (or at least as fast as a normal ASR system), therefore the method should not be too complex. A normal, relatively cheap digital camera is used with a frame rate of 25 frames per second. Each frame was transformed to an image of 200 by 150 pixels, with 24 bit RGB colour depth. Every pixel can be described by a value between 0 and 255 for the red, green and blue component. Furthermore, every frame was compressed in JPEG format, since compression is a very widespread technique when storing large amounts of data.

**Robustness.** The conditions for the speakers were as 'natural' as possible. The test subjects were asked to articulate normally, and no parts of the face were accented with for example lipstick. However, the lighting was virtually the same in all recordings, the position of the face was always at the same distance to the camera and the mouth was always visible. Furthermore, all speakers were white.

Except for the latter, these are conditions that can be manipulated in 'real-life' situations. The illumination can be controlled by placing lights, and the placement of the camera relative to a microphone or a screen can direct the face of a speaker to a predictable location.

Benoît et al. found that of different models, the subjects could speech read the best from the model of the complete face [3]. This model is controlled by six parameters: internal lip height and width, protrusion of the upper and lower lip, lip contact and jaw rotation. For this simple lipreading application a subset of these features was chosen.

- **The region of interest.** The smallest rectangle containing the outer contour of the mouth is considered the region of interest. The height and width of the outer contour of the mouth follow directly form this rectangle. In the software, these features are referred to as `width` and `height`.
- **The inner contour of the mouth.** The opening of the mouth tells more about a viseme than the shape of the outer contour if the mouth, because this information gives a directer insight into whether a mouth was for example opened or closed. But, as we will see in the results, this feature was harder to measure than the outer contour. The width is considered too difficult too estimate, since the corners of the mouth are hard to detect (see section 3.2.4). Only the height is estimated. In the software this feature is referred to as `innerLipHeight`.

Note that 'closed' versus 'opened' and 'rounded' versus 'spread' can be described by these features, and as we have seen these features are very important [27]. To meet these requirements, a software package is developed in Java.

## 3.2 Method

### 3.2.1 Subjects

Eight subjects (four female and four male) were seated in a room, two meters away from the camera. All subjects were native Dutch speakers. They were asked to read aloud a sequence of English digits and to answer a question in Dutch. The latter will be referred to as natural speech. Typically the complete face of the speaker was filmed. However, the speakers moved their head when reading from a paper or answering a question, but the lips are always visible in the recordings. See appendix B for example frames of all eight subjects.

### 3.2.2 How to begin

The video material was transformed into sequences of images using Adobe Premiere. Figure D.1[1] shows original images that will serve as examples for the different steps of the feature extraction. Next, the redness of the lips is used for selecting the part of the image where the lips are. Several methods were tested for selecting the pixels of the proper colour. Parabolic filtering of a specific

[1] Figures with a preceding letter rather than a number can be found in the corresponding appendix.

hue, saturation and brightness[2] gave the best results. See appendix C for a description of the different filter methods that were compared.

### 3.2.3 A simple parabolic filter

The parabolic hue filter selects all pixels with a hue that fall in a certain interval. Te filter is defined as follows [33]:

$$F_{hue}(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & \text{when } |h - h_0| \leq w \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

Here $h_0$ is the base hue value which is filtered out of the image, and $w$ is a width threshold in hue space. Thus, for pixels with a hue that is equal to $h_0$, $F_{hue}$ will be 1[3]. $F_{hue}$ defines a parabola with its peak at $h_0$, it intersects the $h$-axis at $h_0 \pm w$. The half width $w$ defines the strictness of the filter. A similar filter is applied for the saturation and brightness, and the resulting value is calculated by multiplying the outcomes ($F_{hue} \times F_{saturation} \times F_{brightness}$).

Even though all faces are different and not all mouths have the same colour, it was attempted to find a single configuration for the parabolic HSB filter, that suits all eight subjects. As with the skin colour of different people [12], the hue of the lips is the most similar across the different subjects, and the brightness varies the most. This results in a small half width for the hue and a large half width for the brightness. The half width for saturation is in between. The best overall results were given by the settings in table 3.1.

| **Type** | **Center value** (0...255) | **Half width** (0...255) |
|---|---|---|
| Hue | 0 | 11 |
| Saturation | 127 | 30 |
| Brightness | 128 | 128 |

Table 3.1: Settings for the parabolic filter

This hue is characteristic for the red colour of the lips. Shadows within the lip region cause variance in brightness, but the brightness also varies across different subjects. For an example of the results of this filter see figure D.2. With these settings the lip area was generally highlighted the most. In order to select the highlighted area, a binary threshold was set so that the 1.5% that was highlighted the most was made white, and the rest black (see figure D.3). The size of this threshold depends on the size of the lip area relative to the image.

[2]Hue, saturation and brightness are three measures for describing a colour. A colour which is described by a red, green and blue component can adequately be described by hue, saturation and brightness, and vice versa. Hue defines whether a colour is red, orange or yellow, etcetera. It is a gliding, circular scale. By convention, a hue of 0 corresponds to red. Saturation defines whether the colour is gray or vivid. The lower the value, the grayer a colour is. Brightness, also referred to as intensity, defines whether the colour is dark or light. The lower the brightness the darker the colour.

[3]Note that a problem can arise when calculating $h - h_0$. For example if $h_0 = 0.99$, a very slightly blueish red, and $h = 0.0$, simply red, the actual colour distance is 0.01, but calculating $|h - h_0|$ will give 0.99. Calculating the actual colour distance is achieved by checking if $|h - h_0| \leq 0.5$. If not, subtract $|h - h_0|$ from 1, and we get the distance we want.

If the subject would be further away from the camera, this area and thus this threshold should be smaller, and vice versa.

### 3.2.4 Locating the region of interest

At a lower level the region of interest (ROI) is extracted, and at a higher level the width and height of the outer contour of the mouth, and the height of the inner contour of the mouth. The ROI is the smallest rectangle containing (the outer contour of) the lips. Note that the orientation of the head is not taken into account here, so if a subject turns his or her head this can have a negative effect on the results.

First, a density histogram is calculated by counting the white pixels in each column (figure D.4). Second, the histogram is smoothed to reduce the influence of noise (figure D.5). All pictures where the subject was facing the camera have a histogram with a blob in the center. The length of this blob is equal to the width of the mouth. Furthermore, this blob contains the majority of the white pixels, so the median of the histogram is inside the blob. The width is measured by searching outward from the median to both sides for the first value of the histogram which is lower than a certain threshold. This threshold is the minimum amount of pixels (namely two) that is expected to be white in the mouth region. The threshold was optimized so that the width of the mouth was estimated rather somewhat too small than too wide. The corners of the mouth are sometimes left out. This occurs in the top left image in figure D.6.

The same method is applied when determining the height of the mouth, except that the density histogram for the rows is calculated from the pixels only within the estimated vertical region of interest. See figures D.7, D.8 and D.9. The threshold used for selecting the horizontal ROI is lower than for the vertical ROI, namely 0.25 pixels. The threshold was set so low because an opened mouth creates a histogram, with a valley in the middle (see for example the lower right image in figure D.8). This method is not perfect however. We will see that sometimes the upper lip falls outside the region of interest (figure E.3) and sometimes a part of the nose is selected (figure E.4).

### 3.2.5 Measuring higher level features

The width and height of the outer contour follow directly from the ROI. The last feature to be estimated is the height of the opening of the mouth. This estimate is acquired from the center ten columns of the region of interest (see figure D.10). These columns are chosen because a smaller set of columns would too easily be influenced by noise, and a wider set of columns would be influenced too much by the curvature of the mouth. In order to get all red regions in the center strip, also the ones that may have been left out due to threshold filtering as described in section 3.2.3[4], this strip is filtered again. The region is taken from the original image and HSB filtered with the same settings, and a percentage threshold of 50% is applied. Now all rows are selected with one or more white pixels, and all series of selected rows are counted. The distances between the regions are measured. The following six situations can occur:

[4]This is likely to occur when regions other than the mouth are highlighted as well.

1. **A closed mouth.** If only one region is seen, the mouth is regarded as closed and the height of the opening is estimated at zero. In all other situations the mouth is considered opened.

2. **An opened mouth.** When two regions are seen, these will be regarded as the upper and lower lip, and estimate that the distance between these regions is the height of the opening of the mouth.

3. **The lower lip is interrupted by reflective light.** In this case three regions are counted, the distance between the lower two regions is considered thus close that they belong to the same lower lip.

4. **The tongue is seen.** Three regions are detected, and the distance between the lower two regions is too big in order for them to be a lip that was interrupted by reflective light. The perfect threshold for distinguishing whether a tongue is seen or the lower lip was interrupted is hard to find, and still needs some fine tuning. For examples, see figures D.11 and E.9.

5. **The upper lip is outside the region of interest.** There is no white region in the upper third of the middle strip. The system will recognize that no upper lip was found (see the top left image in figure D.9). With this knowledge a more advanced system can look for an upper lip just above the region of interest. In the final implementation this has been left out, with keeping the program simple in mind.

6. **The image is unclear.** Zero or more than three regions are counted. The picture is too unclear to say anything about it. This has not occurred with the test material.

See figure D.11 for some examples of the different categories. The inner lip height can be estimated in the first four cases as the distance between the regions corresponding to the lower and upper lip. Next to the `width`, `height` and `innerLipHeight`, the parameters `state`, `comment` and `fileName` are generated as output. The parameter `state` describes whether the mouth is opened or closed, and the parameter `comment` reports which of these six categories occurred.

### 3.2.6 Test procedure

The system was tested on different sequences of 128 frames. The computer generated estimates for outer width and height and inner height are compared to manually tagged values. The tagging of these values was done by drawing a rectangle around the lips, and drawing a line between the upper and lower lip, both with the mouse. It should be noted that the tagging of the data was done only once, and the standard error for one measurement is estimated at 2 pixels, so the standard error for a measured distance is close to 2.8 pixels[5] This uncertainty can be attributed to two causes: 1.) JPEG compression leads to a loss in detail, and 2.) the images were tagged at a relatively big distance from

[5]Since the measurement of a distance involves the measurement of two points with each a standard error $\sigma_x = 2$, the standard error for the distance is $\sigma_f \approx 2.8$, according to $\sigma_f^2 = \sigma_x^2 + \sigma_y^2$, for $f = x + y$ [4].

the screen, resulting in a very small viewing angle. Because the tagging of these values is a vast task, not all material was used for testing. Of five subjects two sequences were tested, a sample of both the digits read aloud and the natural speech. Of a sixth subject only a sequence of natural speech was measured.

# Chapter 4

# Evaluation

## 4.1 Method of evaluation

From human speech perception and auditory speech perception we have seen that visual information might be useful in speaker detection, judging whether a speaker is speaking at all and in distinguishing (similar) phonemes (see section 2.2.3). Speaker detection has not been implemented, the system will always assume a speaker is there. The aim for this feature extraction module is to estimate values for a set of features that can maximally discern the different visemes. With the use of these features, the system should be able to distinguish similar phonemes. The ultimate goal would be to discern visemes to a degree where each viseme can be mapped to a single phoneme, but if the system can distinguish between groups of visemes it is still useful. For example, when an ASR system is expecting to observe for example either a */k/* or a */p/*, a system that can distinguish bilabial visemes from velar visemes can give a definite prediction[1]. The conventional way to test the usability of this system would be to use a module that classifies the estimated features in visemes or groups of visemes. This is the logical step to take but nevertheless lies beyond the scope of this research. Therefore, the usability of this system can not be tested in the conventional way, but has to be estimated. This is done with the help of graphs with traces of manually tagged values and computer-generated estimates. Furthermore, the variance of the features and the correlation between the traces are a measure of the 'goodness' of the results. A few plots are made of the width and height of the mouth when pronouncing different visemes. These plots are used to see whether the classification from the International Phonetic Alphabet (see appendix A) applies to the estimated values, and whether this classification can thus be used to categorize different visemes. Finally, the possibility for judging whether a speaker is silent or not will be considered. But first we will take a look of some of the common errors that occur when estimating the features.

### 4.1.1 Common errors

The method for selecting the width of the mouth is weak in situations where other regions in the face are red as well. The areas of the lips (often the corners

[1] The graph in figure 2.6 gives an indication of what viseme groups are likely to emerge.

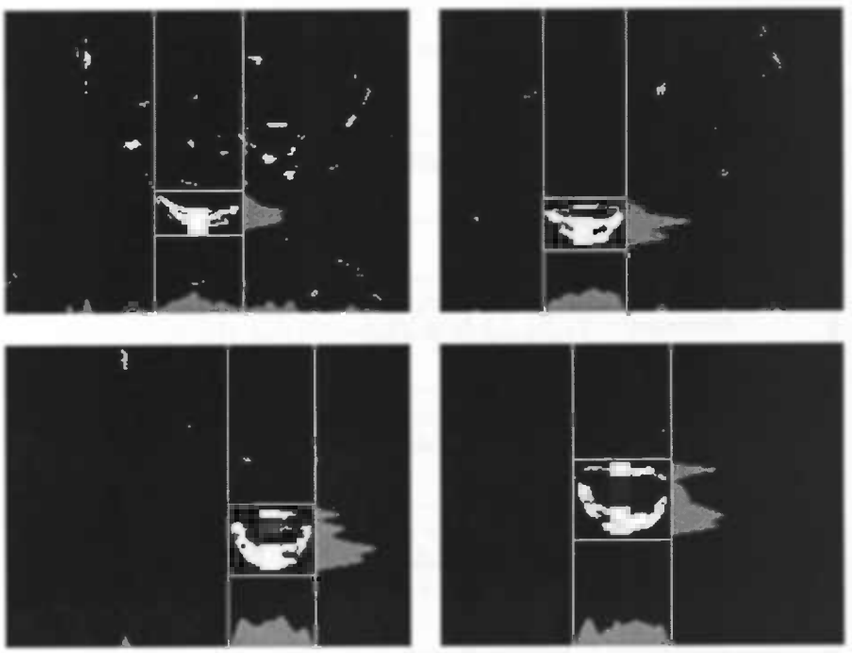


Figure 4.1: Feature extraction performed by the system. The region of interest is selected, and the width and height of the outer contour of the lips follow directly from it. Furthermore, the height of the inner contour of the lips is estimated. (These are the same images as in figure D.11.)

of the mouth) that deviate more from the ideal hue, saturation and brightness fall outside the region of interest. This occurs in the top left image of figure 4.1. Setting the threshold for selecting the width lower is not the answer since it will result in a vertical region interest that is wider than just the mouth. Other settings for the hue, saturation and brightness that are used for selecting the lips will select more regions outside the lips, or less of the lips.

Estimating the height of the mouth is also sensitive to the situation where the corners of the mouth are not selected. When the corners of the mouth are not highlighted, and the mouth is opened, the histogram with values from the rows in the vertical region of interest will have a valley between the two bulks of pixels of the upper and lower lip. In the top left image in figure 4.1 the upper lip is not included in the region of interest, but in the bottom right image the gap is thus small that the upper lip is detected. The density histogram is smoothed and a low threshold is used, but if the histogram is made smoother, or the threshold is set lower, this will often result in a ROI that includes more than just the lips.

The height of the inner contour of the mouth is difficult to measure when in the middle columns of the region of interest instead of two red areas (for the upper and lower lip), three areas are detected. This generally occurs either when the redness of the lower lip is interrupted by reflective light, or when the

tongue is visible. In the top right and the bottom left image both scenarios occur and are correctly recognized, but as we will see further on, this often goes wrong.

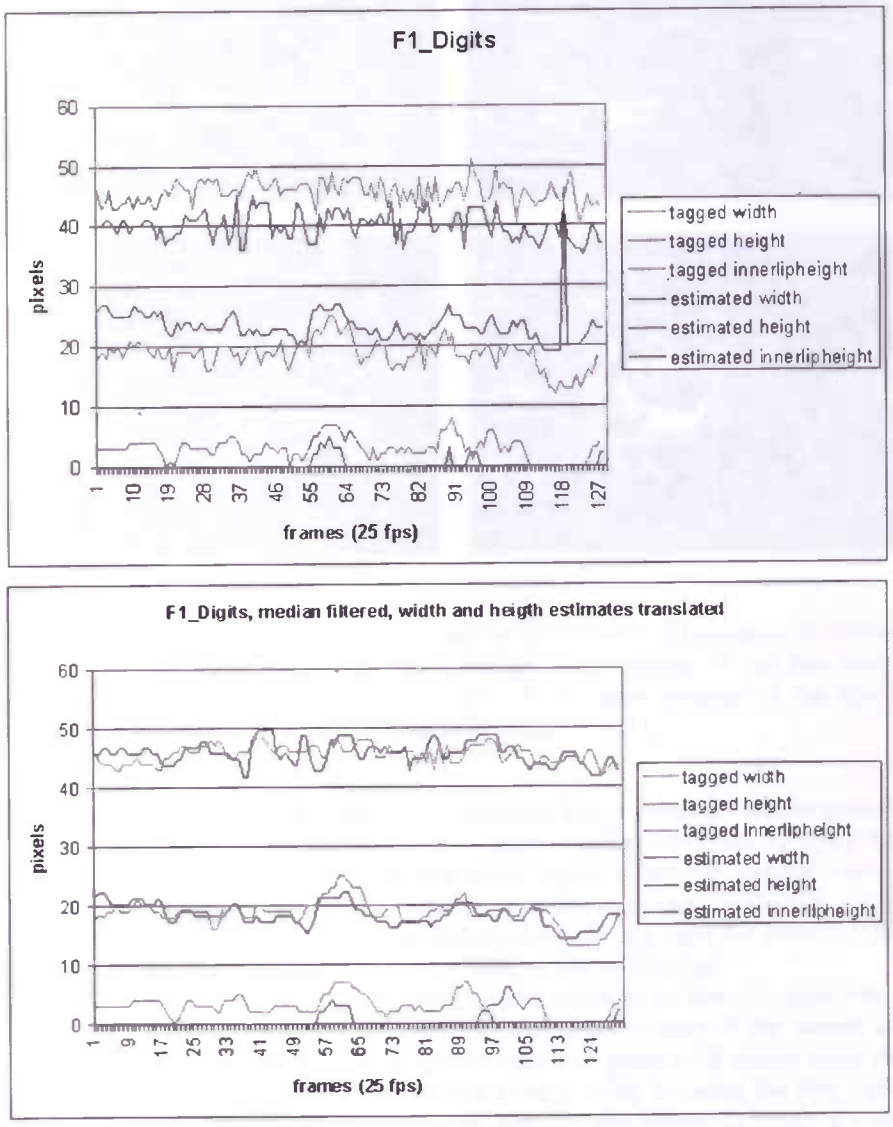



Figure 4.2: Results for F1 Digits, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. See section 4.1.2 for details.

### 4.1.2 Quantitative results

Figure 4.2 shows series of the three estimated features next to the manually tagged features, of subject F1 pronouncing digits. In the graphs, the gray lines represent manually tagged values, and the black lines are computer-generated estimates. The top two traces are the `width`, the middle two are the `height` and the bottom two are the `innerLipHeight`. In order to reduce the influence of the standard error on the tagged traces, and the influence of noisy estimates on the estimated traces, a median filter was applied with a window size of three pixels. This small window adequately removes singular errors, but preserves the rest of the signal. In the bottom image the peak in the height near frame 120 is removed. Typically, the width was estimated narrower, due to the strictness of the system (as described in section 3.2.4). The height, on the other hand, is estimated higher than the tagged traces. This can be attributed to the smoothing of the horizontal histogram, and the low threshold for selecting the horizontal region of interest (described in section 3.2.4 as well). In order to remove this systematic deviation, the mean difference between the tagged and estimated traces is calculated, and added to the estimates. The figure of the traces includes both the original values (top image) and the median filtered and translated values (bottom image). Appendix E contains all traces of the test material. Note that sometimes gaps appear in the trace of the inner lip height. This happens when the system is unable to detect an inner lip height because for example no upper lip height is found, as described in section 3.2.5.

The `innerLipHeight` in figure 4.2 hardly seems to match the manually tagged trace. In this case the subject articulated very modestly (compare for instance the `innerLipHeight` from E.10). With JPEG compression, pixels can be influenced by neighboring pixels and thus the small opening of the mouth is often made red as well. Furthermore, the smoothing of the horizontal density histogram can cause small gaps to disappear. The two `width` traces stay relatively close to each other, but the deviations of the tagged trace do not fall together with the deviations in the estimated trace. This is better with the height. For example there is a peak around frame 60 and a valley around frame 120 in both signals. A measure for the co-occurrence of these 'peaks' and 'valleys', and thus a measure for the goodness of the system, is the correlation coëfficient[2]. All correlation coëfficients are given in table 4.1.

As we compare the different values in this table, what stands out is that the results vary greatly from subject to subject. Compare for instance the correlations from F2 Natural, with the correlations from M4 Natural (the cor-

[2]The correlation coëfficient is calculated as:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.1}$$

where:

$$-1 \leq \rho_{X,Y} \leq 1 \tag{4.2}$$

$\sigma_X$ and $\sigma_Y$ are the standard deviations, the square root of the variance as calculated in footnote 3 and the covariance $Cov(X,Y)$ is calculated as:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) \tag{4.3}$$

where $X$ and $Y$ are the series of estimates and tagged values, $x_i$ And $y_i$ represent a single estimate or tagged value and $n$ is the length of the series.

| | correlation | | |
|---|---|---|---|
| subject | outer width | outer height | inner height |
| F1 (Digits) | 0.39 | 0.66 | 0.40 |
| F1 (Natural) | 0.052 | 0.21 | 0.12 |
| F2 (Digits) | 0.13 | 0.29 | 0.25 |
| F2 (Natural) | -0.20 | 0.30 | -0.079 |
| F4 (Digits) | 0.49 | 0.83 | 0.79 |
| F4 (Natural) | -0.064 | 0.46 | 0.33 |
| M1 (Digits) | -0.051 | 0.81 | 0.87 |
| M1 (Natural) | 0.32 | 0.51 | 0.71 |
| M2 (Digits) | 0.51 | 0.85 | 0.86 |
| M2 (Natural) | 0.75 | 0.82 | 0.83 |
| M4 (Natural) | 0.59 | 0.85 | 0.91 |

Table 4.1: The correlation of the different features, between the tagged and estimated traces.

responding traces can be found in figure E.4 and figure E.11). The correlation coëfficient is a measure for the reliability of the estimate. If the system would be used as an auxiliary module for an auditory speech recognizer, this coëfficient indicates the reliability of the visual stream. Note however that this coëfficient can only be calculated when all data are manually tagged as well.

Aside from the correlation, the manner of articulation is important when distinguishing visemes. When a subject is articulating clearly, the different positions of the mouth are easier to distinguish and vice versa. A measure for the articulateness of pronunciation is the variance[3] of the features. When a subject articulates more clearly, there will be a greater variance in the tagged values, because the height and width of the lips will change more. Table 4.2 contains the variance of the tagged and estimated features. Again the values vary between speakers. Note for example the difference between F1 Digits and M2 Digits (the corresponding traces can be found in figure E.4 and figure E.11). Together, the correlation and variance define the discriminatory power of the system. If the correlation is high, the estimates are good, and if the variance is large as well, it is easy to discern different states. When the variance is low, smaller deviations play a greater role. Take for example the traces of the `width` in the figures 4.2, E.2, and E.7. Here a small variation has a greater impact on the correlation than for example in the figures E.9, E.10 and E.11.

We can judge the manner of articulation also by comparing the measure of spread versus the roundedness of different phonemes. Appendix F consists of four figures where for different phonemes the inner lip height and width are plotted. Note that F1 articulated less clearly than M2, while M1 and M4 are somewhere in between. Compare the corresponding variances in table 4.2. The

[3]The variance is calculated as:

$$\sigma_X^2 = \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} \tag{4.4}$$

| | variance: tagged values (pixel$^2$) | | | variance: estimates (pixel$^2$) | | |
|---|---|---|---|---|---|---|
| subject | outer width | outer height | inner height | outer width | outer height | inner height |
| F1 (Digits) | 2.3 | 5.5 | 3.1 | 3.8 | 3.6 | 0.63 |
| F1 (Natural) | 6.8 | 4.8 | 6.6 | 7.8 | 11 | 12 |
| F2 (Digits) | 3.5 | 3.3 | 7.9 | 6.2 | 20 | 2.2 |
| F2 (Natural) | 3.6 | 7.0 | 5.5 | 46 | 30 | 9.2 |
| F4 (Digits) | 11 | 7.7 | 13 | 7.4 | 8.0 | 11 |
| F4 (Natural) | 7.3 | 10 | 22 | 29 | 13 | 20 |
| M1 (Digits) | 5.1 | 4.0 | 12 | 6.6 | 5.5 | 5.7 |
| M1 (Natural) | 16 | 6.3 | 13 | 4.3 | 17 | 14 |
| M2 (Digits) | 11 | 39 | 36 | 7.8 | 23 | 20 |
| M2 (Natural) | 19 | 20 | 21 | 14 | 15 | 31 |
| M4 (Natural) | 4.7 | 13 | 24 | 7.9 | 15 | 20 |

Table 4.2: The variance of the different features, tagged and estimated traces. The greater the variance of a feature, the more the features change. Note that a great variance of an estimated value does not necessarily mean the subject articulates clearly, since a highly erroneous trace with large deviations also causes great variance.

corresponding traces can be found in figures E.2, E.10, E.8 and E.11, respectively. According to the International Phonetic Alphabet in appendix A, the categorization in table 4.3 can be made. Appendix F consists of figures where the median-filtered and translated estimates of the inner height and outer width are compared for different visemes. If we regard these figures, we would expect to find the phonemes distributed in a way like in figure 4.3 (see also [20]).

| vowel | rounded | spread | consonant | | place |
|---|---|---|---|---|---|
| open | | /a:/ | opened | /t/, /d/ | alveolar |
| open-mid | /O/ | /E/ | ↕ | /f/ | labiodental |
| close-mid | /o:/ | /e:/ | closed | /p/, /b/, /m/ | bilabial |
| close | /u/ | /i/ | | | |

Table 4.3: Some phonemes and their visual aspects

It should be noted that in the traces of 128 frames identical phonemes do not occur often. For a more extensive analysis more data have to be used. However, when investigating these examples, a few cases stand out. In figure F.1 the */e:/* is located at a inner lip height of zero, which can surely not be the case. This suggests faulty recognition. Furthermore, in figure F.4, the */m/* and */p/* seem to be pronounced with an opened mouth! When taking a look at the corresponding frames, it seems that the complete closure of the mouth happens too fast for the camera to record. In this figure there are also two distant areas that correspond to the */a:/*. The area with the smaller width (between 49 and 50) appears even more rounded than the */o:/*. If we take a look at figure F.3, the

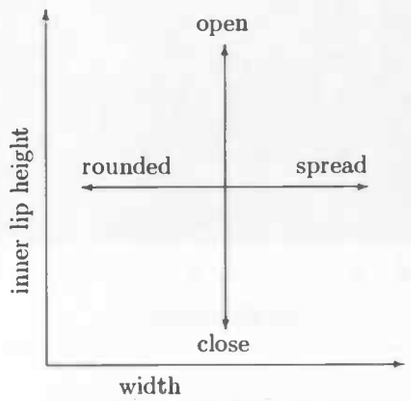



Figure 4.3: Visual distribution of phonemes

height of the /i/ is a much greater than the height for the /O/ and /o:/, while one would expect the opposite. The height of the /O/ should be greater than that of the /o:/ as well. In this case, the /o:/ was pronounced immediately before an /m/, a closed phoneme, which influences the articulation of the /o:/. Apart from these cases, the distribution as suggested by table 4.3 and figure 4.3 holds for the other plotted phonemes. Take, for example figure F.2. For rounded vowels (/O/ and /u/), the width should be smaller than for spread vowels (/E/ and /a:/). The height should be distributed such that the height for /a:/ is bigger than the height for /E/ and /O/, and the height for these vowels should be higher than the height for /u/. To conclude, /p/, /b/ and /m/ should have an inner lip height of zero, and thus smaller than the inner lip height for /t/. All these conditions hold for figure F.2.

We have seen examples of an important aspect of continuous speech, namely that the surrounding visemes greatly influence the articulatory features of a phoneme. Furthermore, differences between speakers are great. As expected, the spread of values for inner lip height for the speaker that articulates the clearest (M2) is greater than for the speaker that articulates in a more subtle way (F1). It is not a surprise that variation exists between the pronunciation of different speakers. Even when speakers are asked to pronounce isolated phonemes articulately, the width and height of the outer contour are different per speaker [20].

### 4.1.3 Judging whether a speaker is silent or not

The system is able to detect when a speaker has a closed mouth for a longer period of time, say longer than a second. In this case the system can recognize

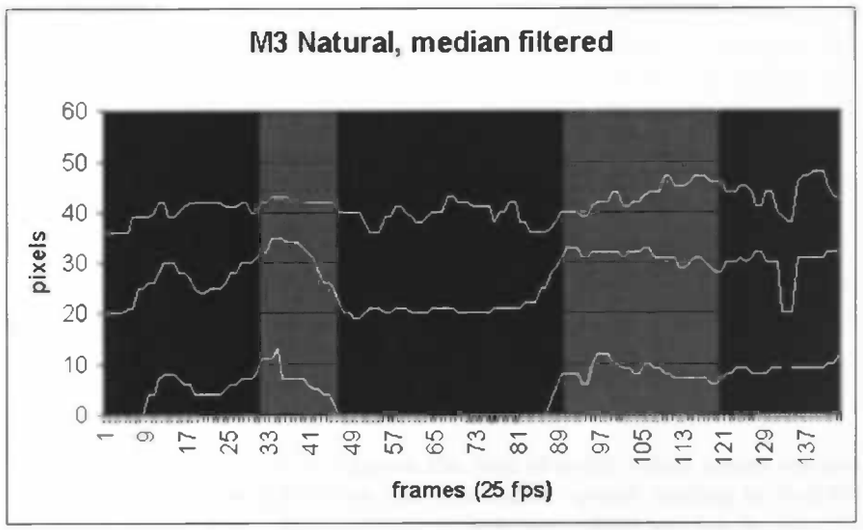



Figure 4.4: A trace of subject M3. The gray background indicates that the subject was speaking, and the black background is where the subject was silent. The upper trace is the width, the middle trace is the height and the lowest trace is the inner lip height. Only in the case where the inner lip height is zero for a longer period of time, it can safely be estimated that the speaker is silent.

that the speaker is silent. However, this is a special case of 'speaker silence'. Take figure 4.4, here a longer stretch of speech is processed, and the background is made gray where the speaker was actually speaking. We see that when a speaker is silent, it does not automatically mean his lips are not moving. He may, for example, be taking a deep breath. The system will not be able to distinguish between the first type of movement and the second. Only in the case of the prolonged closed mouth, the system will be able to detect silence.

# Chapter 5

# Conclusion

The aim of this project is to explore the field of audio-visual speech recognition, and to chart the difficulties that accompany speech reading in 'real-life' situations. It was found that auditory speech recognition has the largest difficulties with discerning the speech signal from background noise. Using visual information can help, specifically with discerning (similar) phonemes, knowing whether a speaker is silent, and detecting whether there is a speaker present at all. The common approach of audio-visual speech recognition is to extract features from visual material, that, combined with auditory features, are used to predict the most likely utterance. In order to gain insight in audio-visual speech recognition, a system was developed that performs feature extraction, and it was evaluated to see if it can utilize visual information in the three ways described above. Since this system should operate in 'real-life' situations, it was required that the program would be computationally simple. Furthermore, 'real-life' visual material was interpreted such that, although the subjects were always facing the camera at a fixed distance, the different subjects were asked to articulate naturally, they were allowed to move their head, and no parts of the face were accentuated with for example lipstick.

The system performs feature extraction at two levels. At a lower level, the system locates the region of interest. The method is simple but very effective in locating that part of the image which contains a certain percentage if the pixels that fit the base colour the best. Although this method is now only used to locate the mouth, it can easily be adapted to locate any region of a specific hue, saturation and brightness. At a higher level, the system estimates the height and width of the outer contour of the mouth, and the height of the inner contour of the mouth. The system performs varyingly across different subjects, and across the different features. The correlation between manually tagged values and computer estimates (table 4.1) is a measure for the reliability of the estimates. A correlation smaller than or equal to 0 corresponds to a reliability of 0%, and a correlation of 1 to a reliability of 100%.

Although the system does not render reliable results for all subjects, for the better cases (for example figures E.10 and E.11) it is interesting to consider two of the three methods for using visual information to improve recognition. First, the system can help with distinguishing between similar phonemes. Take for example the similar sounding */k/* and */p/*, if the `innerLipHeight` is 0, it can only be the bilabial */p/*. Bilabial visemes are the easiest to distinguish from

other visemes, since these are the only sounds pronounced with closed lips. To what degree exactly this system can distinguish visemes deserves further investigation. Second, the system can determine when a speaker is silent or not, but only in the case where the mouth is closed for a longer period of time. At this moment, detecting a speaker is not performed by the system. Since different human faces have a similar hue [12], the methods that have been implemented already can easily be adapted to locate a large area with pixels of a 'face-like' hue.

Regarding the processing of 'real-life' visual material, two major difficulties have been found. First, although for all speakers the lips have a similar hue, the faces differ a lot. One face may be blushing, where another is pale. The distribution of color across a face can vary greatly. Second, articulation between speakers varies (see also [20]), and the pronunciation of a single phoneme is greatly influenced by the surrounding visemes.

To summarize, a preliminary venture into speech reading has been implemented. The system performs good at locating a blob of a certain colour. This simple method may prove successful in a wide range of applications[1]. The method for estimating higher level features still leaves room for improvement, however. The extracted visual features will only make it possible for a few of the subjects to recognize visemes or determine whether a speaker is silent or not. In 'real-life' however, different faces, with different lipsticks and skin colours, and sloppy articulation make it seem impossible for a system like this to speech read accurately.

## 5.1 Recommendations

The system that has been created can serve as a platform for further exploration of the field of audio-visual speech recognition, or at a larger scale digital image processing. An interesting expansion to the system would be the face detection module. The extraction of higher level features can certainly be improved. Concerning the difficulties that occur with different speakers, two different approaches can be taken. On the one hand, the system could use different settings for different speakers. For example, the distribution of the ideal colour in the image could act as a hint to whether the system is dealing with a blushing or a pale face, and use different thresholds accordingly. On the other hand, the system now uses but little a priori knowledge about the mouth. If the method for locating it were to use for example information about the elliptical shape of the mouth and what widths and heights are likely, the fragile system of thresholds could be replaced by a better one.

Furthermore, the sloppy articulation of people in normal situations has proven to make the discerning of different visemes very difficult. A system like this probably will never be able to map a viseme to a single phoneme. The quality of the images is very poor, due to JPEG compression and the small size, so it is advisable to use better material. Furthermore, it may be worthwhile to reconsider the features that are estimated, take for example the width of the inner contour of the mouth. Otherwise, one could ask the subjects to articulate more clearly, but this would comprise the demand of 'real-life' material.

[1] Locating an orange ball on a playing field, for instance.

A logical step to take from here is to create a classification module, that discerns different groups of visemes, based on the features extracted by the current module. Not only will such a system be the necessary next step in automatic speech reading, it is also essential for further evaluation of the discriminatory power of the feature extraction module.

The problems that occur when processing 'real-life' material shed an interesting light on the field of audio-visual speech recognition today. Where the problems that occur with the differences between speakers may very well be overcome by more sophisticated methods, the sloppy articulation in continuous speech appears to make the classification of visemes to a level of individual phonemes impossible. If we want to see a man-made lipreading machine as soon as 2010, this is probably the hardest problem to overcome.

# Chapter 6

# Acknowledgments

First and foremost I would like to thank my counselor dr. Esther Wiersinga-Post for miraculously being able to take care of me, next to three children. I feel our cooperation has been ever pleasant.

I am grateful towards prof. dr. Lambert Schomaker, dr. Tjeerd Andringa and prof. dr. ir. Hendrikus Duifhuis for showing me some right direction when I felt lost. Drs. Judith Grob's diligent scrutinizing of my words also has not gone unnoticed. Speaking of people at the lab, I owe thanks to those such as Albert 'Balbert' van der Heide for creating a pleasant working environment.

I would like to thank my parents Francien and Diek Duifhuis for providing me with both nature and nurture. Life has been a most interesting field of research. Finally, the lovely Floor 't Sas has proven to be the most rewarding topic of research as of yet. Thank you,

I love you.

# Bibliography


[1] Ali Adjoudani & Christian Benoît: *On the Integration of Auditory and Visual Parameters in an HMM-based ASR*, in David G. Stork, Marcus E. Hennecke: *Speechreading by Humans and Machines: Models, Systems and Applications*, p 461-471, 1996

[2] Tjeerd Andringa: *Continuity preserving signal processing*, 2002

[3] Christien Benoît, Thierry Guiard-Marigny, Bertrand Le Goff, Ali Adjoudani: *Which components of the face do humans and machines best speechread?*, in David G. Stork, Marcus E. Hennecke: *Speechreading by Humans and Machines: Models, Systems and Applications*, p 315-328, 1996

[4] Herman Johan Christiaan Berendsen: *Goed meten met fouten*, 2000

[5] Greg I. Chiou, Jenq-Neng Hwang: *Lipreading by Using Snakes, Principal Component Analysis, and Hidden Markov Models to Recognize colour Motion Video*

[6] Ronald A. Cole, Yonghong Yan, Brian Mak, Mark Fanty, Troy Bailey: *The Contribution Of Consonants Versus Vowels To Word Recognition In Fluent Speech*, Proc. ICASSP '96, 1996

[7] Ellen Eide, Benoit Maison, Mark Gales, Ramesh Gopinath, Scott Chen, Peder Olsen, Dimitri Kanevsky, Miroslav Novak, Lidia Mangu: *IBM's 10X Real-Time Broadcast News Transcription System Used in the 1999 HUB4 Evaluation*, Proceedings of the 2000 Speech Transcription Workshop, NIST.

[8] Jonathan Fiscus, William M. Fisher, Alvin Martin, Mark Przybocki, David S. Pallett: *2000 NIST Evaluation of Conversational Speech Recognition over the Telephone*, Proceedings of the 2000 Speech Transcription Workshop, NIST.

[9] Eric Galyon: *C++ vs Java Performance*, http://www.cs.colostate.edu/-cs154/PerfComp/

[10] Laurent Girin, Jean-Luc Schwartz, Gang Feng: *Audio-visual enhancement of speech in noise*, Journal of the Acoustical Society of America, vol 109, nr 6, 2001

[11] Alan J. Goldschen, Oscar N. Garcia, Eric Petajan: *Continuous Optical Speech Recognition by Lipreading*, 1994

[12] Marcus E. Hennecke, David G. Stork, and K. Venkatesh Prasad: *Visionary Speech: Looking Ahead to Practical Speechreading Systems*, in David G. Stork, Marcus E. Hennecke: *Speechreading by Humans and Machines: Models, Systems and Applications*, p 331-349, 1996

[13] The IEEE History Center: *website on the history of Automatic Speech Synthesis & Recognition*, http://www.ieee.org/organizations/history_center/-sloan/ASSR/assr_index.html, 2001

[14] The International Phonetics Association, http://www.arts.gla.ac.uk/ipa/-ipa.html

[15] John MacDonald and Harry McGurk: *Visual influences on speech perception processes*, Perception and Psychophysics, 1978, Vol. 24 (3), 253-257

[16] Iain Matthews, Tim Cootes, Stephen Cox, Richard Harvey, and J. Andrew Bangham: *Lipreading using shape, shading and scale*, 1998

[17] Harry McGurk and John MacDonald: *Hearing lips and seeing voices*, Nature Vol. 264 December 23/30, 1976.

[18] Microsoft Research, Speech Technology website, http://research.microsoft.com/research/srg/default.aspx, Juli 2002

[19] George A. Miller and Patricia E. Nicely: *An Analysis of Perceptual Confusions Among Some English Consonants*, The Journal of the Acoustical Society of America, volume 27, nr 2, 1955

[20] Allen A. Montgomery, Pamela L. Jackson: *Physical characteristics of the lips underlying vowel lipreading performance*

[21] Ara V Nefian, Lu Hong Liang, Xiao Xing Liu, Xiaobo Pi: *Visual Interactivity: Audio-Visual Speech Recognition*, Intel Research, 2003, http://www.intel.com/research/mrl/research/avcsr.htm

[22] Chalapathy Neti, Gerasimos Potamianos, Jeurgen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, Azad Mashari, and Jie Zhou: *Audio-Visual Speech Recognition*, CLSP Workshop final report, 2002, Available online from: http://www.clsp.jhu.edu/ws2000/.

[23] Long Nguyen, Spyros Matsoukas, Jason Davenport, Jay Billa, Rich Schwartz, John Makhoul: *The 1999 BBN BYBLOS 10xRT Broadcast News Transcription System*, Proceedings of the 2000 Speech Transcription Workshop, NIST.

[24] The NIST Speech Group: *website dedicated to the advancement in the field of automatic speech recognition.* http://www.nist.gov/speech/index.htm, 2003

[25] Mosur Ravishankar, Rita Singh, Bhiksha Raj, Richard M. Stern: *The 1999 CMU 10X Real Time Broadcast News Transcription System*, Proceedings of the 2000 Speech Transcription Workshop, NIST.

[26] Lionel Revéret, Frederique Garcia, Christian Benoît, Eric Vatikiotis-Bateson: *An Hybrid Approach to Orientation-Free Liptracking*, Proc. of the First ESCA Workshop on Audio-Visual Speech Processing, AVSP'97, p. 117-120, Rhodes, Greece, Sept. 26-27, 1997

[27] Jordi Robert-Ribes, Jean-Luc Schwartz, Tahar Lallouache, and Pierre Escudier: *Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise*, Journal of the Acoustical Society of America, vol 103, nr 3, 1998

[28] UCL Phonetics and Linguistics: *SAMPA; computer readable phonetic alphabet*, http://www.phon.ucl.ac.uk/home/sampa/home.htm

[29] Lambert Schomaker et al.: *A Taxonomy of Multimodal Interaction in the Human Information Processing System*, Esprit Project 8579/MIAMI, http://hwr.nici.kun.nl/ miami/taxonomy/taxonomy.html, 1995

[30] W.H. Sumby and Irwin Pollack: *Visual Contribution to Speech Intelligibility to Speech in Noise*, The Journal of the Acoustical Society of America, 1954

[31] Quentin Summerfield: *Some preliminaries to a Comprehensive Account of Audio-Visual Speech Perception*, In B. Dodd and R. Campbell, editors, Hearing by Eye: The Psychology of LipReading, pages 97–113, Hillside, 1987. Lawrence Erlbaum Associates.

[32] Brian E. Walden et al.: *Effects of training on the visual recognition of consonants*, Journal of Speech and Hearing Research, 20, 1977

[33] Jacek C. Wojdel, Leon J. M. Rothkrantz: *Robust video processing for lipreading applications*, Proceedings of 6th annual scientific conference on web technology, new media, communications and telematics theory, methods, tools and applications (EUROMEDIA 2001) Valencia, Spain, p 195-199

[34] Su-Lin Wu: *Incorporating Information From Syllable-length Time Scales Into Automatic Speech Recognition*, 1998

[35] Steve Young: *Hidden Markov Models in Speech and Language Processing*, ELSNET Summer School, 1994

# Appendix A

# International Phonetic Alphabet

The chart in figure A.1 is taken from the International Phonetic Association [14]. The aspects *place* and *manner* of the consonants */b/*, */d/* and */g/*, as referred to in section 2.2.1, correspond to the descriptions in the upper row (for place) and the left-most column (for manner) of the chart for the pulmonic consonants. Throughout this thesis however, the computer-readable SAMPA alphabet is used [28].

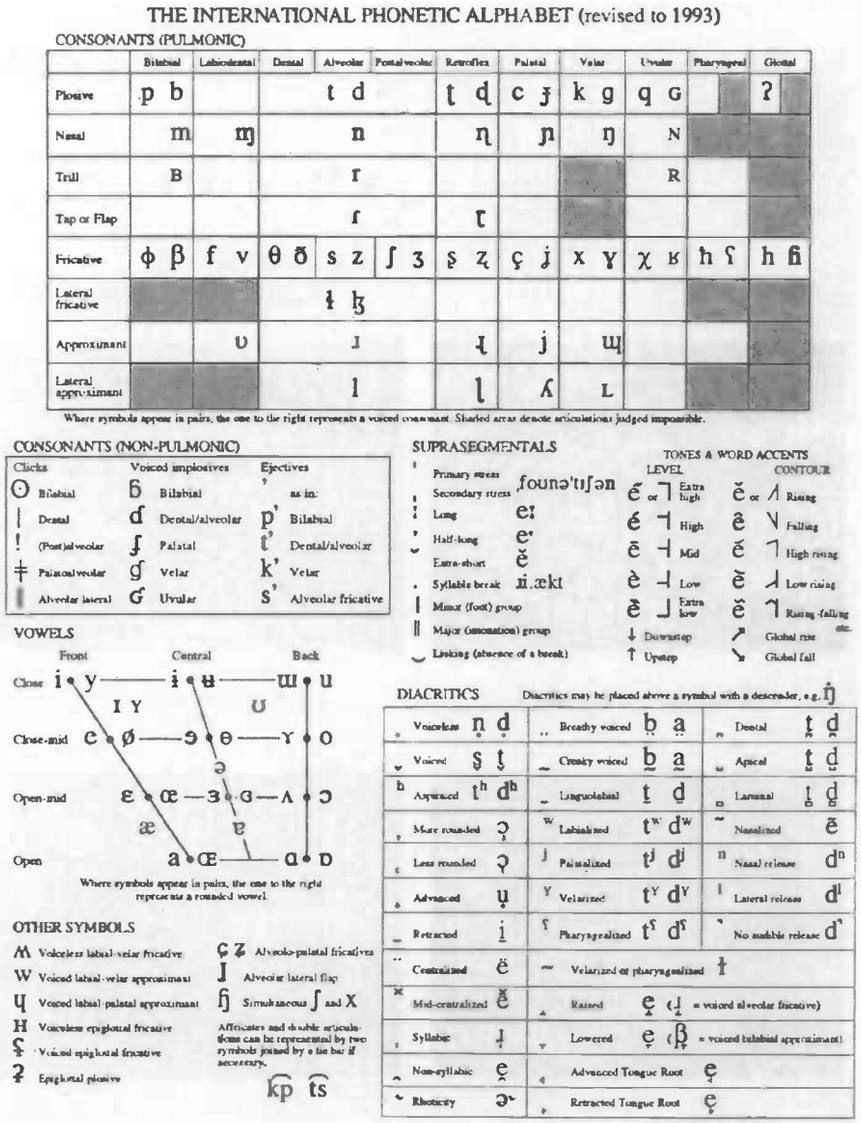



Figure A.1: Chart of the International Phonetic Alphabet

# Appendix B

# The subjects

Figure B.1: Subject F1

Figure B.2: Subject F2

Figure B.3: Subject F3

Figure B.4: Subject F4

Figure B.5: Subject M1

Figure B.6: Subject M2

Figure B.7: Subject M3

Figure B.8: Subject M4

# Appendix C

# Comparing filters

## C.1 Overview

Several filters are compared concerning their effectiveness in selecting the lip region for different subjects. An application was created where the different filters can be applied to images, and the settings for the filters can be tweaked. The following five filters were tested:

1. Parabolic, hue-based
2. Parabolic, hue and saturation-based
3. Parabolic, hue, saturation and brightness-based
4. Red/green threshold
5. Red/green colour burn

All methods make use of the redness of the lips, and select pixels based on their colour.

## C.2 Parabolic filtering with hue, saturation and brightness

The filter used for hue-based filtering is the parabolic filter as described in greater detail in section 3.2.3. The parabolic hue filter selects all pixels with a certain hue. Pixels that have the same hue as the base hue for the filter are made white, and the further the hue of pixels is away from the base hue, the darker gray the pixel is made. If the distance between the base hue of the filter and the actual hue of a pixel is greater than a certain 'half width', the pixel becomes black.

In order to get a filter that selects just the lips, it was attempted to create a filter that selects no other colours than those that occur in the lip region. In other words, it was attempted to find a hue and half width, where the half width is as small as possible.

The hue of a pixel tends to take on unpredictable values in darker areas [33], and elements of the image such as a bright red sweater can be the same hue,

but with a different saturation. Therefore, next to the hue, a similar filter is applied to the saturation and brightness.

Hue is the best discerning measure, then saturation, and in the last place brightness. The brightness is less discriminatory, since there is a big variance in lighter and darker areas in a mouth. Faces and lips still vary in colour, the filter selects one person's lips better than another's. With the settings in table 3.1 overall the lips are highlighted the most. A percentage threshold filter is applied so that for all frames the same amount of pixels is selected. Pictures taken in indoor office situations give good results, but problems occur with other red areas in the face, such as the nose, the cheeks and sometimes the ears.

## C.3 Red/green threshold

In this method, the red component of a pixel is divided by the green component, and a lower and upper threshold are applied to the resulting value [5]. When testing this method it was found that the upper threshold does not contribute to better results, since the lips are in the part of the picture with the highest red/green ratio. The lower threshold, on the other hand, greatly influences the results. Finding a single good threshold was a problem, since a value that works well for one picture, can be too high for the next. This might be solved with a percentage threshold as described above, but next to this problem, this filter selects pixels outside the lip region more often than the previous method. Therefore the Red/Green threshold method was not selected for further processing.

## C.4 Red/green colour burn

'Colour burn' is an operation over two colours, where the lightness of the blend colour defines the lightness of the base colour in the result. When the blend colour is dark, the result will be dark, when the blend colour is light, the result will be exactly the value of the base colour. The formula for colour burn is:

$$F_{colourburn}(c_{base}, c_{blend}) = 1 - \frac{1 - c_{base}}{c_{blend}} \tag{C.1}$$

Red is used as the base colour and inverse green as the blend colour. Inverting the green value will give a higher result when the green value is lower, and a lower result when the green value is high:

$$F_{colourburn}(c_{red}, (1 - c_{green})) = 1 - \frac{1 - c_{red}}{(1 - c_{green})} \tag{C.2}$$

Here $c_{red}$ is the red component of a pixel on a scale of 0 to 1, and $c_{green}$ the green component. After applying the colour burn filter, a percentage threshold is applied as well. The method is very sensitive to darker regions, and tends to select for example the nostrils or areas of the face that are shaded in another way.

To conclude, the parabolic HSB filter was chosen because it is the best at selecting just the pixels that belong to the lip area. The hue, saturation and

brightness scale seems the most insightful approach to describing a colour, so it is easy to adjust a filter if one has an idea about the colour that should be selected. What stands out as well is that in all images the hue of the lips is very similar, the filter for selecting the hue has a relatively small half width.

# Appendix D

# Hue-based Automatic Lip-detection: Images

Figure D.1: The original images of four of the subjects, gray scaled.

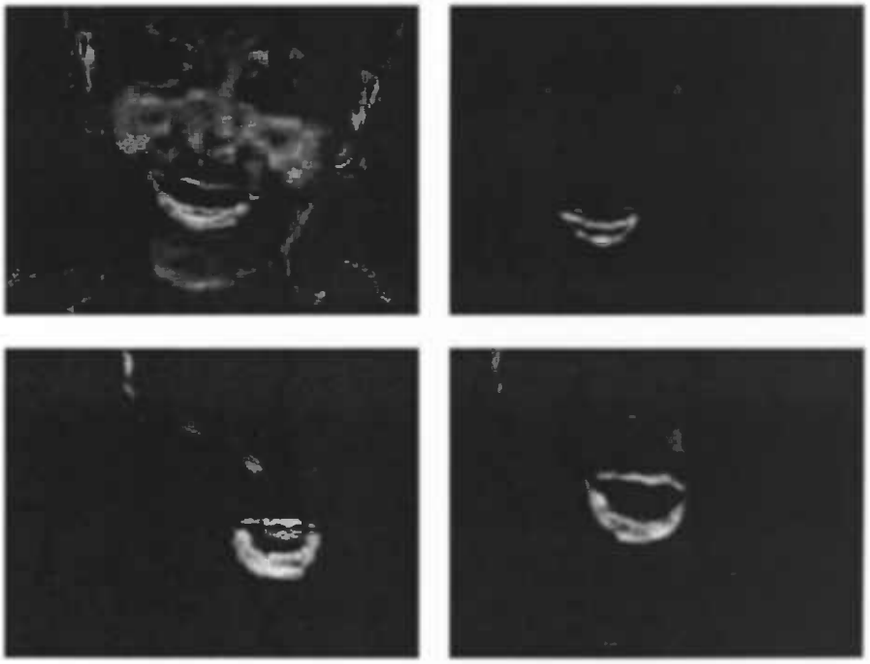


Figure D.2: The effect of applying the parabolic HSB filter. Note the wide variety in intensity of the selected pixels. The girl in the upper left image was blushing, so a big part of the face 'lights up'. Furthermore, the girl in the upper right image had a relatively pale face, compared to the others.

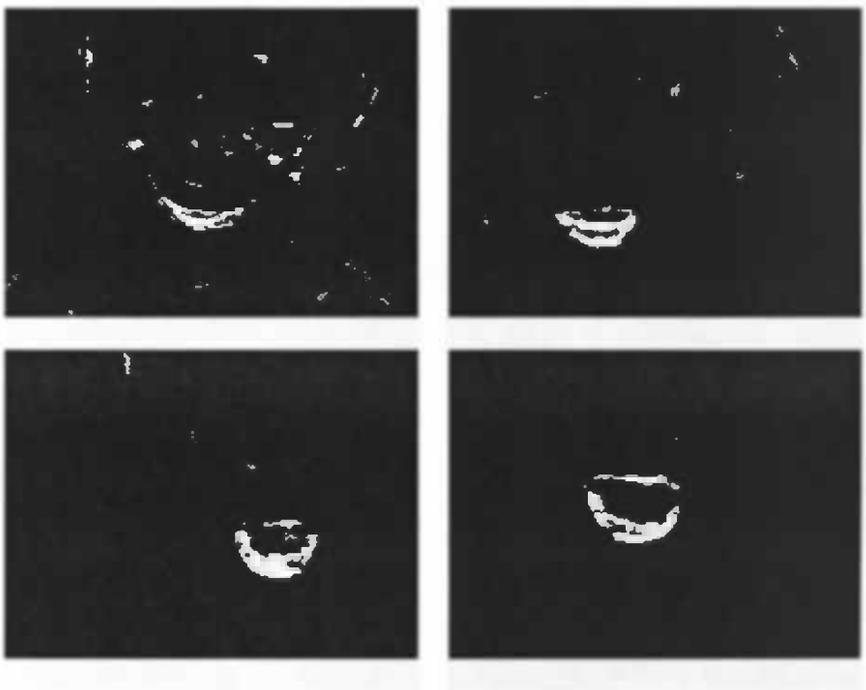


Figure D.3: Applying a percentage threshold on the filtered images. The differences between the intensities as in figure D.2 are normalized, but in the **upper left** image a lot of unwanted pixels are highlighted, just as some pixels (mainly in the corners of the mouth and the upper lip) are not highlighted.

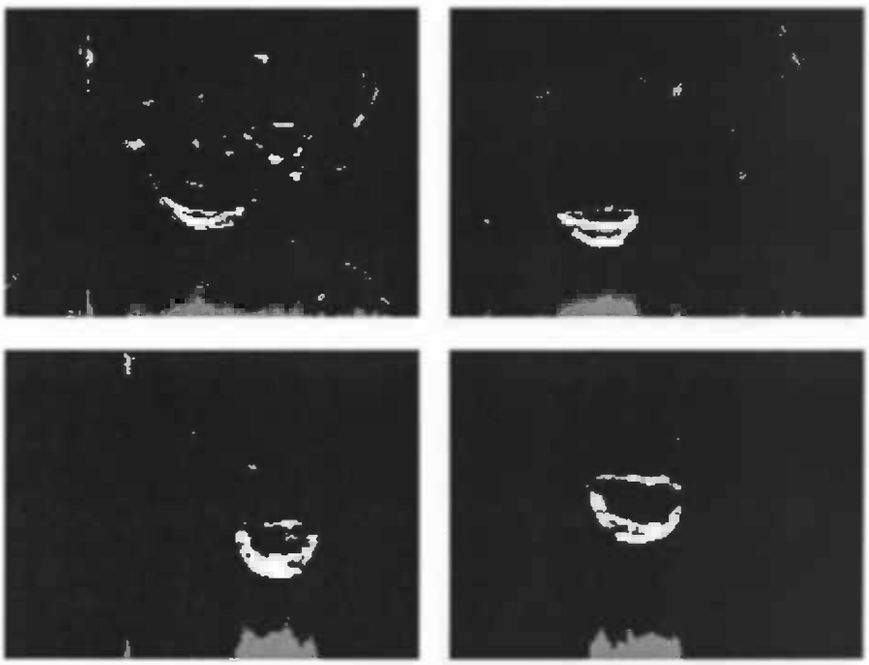


Figure D.4: Vertical density histogram. Note the blob in the center that relates to the width of the mouth.

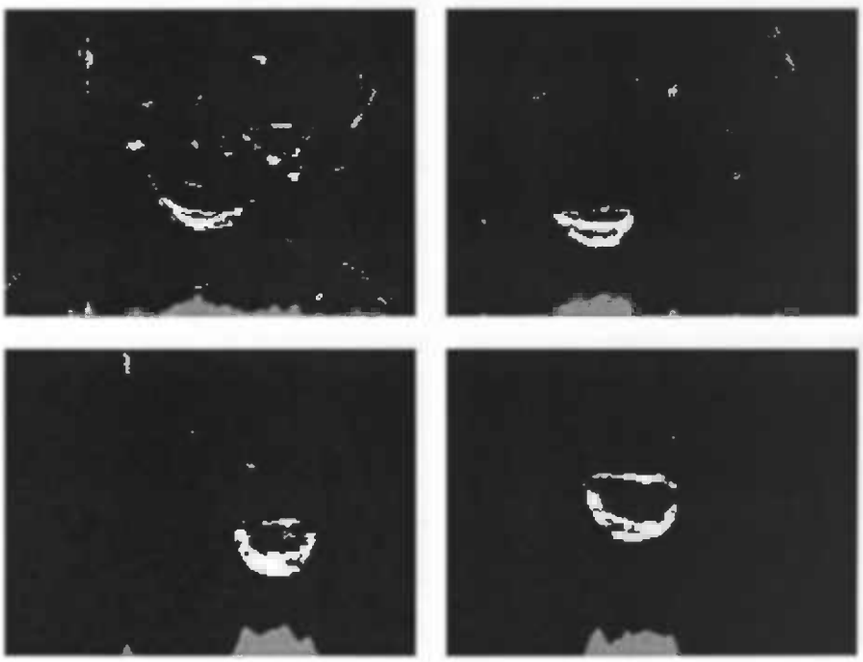


Figure D.5: Smoothed vertical density histogram.

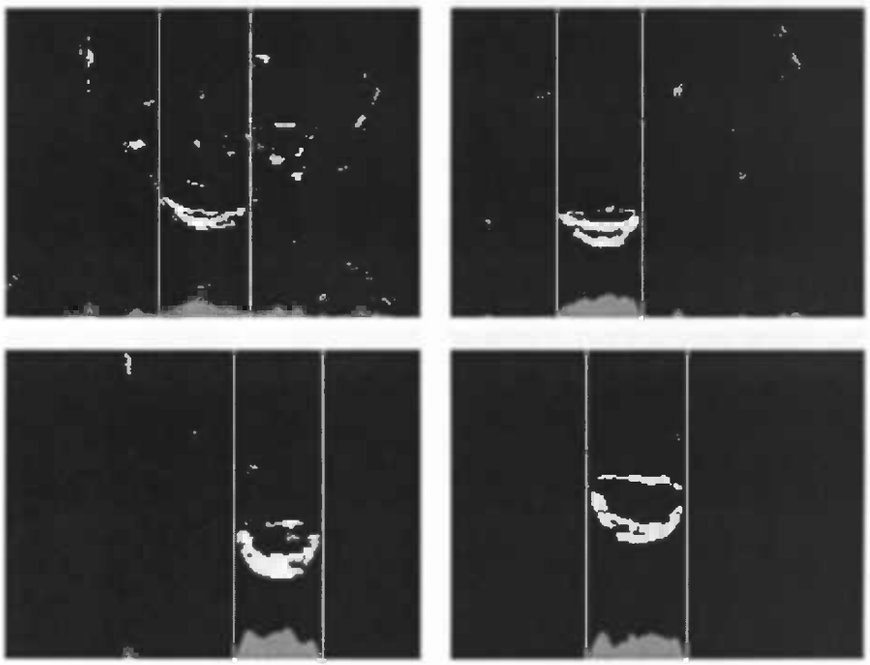


Figure D.6: Vertical region of interest. Note that the corners of the mouth in the upper left image fall outside the ROI.

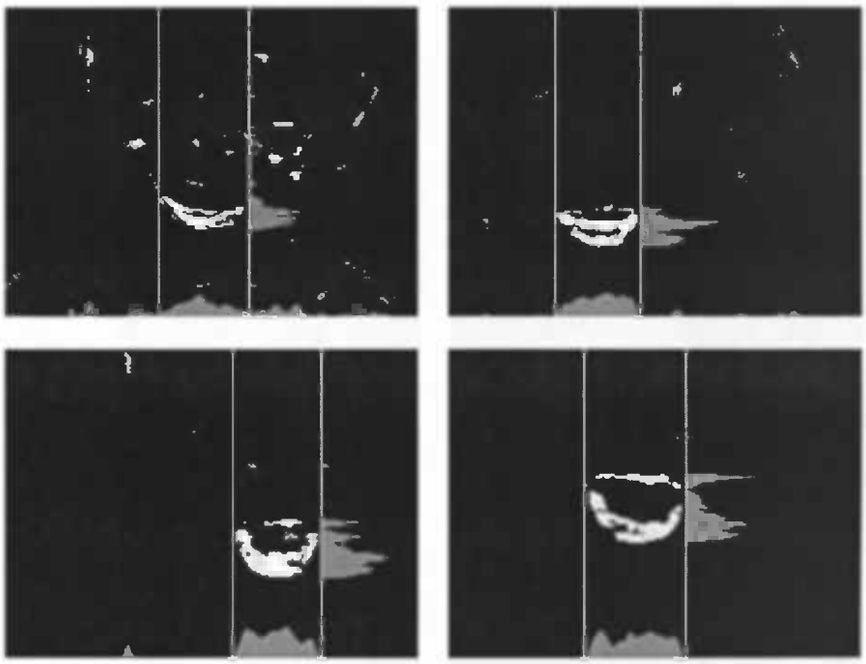


Figure D.7: Horizontal density histogram. Because only the pixels in the vertical region of interest are taken into account, many unwanted highlighted pixels are ignored.

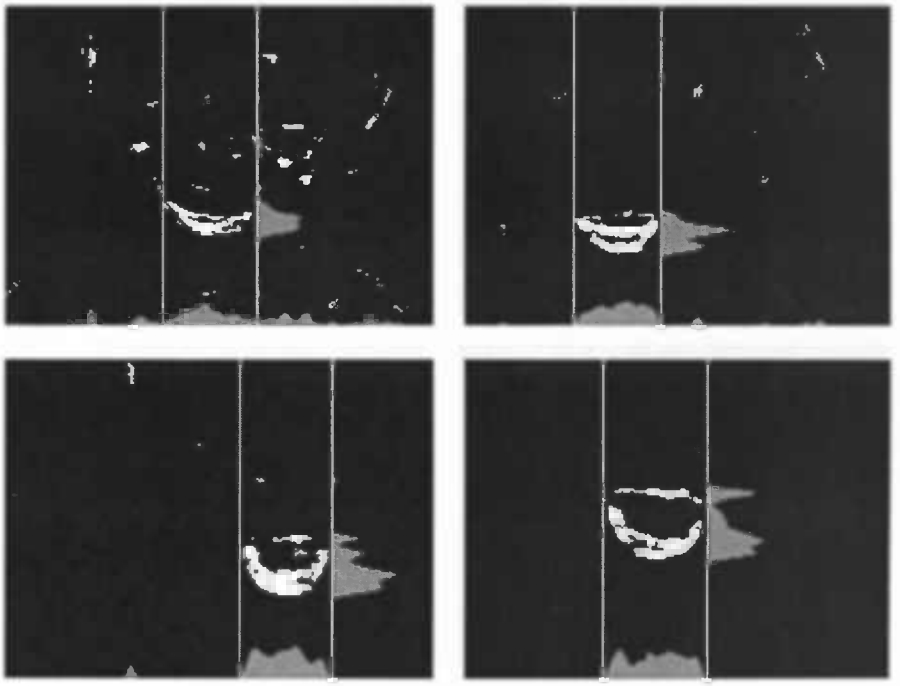


Figure D.8: Smoothed horizontal density histogram.

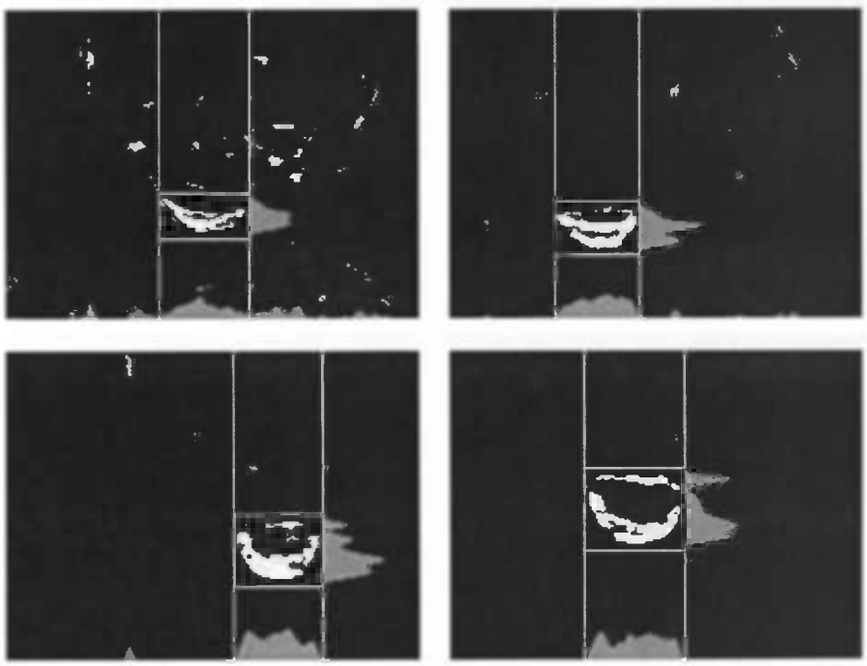


Figure D.9: Horizontal region of interest. Note that in the **upper left** image the upper lip is outside the region of interest.

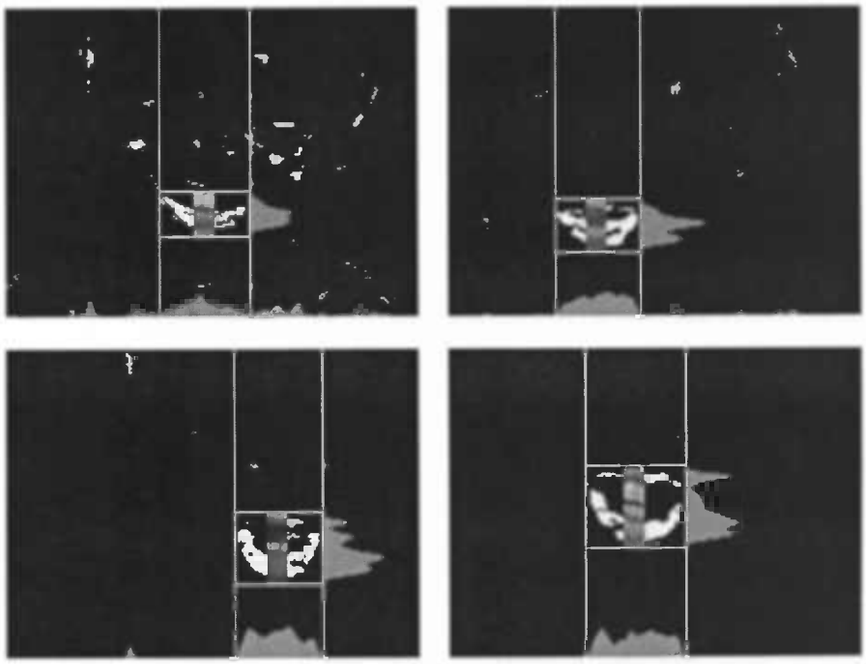


Figure D.10: A closer look at the center 10 columns in the region of interest. Mind the interrupted lip (see figure D.3) in the **upper right** image and the tongue (see figure D.2) in the **lower left** image.

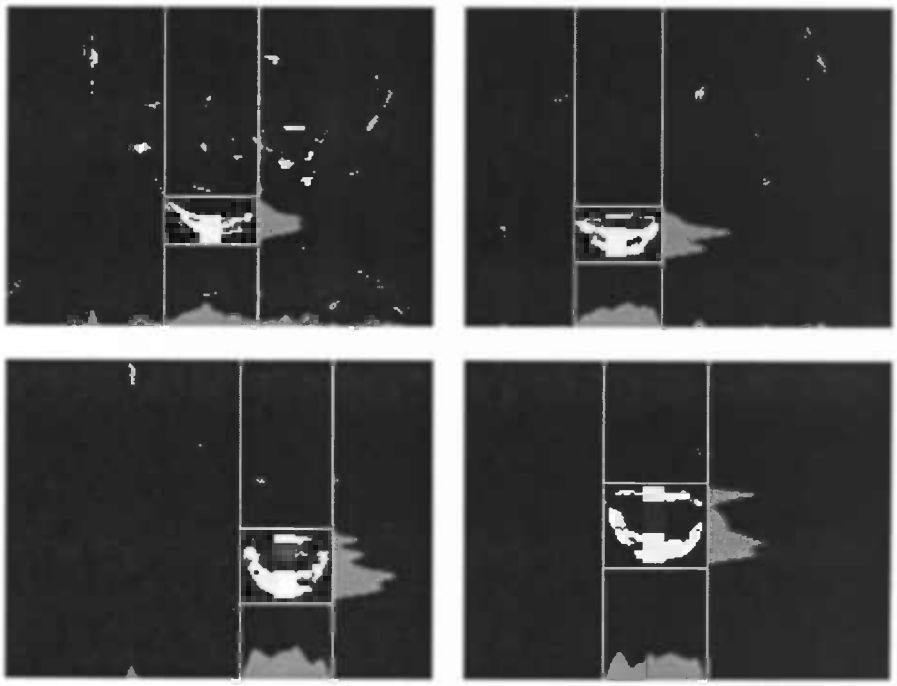


Figure D.11: Determining the inner lip height. **Upper left:** No upper lip is detected. **Upper right:** The lower lip was interrupted. **Lower left:** The tongue was recognized. **Lower right:** A correctly recognized opened mouth.

# Appendix E

# Results: traces

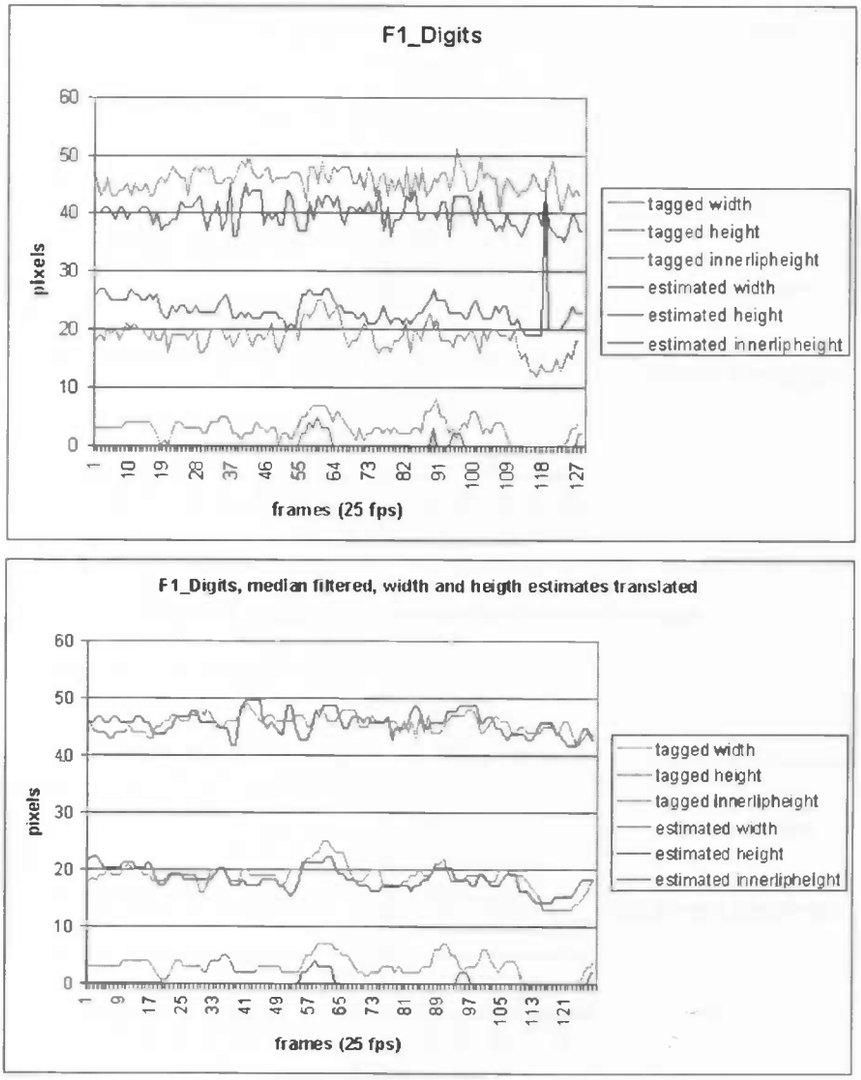



Figure E.1: Results for F1 Digits, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. Note that median filtering adequately filters out the spike in the estimated height near frame 118. The small variation of the features makes it hard to distinguish different states of the mouth.

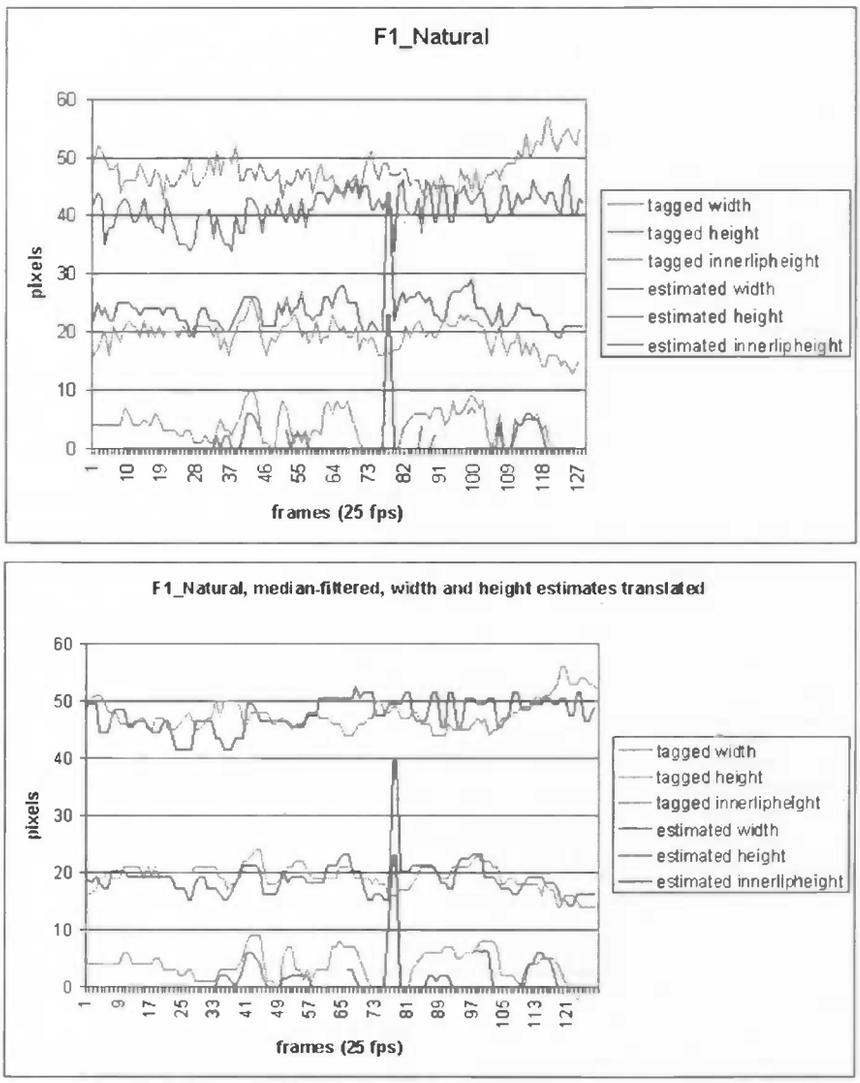



Figure E.2: Results for F1 Natural, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. The small variation of the features makes it hard to distinguish different states of the mouth.

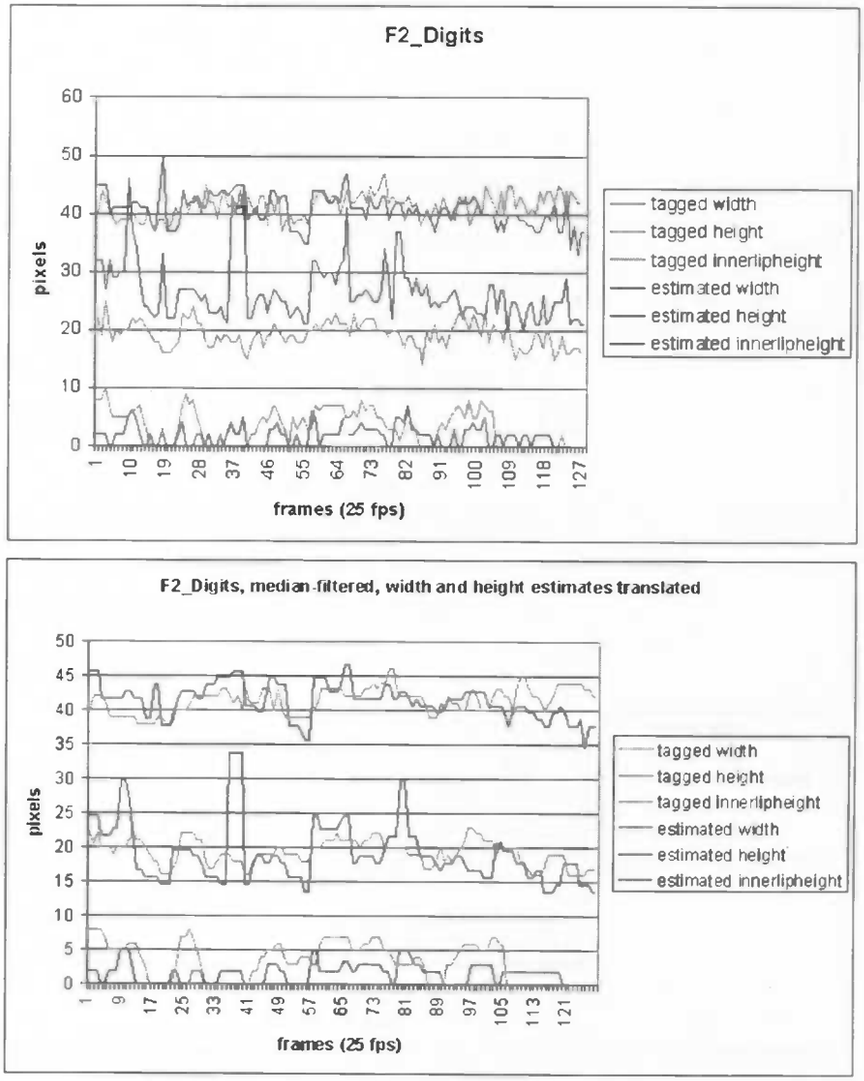



Figure E.3: Results for F2 Digits, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. Between frame 33 and 41 the region of interest includes a red part of the nose; the height is estimated too high.

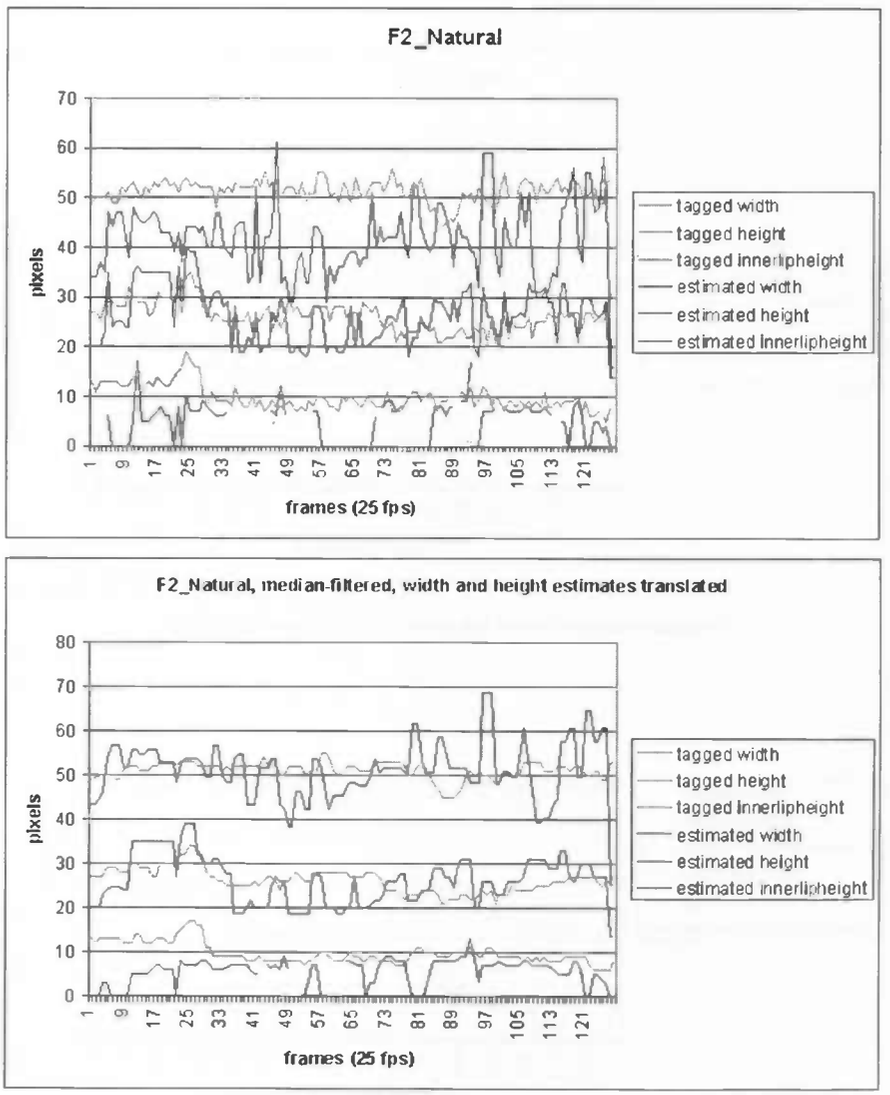



Figure E.4: Results for F2 Natural, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. In this trace the subject was blushing, which disrupted the selecting of the lips, since more parts of the face were red. Between frame 41 and frame 65 three situations occur where the upper lip falls outside the region of interest.

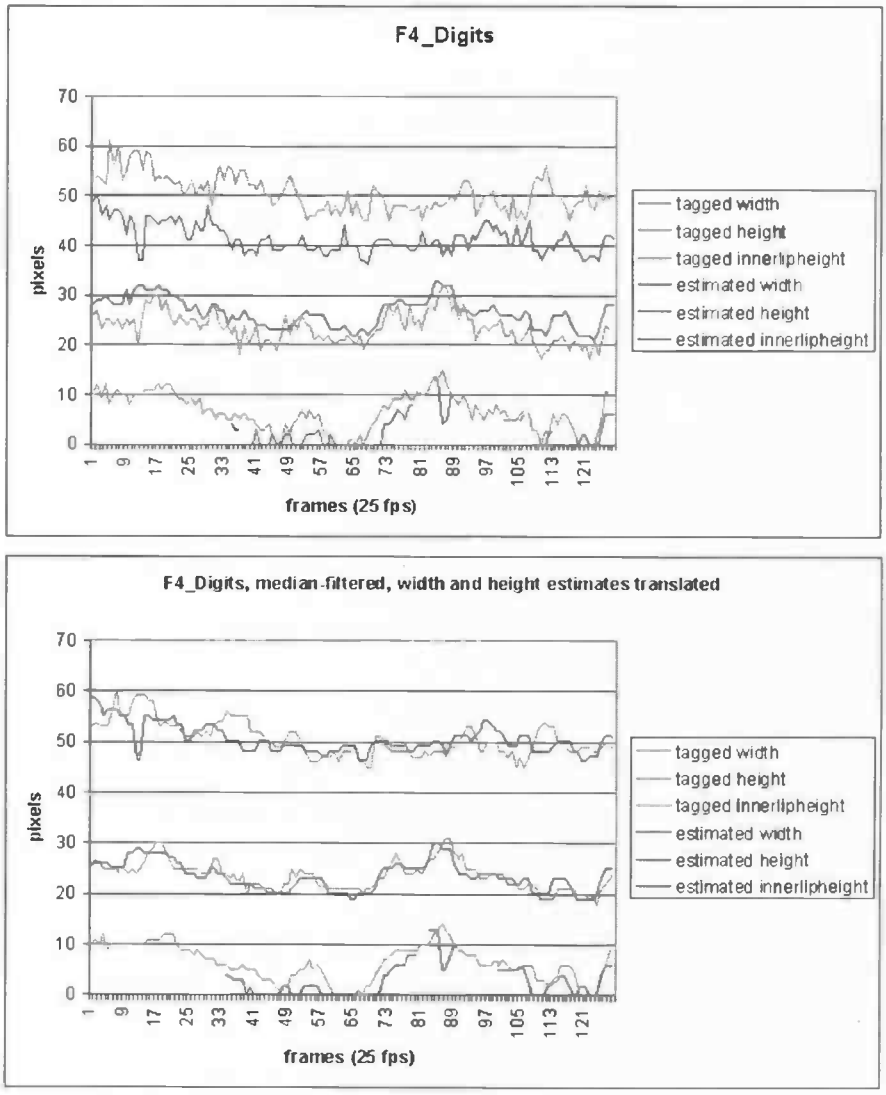



Figure E.5: Results for F4 Digits, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height.

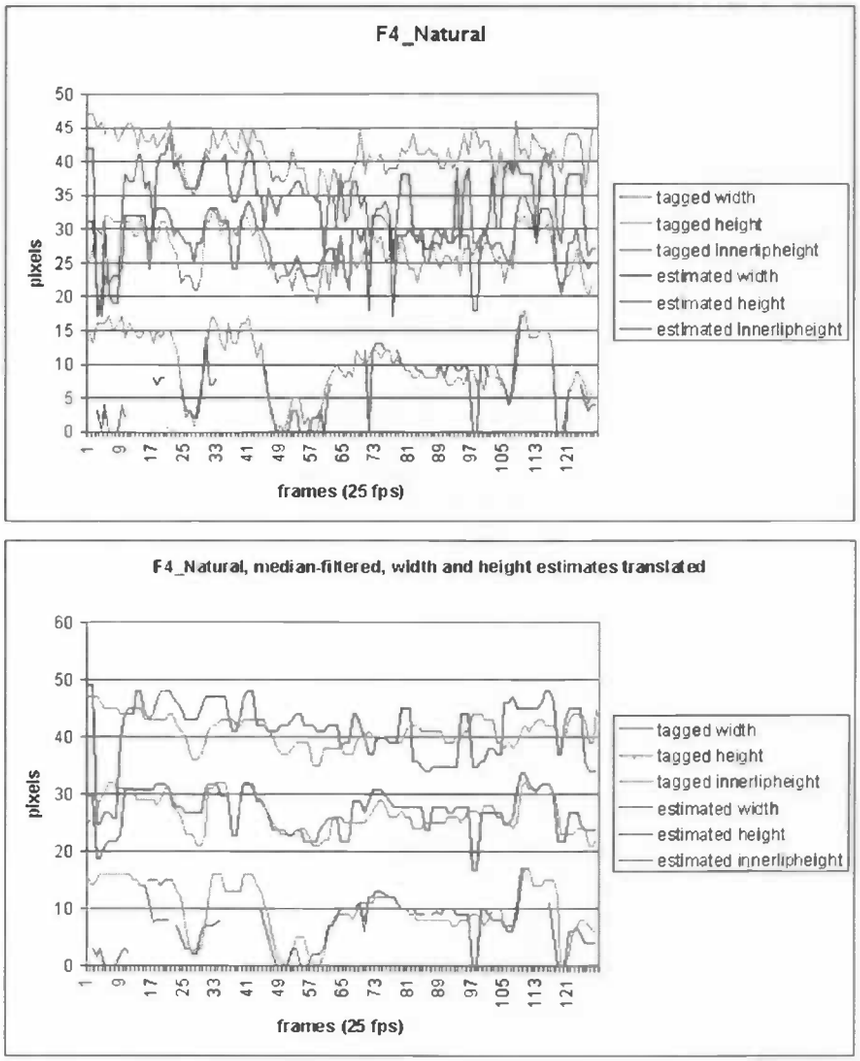



Figure E.6: Results for F4 Natural, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. Pale lips are harder to detect.

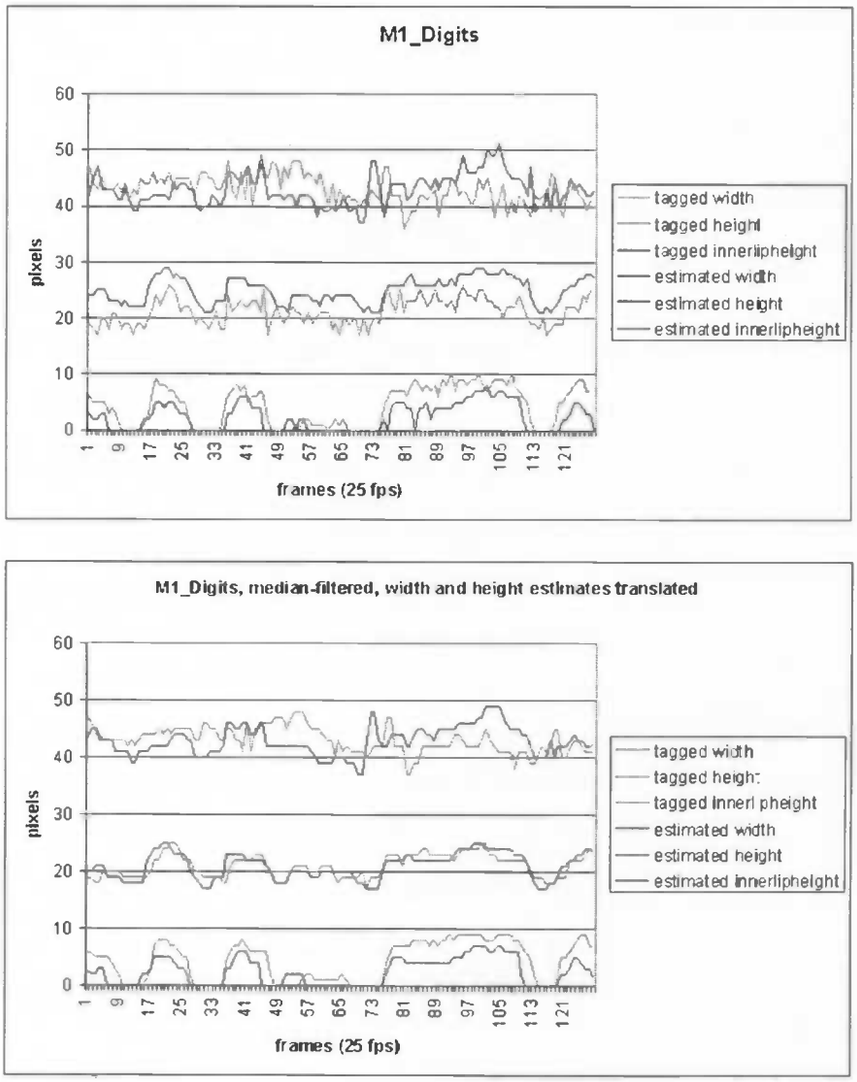



Figure E.7: Results for M1 Digits, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. The small variation of the width makes it hard to distinguish different states of the mouth.

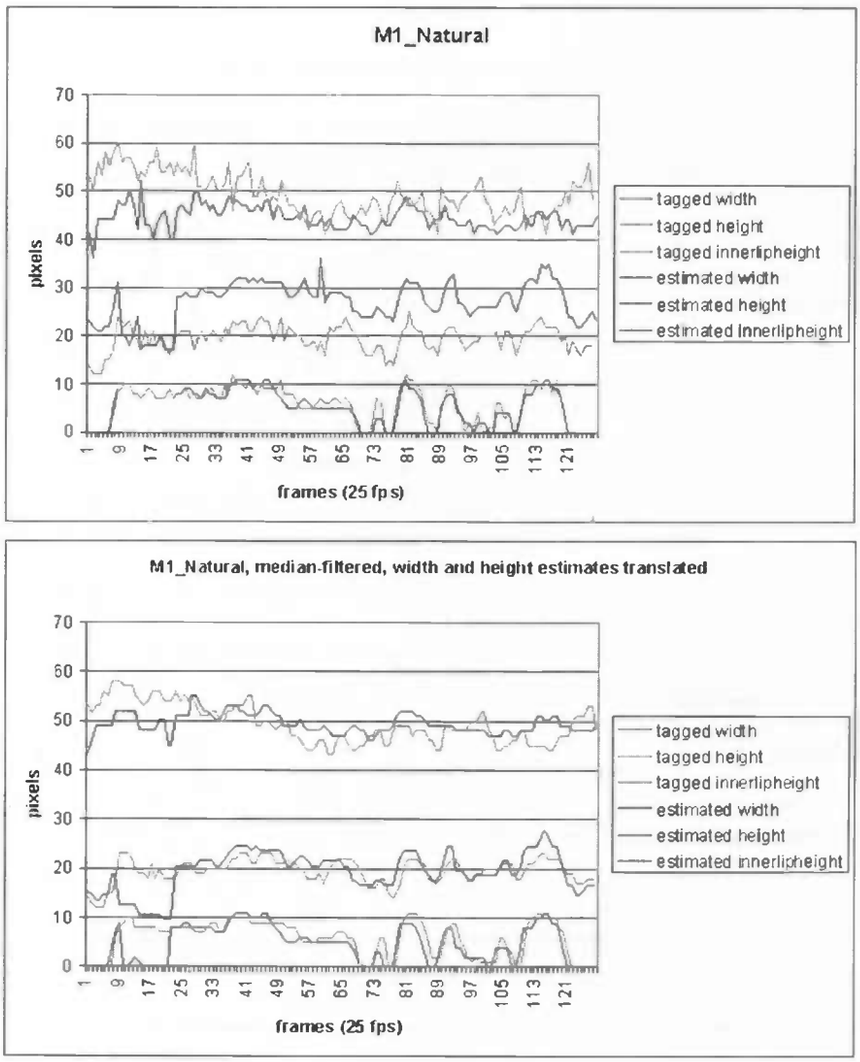



Figure E.8: Results for M1 Natural, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. Between the tenth and twentieth frame the upper lip falls outside the region of interest.

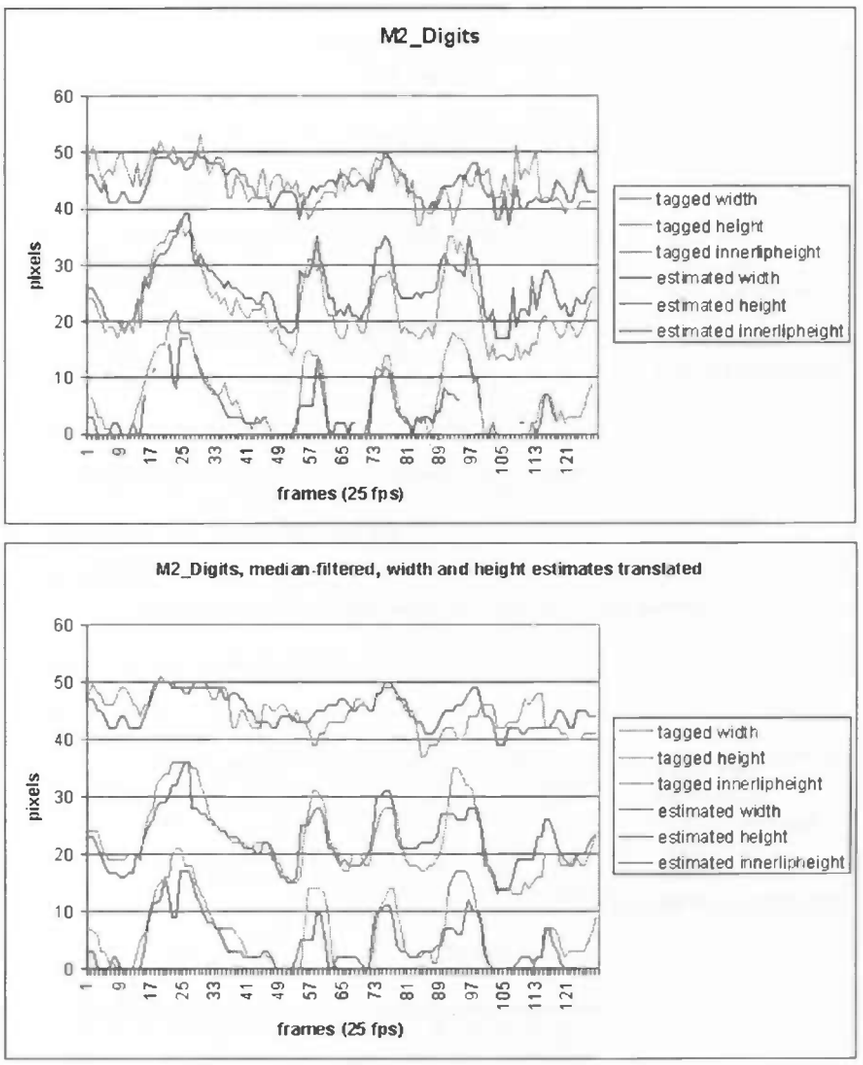



Figure E.9: Results for M2 Digits, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height. At frame 21 and 22 the red region of a tongue is confused for a part of the lower lip, which causes the sudden drop in the estimated inner lip height. In the surrounding frames the tongue is properly recognized.

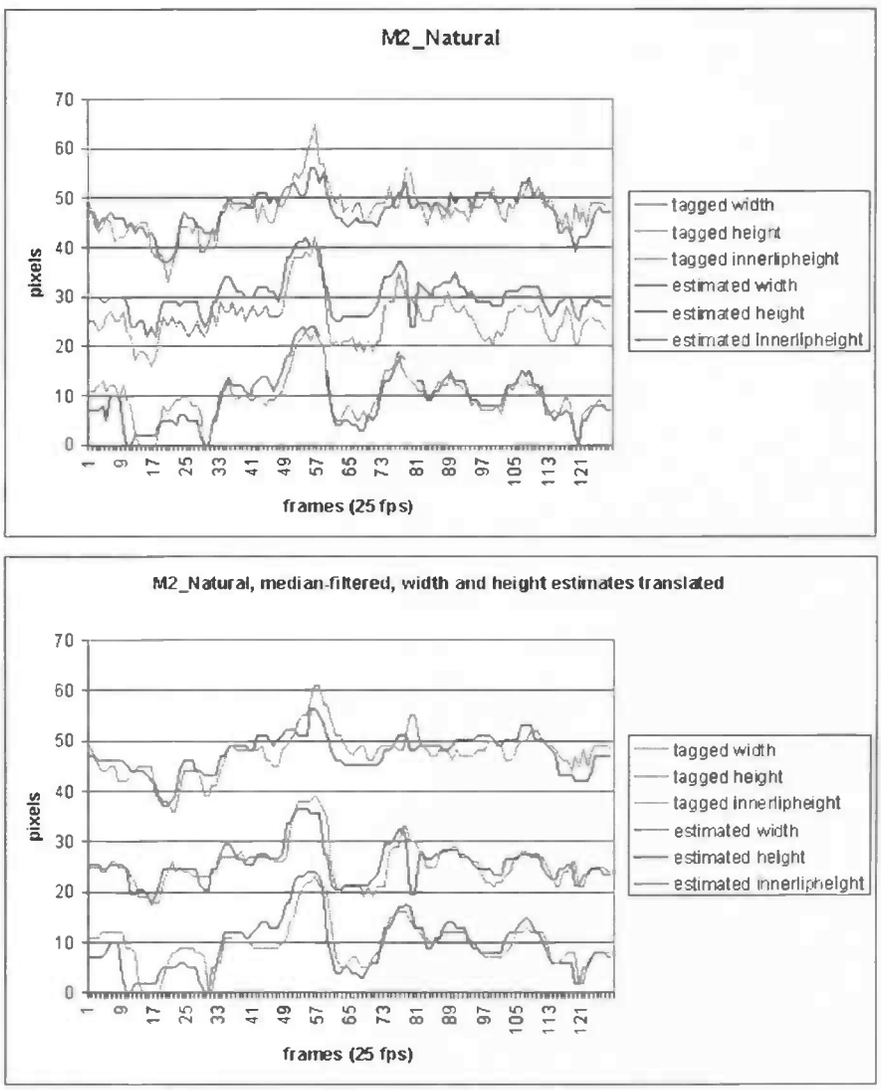



Figure E.10: Results for M2 Natural, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height.

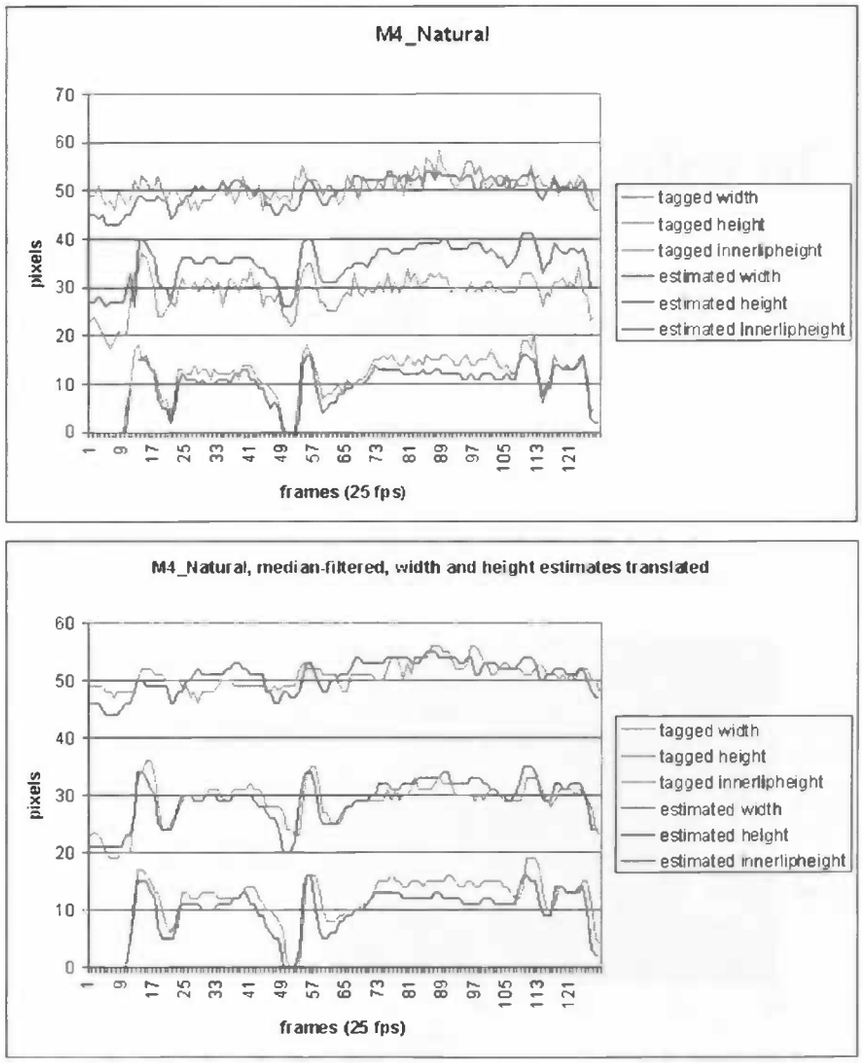



Figure E.11: Results for M4 Natural, in the bottom graph the traces are median filtered, with a window of 3 pixels. The higher trace is the width, the middle trace is the height and the lower trace is the inner lip height.

# Appendix F

# Results: some examples of phonemes and the corresponding shape of the mouth

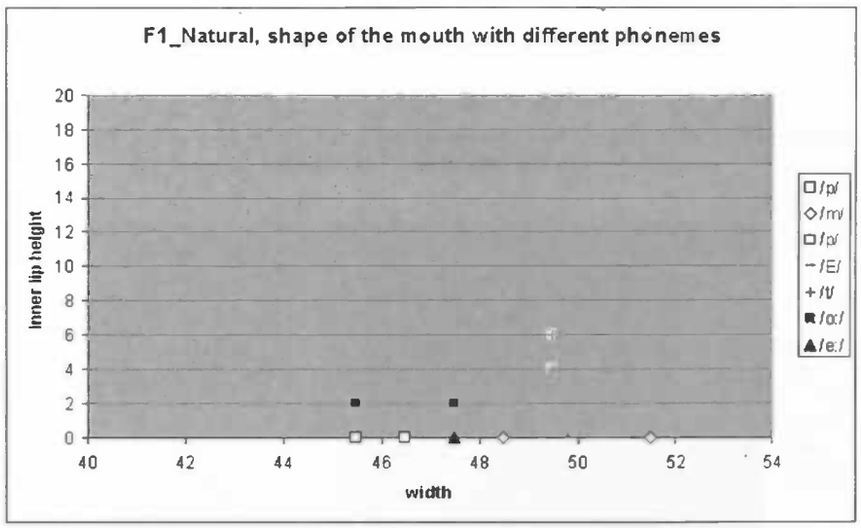



Figure F.1: Some phonemes for F1 Natural. The inner lip height and the width are the translated and smoothed estimates.

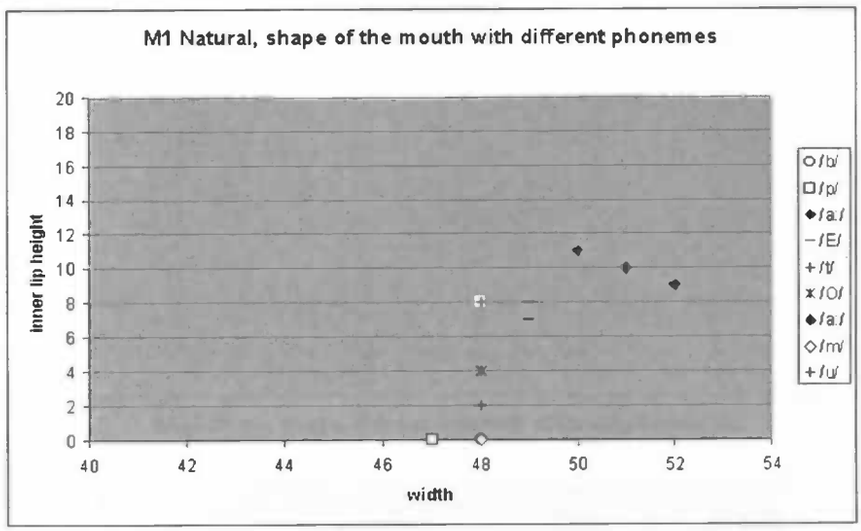



Figure F.2: Some phonemes for M1 Natural. The inner lip height and the width are the translated and smoothed estimates.

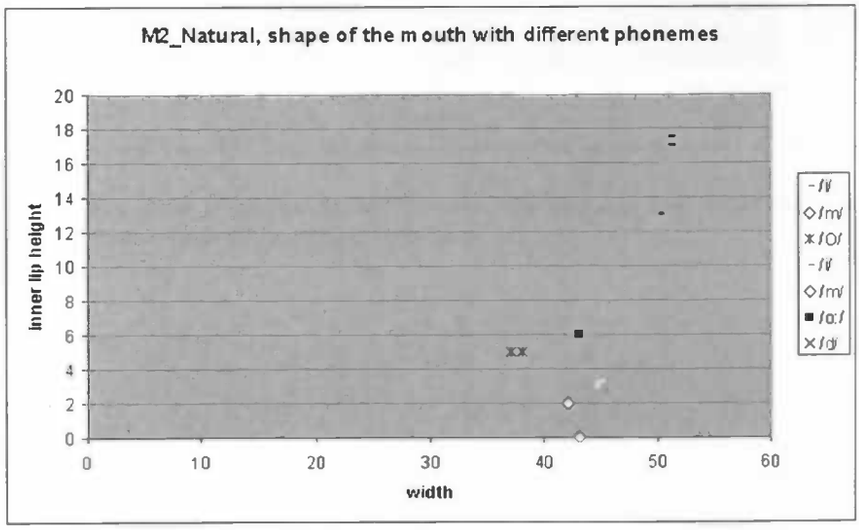



Figure F.3: Some phonemes for M2 Natural. The inner lip height and the width are the translated and smoothed estimates.

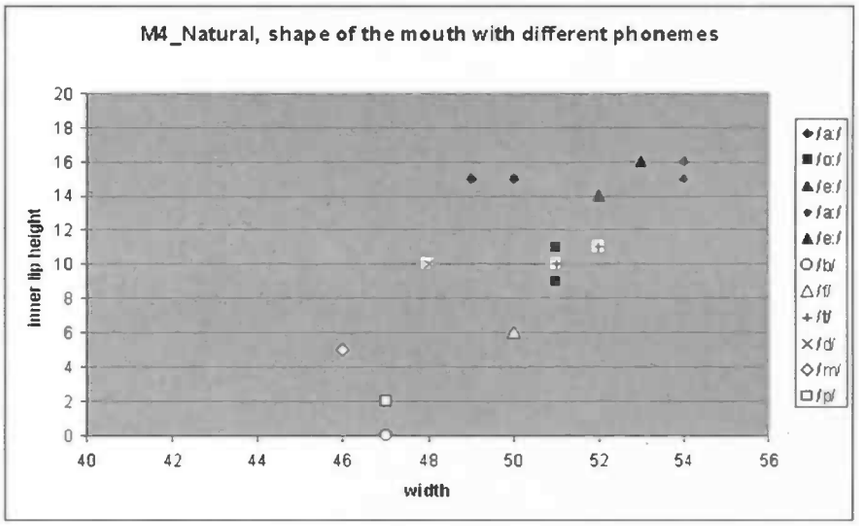



Figure F.4: Some phonemes for M4 Natural. The inner lip height and the width are the translated and smoothed estimates.

# Appendix G

# Software

The source code and documentation will be made available on http://www.ai.rug.nl/-~spiff/hal/.