Articulatory Features for Robust Visual Speech Recognition

by

Ekaterina Saenko

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2004

Author .....
Department of Electrical Engineering and Computer Science
August 15, 2004

Certified by .....
Trevor Darrell
Associate Professor
Thesis Supervisor

Certified by .....
James Glass
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Articulatory Features for Robust Visual Speech Recognition

by

Ekaterina Saenko

Submitted to the Department of Electrical Engineering and Computer Science
on August 15, 2004, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

This thesis explores a novel approach to visual speech modeling. Visual speech, or a sequence of images of the speaker's face, is traditionally viewed as a single stream of contiguous units, each corresponding to a phonetic segment. These units are defined heuristically by mapping several visually similar phonemes to one visual phoneme, sometimes referred to as a *viseme*. However, experimental evidence shows that phonetic models trained from visual data are not synchronous in time with acoustic phonetic models, indicating that visemes may not be the most natural building blocks of visual speech. Instead, we propose to model the visual signal in terms of the underlying *articulatory features*. This approach is a natural extension of feature-based modeling of acoustic speech, which has been shown to increase robustness of audio-based speech recognition systems. We start by exploring ways of defining visual articulatory features: first in a data-driven manner, using a large, multi- speaker visual speech corpus, and then in a knowledge-driven manner, using the rules of speech production. Based on these studies, we propose a set of articulatory features, and describe a computational framework for feature-based visual speech recognition. Multiple feature streams are detected in the input image sequence using Support Vector Machines, and then incorporated in a Dynamic Bayesian Network to obtain the final word hypothesis. Preliminary experiments show that our approach increases viseme classification rates in visually noisy conditions, and improves visual word recognition through feature-based context modeling.

Thesis Supervisor: Trevor Darrell
Title: Associate Professor

Thesis Supervisor: James Glass
Title: Principal Research Scientist

Acknowledgments

First of all, I would like to thank my husband for supporting me in pursuing my dreams, and for being a constant source of strength and inspiration throughout this work. To my mother, who is my biggest role model, - thank you for always being there for me, and for teaching me to love learning.

I am deeply grateful to my advisors Jim Glass and Trevor Darrell for their guidance and encouragement. Special thanks to Karen Livescu for sharing her ideas and for helping me adapt her feature-based DBN model and carry out the word recognition experiments. Thanks also to T.J. Hazen, Chia-Hao La, and Marcia Davidson for their work on data collection for the AVTIMIT corpus.

I would also like to thank Ozlem for keeping me sane and well-nourished; my officemates, Kevin, Louis-Philippe, Mario, Mike, Neal and Tom, for providing me with great advice and a fun atmosphere to work in; the members of the SLS and VIP groups for their mentorship and technical assistance; and all the wonderful people who have made my experience at MIT so enjoyable.

This research was supported by ITRI and by DARPA under SRI sub-contract No. 03-000215.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A major weakness of current automatic speech recognition (ASR) systems is their sensitivity to environmental and channel noise. A number of ways of dealing with this problem have been investigated, such as special audio preprocessing techniques and noise adaptation algorithms [3]. One approach is to take advantage of all available sources of linguistic information, including nonacoustic sensors [40], to provide greater redundancy in the presence of noise. In particular, the visual channel is a source that conveys complementary linguistic information without being affected by audio noise.

Using the images of the speaker's mouth to recognize speech is commonly known as lipreading. Long known to improve human speech perception [53], lipreading has been applied to ASR extensively over the past twenty years. The result is the emergence of two closely related fields of research. The first, *Visual Speech Recognition*, sometimes also referred to as *automatic lipreading* or *speechreading*, uses just the visual modality to recognize speech. The second, *Audio-Visual Speech Recognition (AVSR)*, combines both the audio and visual modalities to improve traditional audio-only ASR. Current AVSR systems are able to achieve an effective signal-to-noise (SNR) gain of around 10 dB over traditional audio-based systems [47].

1.1 Motivation

Overall, automatic lipreading promises to add robustness to human-machine speech interfaces. In practice, however, the visual modality has yet to become mainstream in spoken human-computer interfaces. This is partly due to the increased processing and storage demands, and partly to the relative novelty of the field. In particular, the lack of large, commonly available audio-visual corpora has hindered the development of practical algorithms. Furthermore, the reliance of current systems on high-quality video, recorded in controlled environments where the speaker is always facing the camera, is a major issue in practice. In fact, in field situations where acoustic channel noise can become a problem, it is possible that the visual channel will also become corrupted by noise, for example, due to inferior quality of recording equipment.

The need for improving the robustness of visual feature extraction algorithms is starting to attract attention in the research community. A recent study compared the performance of a state-of-the-art AVSR system on a typical "visually clean" studio database and a more realistic database recorded in offices and cars using an inexpensive web camera [46]. The results show that, although the visual modality remains beneficial even in such challenging conditions, the visual-only word error rate (WER) approximately doubles when moving from the studio to the office environment, and triples on the automobile data. This brings up an interesting research question of how to adapt systems trained on clean studio data to the varying levels of visual noise encountered in practice.

One of the other open problems in AVSR is the joint modeling of audio and visual information. There is an on-going debate about whether *Early Integration (EI)*, which assumes conditional dependence between the modes, or *Late Integration (LI)*, which assumes their conditional independence, is the correct model. It has been shown that, for certain architectures, asynchronous modeling of audio and visual data streams outperforms synchronous modeling. In order to allow for this observed *audio-visual asynchrony*, various extensions of the multi-stream HMM have been proposed. Such models normally assume that there is an underlying "visual process" generating

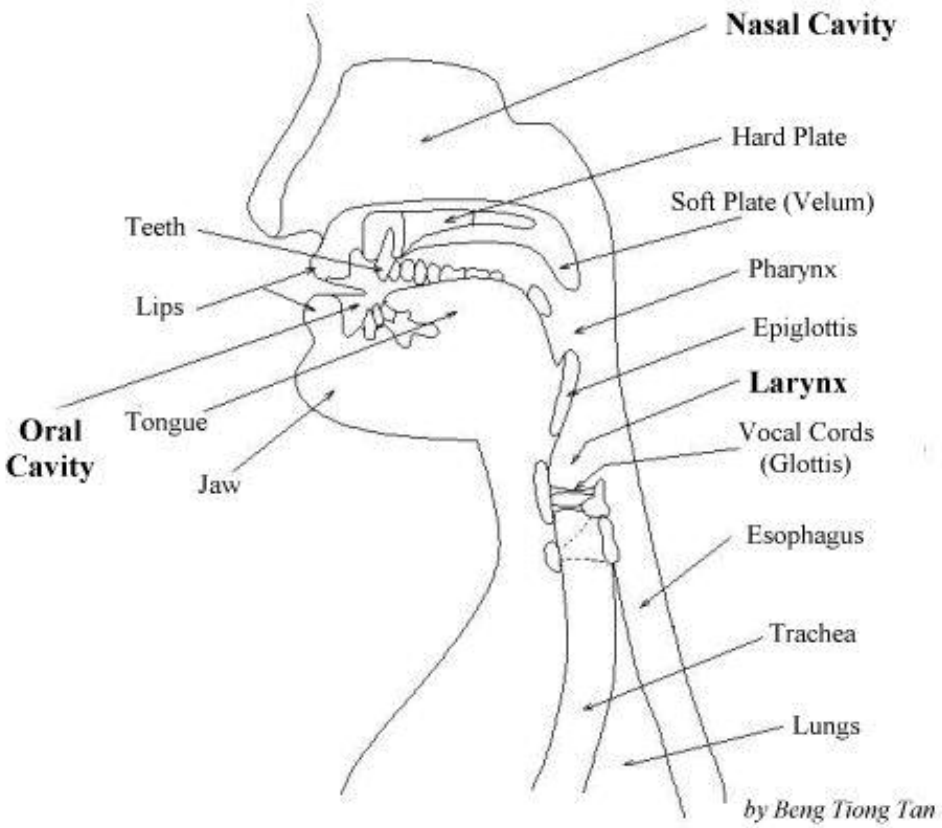


Figure 1-1: Human speech production.

the visual observations, and an underlying "acoustic process" generating the acoustic observations. The two processes correspond to the two hidden-state streams of the HMM. However, this view conflicts with the fact that a single underlying articulatory process generates both streams of observations. Thus, it seems incorrect to assume the existence of a strictly "visual" process that only influences visual observations. As this is still an area of active research, the question of what is the optimal audio-visual integration model remains to be settled.

1.2 Distinctive Features

The majority of automatic speech recognition systems developed in the past twenty years subscribe to the so called "beads-on-a-string" model of speech. In particular, they assume that speech consists of a sequence of contiguous basic units, or *phonemes*.

In the case of visual speech, the basic units correspond to the visually distinguishable phonemes, also known as *visemes*.

The above view is consistent with the early theory of generative phonology. However, it is being questioned as research moves away from laboratory-recorded corpora toward noisier, more spontaneous speech. The alternative view, as proposed by linguistic theory, is that *distinctive features* are the more fundamental atomic units of language. Distinctive features capture the natural classes in phonology. The existence of such classes has been motivated by several different phenomena: (i) acoustic regularities among speech sounds (e.g. [21]); (ii) phonemes behaving as a class as they participate in phonological processes (e.g. [7]); (iii) articulatory commonalities between phonemes (e.g. [25]); and (iv) the block diagonal structure of confusion matrices in human perception experiments (e.g. [38].)

Feature-based or acoustic phonetic approaches to automatic speech recognition were attempted in the 1970's. However, most systems used a rule-based recognition framework and were therefore not successful in dealing with the inherent variability of speech. Recently, the approach has been revisited using a more modern statistical learning framework (see [52], [27], [28], [31], [55], and [23] for some recent examples.)

In general, distinctive features may be defined in terms of acoustic or articulatory features. One particular theory describes speech as the combination of multiple streams of hidden *articulatory features (AFs)* [26]. *Articulation* is the process of changing the shape of the vocal tract using the articulators, i.e. the glottis, velum, tongue, lips and jaw (see Figure 1.2,) to produce different sounds [15]. From the point of view of articulation, each speech sound is described by a unique combination of various articulator states, for example: the presence or absence of voicing, the position of the tongue body and tongue tip, the opening between the lips, and so on. A word can be viewed as a sequence of articulator targets. Note that, in the phoneme-based approach to speech modeling, a simplifying assumption is made that words can be broken up into a sequence of phonemes, each of which maps to a canonical articulatory configuration.

Several motivating factors have been identified for the use of distinctive (or artic-

ulatory) features in speech recognition systems [41]. One is a belief that distinctive features will minimize extra-linguistic variability related to speaker identity and signal distortion. Another is that features provide better modeling of co-articulation and pronunciation. A third motivation is the ability to represent sounds of any language using a compact set of distinctive features. Thus, existing models become portable to a new language, where a phoneme model would have to be re-trained.

One of the advantages of representing speech as multiple streams of articulatory features is the ability to model each feature independently, and even to allow them to de-synchronize. It has been noted that spontaneous, conversational speech is difficult to transcribe in terms of conventional phoneme units, and presents a challenge for existing ASR systems [16]. On the other hand, feature-based pronunciation models have been shown to be better at accounting for the types of pronunciation variations that occur in spontaneous speech, partly due to their ability to model the asynchronous nature of articulation [30].

Another advantage of AF-based modeling is its robustness in noisy environments. Experiments in acoustic speech recognition have shown that articulatory-feature systems can achieve superior performance at high noise levels [27]. The de-compositional nature of the approach can help increase robustness in two main ways. First of all, it combines several sources of information about the underlying speech process, derived independently via parallel classifiers. Therefore, it can take advantage of the fact that some of the features may be easier to classify than others under conditions of image corruption, low resolution, or speaker differences. Confidence values can be used to assign each feature a different weight, effectively reducing the overall number of distinguishable classes. The second advantage is that, because there are fewer possible values for each feature class than there are phonemes, the training data set generates more instances of each feature class value than each phoneme. This, in turn, leads to a larger amount of training data per feature value.

1.3 Overview of Proposed Approach

As described in the previous section, articulatory feature modeling is a promising alternative to the "beads-on-a-string" model that is actively being explored by researchers for acoustic ASR. It would be interesting to see if the AF approach also has potential to improve robustness, or to help model pronunciation, in the case when the input is visual. And, if the input is audio-visual, perhaps a hidden feature model is a better underlying structure for A/V integration? With these questions in mind, we propose to study the application of the AF approach to visual speech recognition. In the rest of this section, we provide a general overview of the proposed approach, which is derived from human speech production and inspired in part by the distinctive feature models described in the previous section. Our hypothesis is that the benefits of feature-based recognition will also apply in the case of visual speech.

1.3.1 A Production-based Model of Visual Speech

Rather than using visemes as basic recognition units, we suggest representing visual speech classes in terms of the underlying articulatory processes, or articulatory features. The features are associated with articulatory gestures and have both visual and acoustic consequences. Both the low-level viseme units and the higher-level word units can be represented as a combination of multiple streams of such features.

Of course, since we are dealing with the visual modality, we are limited to the modeling of visible articulators. From the video of the speaker's lower face region, we can obtain information about the position and relative configuration of the jaw, lips, teeth, and tongue. Also, in addition to static features, the video contains dynamic articulatory features, for example, lips closing and opening, tongue protruding and retracting through teeth, lower lip touching upper teeth, lips protruding, and so on. However, the rest of the articulators are not visible under normal circumstances.

The typical process of visual speech recognition goes through three stages, illustrated in Figure 1-2: 1) face detection and mouth tracking, 2) low-level image feature extraction, 3) categorization into viseme classes, and 4) the combination of frame-

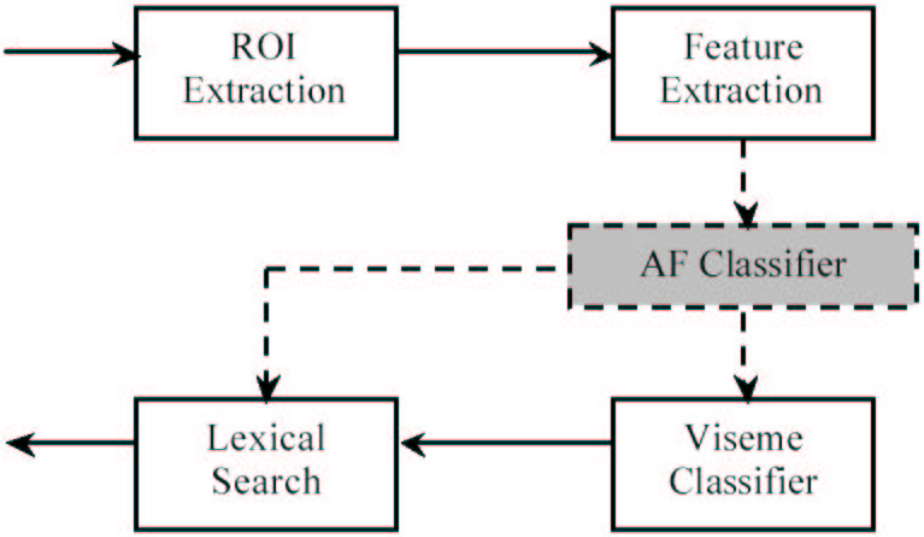


Figure 1-2: Articulatory-Feature approach to visual speech recognition.

level scores over time in order to find the most likely word sequence. We introduce an extra step after the initial preprocessing of the image, but before the viseme scores are computed. In this step, the input data are classified in terms of several articulatory features by a set of parallel statistical classifiers. Afterwards, the lexical search can either proceed right away, using the obtained articulatory feature scores, or follow an additional step of classification into the higher-level visemic categories.

Our approach is in many ways identical to the multi-stream articulatory-feature approach to audio speech modeling. We are essentially proposing to model visual speech as multiple streams of visible linguistic features, as opposed to a single stream of visemes. In fact, most of the articulatory events described above have direct equivalents in the feature set used for pronunciation modeling in [30]. For example, the visual feature of the lips closing and opening corresponds to the LIP-OPEN feature. Therefore, an integrated AF-based audio-visual speech recognizer can use the same underlying feature set. However, due to the complementary nature of the two modalities, some features may be easier to derive from the audio stream, and others from the video stream, especially in the presence of noise. For instance, it is

Figure 1-3: Full bilabial closure during the production of the words "romantic" (left) and "academic" (right).

known from perceptual studies that acoustic noise affects the detection of place of articulation (e.g. glottal, bilabial) more than voicing [38]. On the other hand, since place information is highly distinguishable visually, it might be less affected by visual noise than other features.

The difference between our method of classifying articulatory features and the conventional method of classifying visemes is illustrated by the following example. Suppose we were to model the phoneme /m/ in two different phonetic contexts, romantic and academic, using a single, context-independent visual model. The image snapshot taken at the moment of complete closure during the production of /m/ in each context is shown in Figure 1-3. Both examples would be considered to belong to a single viseme class (the bilabial viseme) and to have the same open/closed feature value (fully closed.) However, their appearance is different: in the second context, the distance between the mouth corners is roughly 25% wider. This suggests the presence of contextual information. In fact, the preceding /ow/ in romantic causes the /m/ to be rounded, whereas the preceding /eh/ in academic does not. Thus, modeling lip rounding and lip opening as two separate articulatory features would allow us to recover more information than just modeling the /m/ viseme.

Our proposed method of extracting articulatory feature information from visual input is similar in spirit to the method of extracting geometric mouth parameters often used in automatic lipreading systems. However, there is a subtle difference between what we call *visual articulatory features* and what is referred to as *visual features* in the literature, for example, in [58]. In the latter work, a set of visual features, including

mouth width, upper/lower lip width, lip opening height/width, etc., are extracted by tracking a set of points on the lips. These features, plus features indicating the presence of teeth and tongue obtained by color segmentation of the mouth region, form the input observation vector to the HMM that performs word recognition. In contrast, we treat articulatory features as the hidden states underlying the production of the surface observations. Thus, our model can use the same preprocessing techniques as the one described above to produce a *surface* feature vector, which can be used as input to a statistical classifier. The classifier then assigns abstract class labels to the input vectors that correspond to the underlying articulatory features, such as "lip-open", "lip-rounded", "fricative", etc. One of the potential benefits of our approach is the ability to use different low-level measurements for each articulatory feature. For example, the classifier for "lip-rounded" could take optical flow measurements as input, while the "teeth" classifier could use color information.

1.3.2 Feature-based Audio Visual Integration

One of the main open problems in AVSR is the joint modeling of audio and visual information. At a high level, there are two possible approaches: 1) use models based on human perception, and 2) use models based on human speech production. Although the speech production mechanism is fairly well understood, scientists do not yet know exactly how humans process and integrate perceived acoustic and visual speech. Nevertheless, since the discovery of the McGurk effect in the 70's, there has been active research in this area.

The AVSR research community has focused primarily on models motivated by human perception. There is an on-going debate about whether Early Integration (EI), which assumes conditional dependence between the modes, or Late Integration (LI), which assumes their conditional independence, is the correct model. In fact, there is perceptual evidence to support both. Several studies regarding the McGurk effect and Voice Onset Time perception indicate that humans integrate early, before categorizing speech phonetically. On the other hand, studies of acoustic speech perception indicate that humans perform partial recognition independently across different frequency

bands. Note that the question of audio and visual stream dependence is closely related to the concept of audio-visual asynchrony. If the audio and visual cues are always synchronous in time, then they are temporally dependent and therefore fit into the EI model. If the audio and visual cues are asynchronous, then they are temporally independent and fit into the LI model.

Unfortunately, many papers in the AVSR field have not provided a lot of motivation for the proposed integration models, either in the form of psycho-linguistic studies, or models of speech production. A popular approach is to train various extensions of the multi-stream HMM on a large dataset and see which one produces better recognition results, although, the authors claim that the extensions are based on studies of human perception. In fact, several human intelligibility experiments (Massaro and Cohen [34], Smeele, et al [51], etc.) do show that the human integration process is robust to artificially introduced asynchronies between the audio and video of up to 200 ms.

In the paper titled "Asynchrony modeling for audio-visual speech recognition" [18], Gravier, et al motivate the state-asynchronous Product HMM by the following statement: "it is well known that, although the visual activity and the audio signal are correlated, they are not synchronous. As a matter of fact, the visual activity often precedes the audio signal by as much as 120 ms." This last result is from the paper titled "Eigenlips for Robust Speech Recognition" by Bregler and Konig [4], who studied the cross-modal mutual information between the acoustic and visual feature vectors, offset from each other by various amounts. They found that the maximum mutual information occurs when the visual features are offset by a negative 120 ms relative to the audio. Bregler and Konig explain their findings by the "forward articulation" effect, which they say has been confirmed by psychological experiments by Benoit. However, since Bregler and Konig used sequences of 3-8 spelled letters as their training data, their result may not apply to continuous, large-vocabulary speech. In fact, if there were silences between the letters, then the "beginning of word" situation, where the lips start moving before any sound is made, would be predominant. On the other hand, our own observations of continuous speech indicate

that, in addition to the "forward articulation" effect, the "backward articulation" effect is also present. Some evidence of this more general de-synchronization model can be found in "Audio-Visual Speech Modeling for Continuous Speech Recognition" by Dupont and Luettin [14]. Their (preliminary) studies of asynchrony between the acoustic and visual HMM streams showed that, in some cases, the visual state transitions were delayed, and in some cases, the acoustic transitions were delayed.

Besides Bregler and Konig's forward articulation effect and studies of human perception of asynchronous signals, there is little other explanation of how the proposed HMM variants are modeling the underlying speech process. It is normally assumed that there is a "visual process" generating the visual observations and an "acoustic process" generating the acoustic observations. These two processes are modeled by the two hidden-state streams of the HMM. However, from the point of view of human speech production, there is a single underlying articulatory process that generates both streams of observations. It seems incorrect to assume that there is a "visual" process that influences only visual observations.

In fact, if we view the problem from the speech production perspective, the two "modalities" are really just two different ways to observe the same underlying process. Thus, audio-visual speech is a multi-media rather than a multi-modal phenomenon. The underlying process producing the observations is the behavior of the vocal tract and the actions of several articulators, e.g. velum, lips, and tongue. Of these articulators only some are visible, influencing the visual observations. In some cases, such as when the vocal folds are vibrating, the visible articulators can also influence the resulting sound, thus affecting the acoustic observations. At other times, such as during periods of silence, the speaker can move the visible articulators in order to anticipate the following sound, in which case they have no effect on the acoustic observations. The above reasoning suggests that it is more appropriate to model the independence (asynchrony) of certain underlying articulators, as opposed to the independence (asynchrony) of the separate "visual" and "audio" processes.

If we look closely at the physiology of the speech production mechanism, we see that the apparent asynchrony of the visual and acoustic observations is caused by 1)

co-articulation and 2) the fact that visible articulators are not always involved in the production of a phone. Co-articulation can be described as follows. During speech generation the phonetic articulators move from target positions of one phone to target positions of the succeeding phone. These movements are planned by the brain in such a way that the effort of the muscles is kept to a minimum. If an articulator must reach a certain target position to produce a phone and the preceding phone does not need that articulator, then the articulator itself may anticipate its movement toward the next target position before the production of the previous phone is finished. This effect is known as "anticipatory coarticulation." On the contrary, an articulator may wait to release a target position corresponding to the pronounced phone, if it is not required by the next phone. This effect is known as "preservatory coarticulation." If, in addition, the articulator in question is visible, but the target phone is produced mainly by invisible articulators (eg. /n/), then the result is the apparent disagreement of the visible and audible phone class. For example, the acoustic /n/ in "and thread" can look like the following /th/, or, in "don't", like the preceding rounded /ow/.

In order to better relate the audio-visual information to the underlying speech mechanisms, we may wish to model the state of the articulators, as opposed to visemes and phonemes. This would allow us to model audio-visual asynchrony as the de-synchronization of the underlying articulators.

1.4 Goals and Outline

To summarize, the main motivation behind this work is to develop a production-based approach to visual speech modeling. We would like to depart from the "beads-on-a-string" phoneme/viseme model and create a visual extension of the feature-based framework for audio speech recognition. We hope that such a framework will provide a better solution to the problems of noise robustness, context modeling, and audio-visual integration.

The main goals of this work are: i) to analyze multi-speaker data in order to determine distinguishable visual features; ii) to ascertain the influence of different

signal representations on the features; iii) to construct articulatory features detectors; iv) to incorporate these detectors into a word recognition system; and v) to compare performance to the baseline viseme model.

First, Chapter 2 will present an overview of related research. Then, Chapter 3 will provide a detailed description of the proposed recognition system. Chapter 4 will deal with the issues surrounding the design of an articulatory feature set, and Chapter 5 will describe the results of experimental evaluation. Chapter 6 will present our conclusions and talk about future work directions.

Chapter 2

Related Work

In this chapter, we provide an overview of previous work related to the fields of audio-visual speech processing and feature-based speech recognition.

2.1 Audio-Visual Speech Processing

The first audio-visual speech recognizer was designed by Petajan in 1984 [44]. Since then, over one hundred research articles have been published on the subject. Applications have ranged from single-subject, isolated digit recognition [44], to speaker-independent, large-vocabulary, continuous speech recognition [39]. The majority of reported AVSR systems have achieved superior performance over conventional ASR, although the gains are usually more substantial for small vocabulary tasks and low signal-to-noise ratios [46].

The main issues involved in the development of AVSR systems are 1) visual feature design and extraction, 2) the choice of speech units, 3) classification, and 4) audio-visual integration. Although the second and third issues also apply to audio-only systems and are therefore often resolved in the same way for both modalities, the first and the last issues are unique to audio-visual systems.

2.1.1 Visual Feature Extraction

Visual feature design falls into three main categories: appearance-based, shape-based, and a combination of the two. Appearance-based approaches treat all intensity and color information in a region of interest (usually the mouth and chin area) as being relevant for recognition. The dimensionality of the raw feature vector is often reduced using a linear transform. Some examples of this "bottom-up" approach include simple gray levels [17]; principal component analysis of pixel intensities [4]; motion between successive frames [33]; transform-based compression coefficients [48]; edges [1]; and filters such as sieves [36].

In contrast, shape-based methods usually assume a top-down model of lip contours. The parameters of the model fitted to the image are used as visual features. Some examples of shape-based features include geometric features, such as mouth height and width [44], [1], [5], [58]; Fourier and image moment descriptors of the lip contours [19]; snakes [24]; and Active Shape Models (ASM) [10]. In general, lip contours alone lack the necessary discriminative power, so they are often combined with appearance. For example, it was shown that the addition of appearance to shape significantly improves the lipreading performance of the ASM [36]. The result is an Active Appearance Model (AAM) [9], which combines shape and appearance parameters into a single feature vector. A similar model is the Multidimensional Morphable Model (MMM), developed in [22].

2.1.2 Classification

Visual speech recognizers can differ in their choice of classification techniques. Due to the dynamic nature of speech, the most common classifier used is a Hidden Markov Model (HMM), which allows statistical modeling of both the temporal transitions between speech classes, and the generation of class-dependent visual observations [39]. Although most HMMs use a Gaussian Mixture Model classifier for the latter task, several other classification methods have been suggested, including simple distance in feature space [44], neural networks [29] and Support Vector Machines (SVMs) [17].

In this work, we employ SVMs, which are capable of learning the optimal separating hyperplane between classes in sparse high-dimensional spaces and with relatively few training examples. Details of the SVM algorithm can be found in [57].

2.1.3 Audio-Visual Integration

In the case of audio-visual speech recognition, a major area of ongoing research is the integration of the two modalities in such a way that the resulting recognizer outperforms both the visual-only and audio-only recognizers. Integration algorithms generally fit into one of two broad categories: feature fusion and decision fusion, sometimes also referred to as early integration and late integration. Feature fusion involves training a single classifier on the fused bimodal data vectors [56], whereas decision fusion involves training separate single-modality classifiers and then combining their outputs, for instance, as a weighted sum [14]. Decision fusion can occur at any level (e.g., HMM state, phoneme, word, or sentence,) although very early stage fusion techniques are commonly referred to as hybrid fusion [47], [8].

Although we do not directly address the issue of audio-visual integration in this thesis, the proposed articulatory-feature model could be extended to include both acoustic and visual observations. A feature based framework has the advantage of providing a natural common model whose parameters may be jointly estimated from visual and acoustic cues simplifying the task of data fusion from multiple modalities [42].

2.2 Audio-Visual Speech Corpora

Unfortunately, no common large AVSR corpus has been publicly available, making the majority of reported algorithms difficult to compare. Several corpora have been created by researchers in order to obtain experimental results for specific tasks. Those made available for public use have come mostly from universities, and are generally not as extensive as the ones collected by private research labs. Many of the former contain recordings of only one subject [5]. Those with multiple subjects are usually

limited to small tasks, such as isolated letters [35] or digits recognition [45], [8]. Only two of the A/V corpora published in literature (including English, French, German and Japanese) contain both a large vocabulary and a significant number of subjects. The first is IBM's private, 290-subject, large-vocabulary AV-ViaVoice database of approximately 50 hours in duration [20]. The second is the VidTIMIT database [49], which was recently made available by LDC. It consists of 43 subjects reciting 10 TIMIT sentences each, and has been used for multi-modal person verification [50]. In general, the freely available corpora are inadequate for evaluating large-vocabulary, speaker-independent AVSR algorithms.

Furthermore, since most databases were recorded in carefully controlled conditions, they are not suitable for evaluating the robustness of a visual feature set with respect to image noise, lighting, and pose variation. One exception is the CUAVE corpus recently collected at Clemson University. It consists of 36 English speakers who were asked to speak digits while shifting their body position and head pose [43]. Also, researchers at IBM published the results of benchmarking of their current AVCSR system on two challenging visual corpora: one recorded in office conditions and the other in a moving vehicle. The results show that, while the visual-only WER doubled for the first and tripled for the second corpus [46], the visual modality remained beneficial to ASR, at least in the case of connected digits.

In Appendix A, we describe a multi-speaker continuous-speech audio-visual corpus that we have collected to facilitate this work.

2.3 Feature-based Automatic Speech Recognition

In the following section, we present a brief overview of three representative research articles that have applied the distinctive feature approach to statistical speech recognition in the audio domain. The feature-based systems described in these articles show improvements in noise robustness, language portability, coarticulation modeling and pronunciation modeling. To the best of the author's knowledge, no articulatory-feature visual speech recognition systems have been reported in the literature.

In [27], a hybrid HMM/ANN word recognition system was created using a set of five features, with each feature having anywhere from three to ten values. A separate neural network was trained to classify each distinctive feature. The outputs of the feature ANNs were then used to train another network, which learned the mapping from distinctive features to phonemes. This distinctive-feature system had similar performance to a baseline acoustic HMM/ANN system. When the baseline and the distinctive-feature systems were combined at the phoneme level by multiplying the neural-network outputs, the resulting system achieved significantly better word recognition rates across a range of noise levels.

In [28], a distinctive-feature HMM system was developed for consonant recognition in English, German, Italian and Dutch. Kohonen networks were used to classify three distinctive features (Place, Manner and Voicing) into a total of fourteen values, using standard cepstral coefficients as input vectors. The outputs of the Kohonen networks were then used as inputs to the HMM. The system significantly outperformed a baseline HMM system on infrequently occurring consonants, especially language-specific consonants.

In [12] and [13], a flexible coarticulation model was proposed based on overlapping articulatory features. This system used five features related to articulator positions: lips, tongue blade, tongue dorsum, velum and larynx. Five values were used for lip positions, seven for the tongue blade, twenty for the tongue dorsum, two for the velum and three for the larynx. Articulatory features were used as an intermediate stage between the acoustic signal and the phonetic representation; separate HMM states were used to model possible feature combinations. Although the feature models were context-independent, changes in the feature values were not required to synchronize at the phoneme boundaries, allowing the system to model coarticulation. Evaluation on the TIMIT corpus resulted in superior performance compared to a context-independent baseline HMM, and similar performance compared to a context-dependent HMM.

In [30], a flexible feature-based pronunciation model was developed using dynamic Bayesian networks. The system explicitly models the evolution of several streams of

linguistic features: degree of lip opening, tongue tip location and degree of opening, tongue body location and degree of opening, velum state, and voicing state. Changes in pronunciation were accounted for by allowing features to desynchronize and change values, as opposed to the standard approach of allowing phone substitutions, insertions, and deletions. The following synchrony constraints were imposed: i) all tongue features are synchronized; ii) the lips can desynchronize from the tongue; iii) the glottis and velum are synchronized; and iv) the glottis and velum can desynchronize from from the tongue and the lips. A pilot study using transcriptions of the Switchboard corpus manually converted to feature values showed an improvement over a baseline system that employed an extensive set of phonological pronunciation rules.

Chapter 3

AF-based Visual Speech Recognizer

In this chapter, we describe our proposed design of a feature-based visual speech recognition system. The first step in building the full recognition system is to create classifiers for the articulatory features. We describe how features are detected in Section 3.1. Then, in section 3.2, we describe how the outputs of these detectors are integrated over the entire length of the input sequence to produce the final word hypothesis.

3.1 AF Classification

The problem of classifying articulatory features from an input image can be cast as that of supervised learning. We assume that we are given a set of training examples, containing pairs of observation vectors derived from images of mouths and the corresponding articulatory feature labels. Since there are multiple AFs, each image will have several discrete labels, one for each AF. We seek a function that will map novel image observations to AF labels. The problem is similar to that of classifying visemes from input images, except that, instead of only one category of viseme labels, there are multiple categories of AF labels. Various machine learning techniques exist for learning such functions from training data, including artificial neural networks, boost-

Table 3.1: Comparison of classification rates on a single-speaker viseme classification task achieved by two classifier architectures: an SVM with an RBF kernel and a Gaussian classifier with diagonal covariance. The viseme set used is the same as the one shown in Table 4.2. N is the dimension of the data vectors. SVM-S and Gauss-S used static single-frame observations, while SVM-D and Gauss-D used dynamic three-frame observations.

N	SVM-S	SVM-D	Gauss-S	Gauss-D
5	32%	33%	27%	28%
10	34%	34%	28%	33%
50	39%	37%	37%	37%
75	39%	41%	36%	37%
100	35%	41%	35%	35%

ing, etc. In the context of frame-level classification for speech recognition, Gaussian classifiers or Gaussian Mixture Models (GMMs) are popular choices.

In recent years, the success of the Support Vector Machine (SVM) in various pattern recognition applications, including object recognition from images, has lead to its increased use for both binary and multi-class classification tasks. SVMs are powerful linear learning machines capable of finding the optimal separating hyperplane between classes in sparse high-dimensional spaces and with relatively few training examples. In preliminary experiments, we have found that SVMs outperform Gaussian classifiers on the task of viseme classification for a single speaker. Table 3.1 shows that the Gaussian classifier obtains its highest classification rate of 37% on 50-dimensional static observation vectors, while the SVM achieves a peak rate of 41% on 75-dimensional dynamic observations. Based on their superior performance, we have chosen to employ SVMs as the classification technique in the articulatory feature recognizer.

In the rest of this section, we will provide an overview of support vector machines.

3.1.1 Support Vector Machines

Support vector machines employ a learning strategy that simultaneously optimizes the empirical error and the complexity of the classifier. In the following, we assume

that the training set contains instance-label pairs (x_i, y_i), $i = 1, ..., l$, where $x_i \in R^n$ and $y \in \{1, -1\}^l$. Since SVMs are linear learning machines, they use a hypothesis space of linear decision functions of the form

$$f(x) = w^T x + b. \tag{3.1}$$

$f(x)$ is a real-valued function, therefore to classify a novel sample its output is converted to either a positive or negative label using the sign function $sgn()$. Geometrically, $f(x)$ divides the input space $X \subseteq R^n$ into two parts using a *hyperplane*, or an $n-1$ dimensional affine subspace, defined by the equation $w^T x + b = 0$. Points that fall on one side of the boundary are labeled as the positive class, and points that fall on the other side are labeled as the negative class. Although the resulting classifier is binary, it can be extended to handle the multi-class case. In this work, we use the "one-against-one" multi-class method, which combines the decisions of binary classifiers trained on each pair of classes using a simple voting technique [6].

An important property of the above classifier is that the hypothesis can be expressed in the *dual form*, or as a linear combination of the training points:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i x_i^T x + b. \tag{3.2}$$

The goal of the SVM algorithm is to find the *maximal margin hyperplane*, i.e. the hyperplane that maximizes its distance to each training point. This distance, or the *margin* of a point (x_i, y_i) with respect to the hyperplane (w, b), is the quantity $\mu_i = y_i(w^T x_i + b)$. A positive margin means the point lies on the correct side of the hyperplane and was thus classified correctly.

In the case where the data set is not linearly separable, a margin *slack variable* of an example (x_i, y_i) with respect to the hyperplane (w, b) and target margin μ is defined as $\xi_i = max(0, \mu - y_i(w^T x + b))$. Intuitively, this quantity measures by how much a point fails to have a margin of μ.

Linear machines have limited computational power, however, SVMs can overcome

this limitation by projecting the input data into a high-dimensional space and using a linear separating boundary in that space. For example, the transformation $\phi(\mathbf{x}) = (\phi_\mathbf{1}(\mathbf{x}), ..., \phi_\mathbf{N}(\mathbf{x}))$ maps sample points from the input space X to a higher-dimensional feature space F. With this mapping, the hypothesis can be written as

$$\sum_{i=1}^{l} \alpha_i y_i \phi(x_i)^T \phi(x) + b. \tag{3.3}$$

SVMs can operate in this high dimensional feature space without increasing the number of free parameters because the projection is done implicitly. Since the training examples only appear in (3.3) as inner products, the mapping can be performed by replacing the inner product with a kernel function of the original inputs. A *kernel* is a function K such that $K(x_i, x_j) = \phi(x_i)^T \phi(x)$. Kernels can be derived by choosing functions that satisfy certain mathematical properties. In practice, a common approach is to use one of several well-known kernels, such as

- the linear kernel

$$K(x_i, x_j) = x_i^T x_j, \tag{3.4}$$

- the polynomial kernel

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0, \tag{3.5}$$

- and the Radial Basis Function (RBF) kernel

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \quad \gamma > 0, \tag{3.6}$$

Finally, a support vector machine is formally defined as the solution to the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i, \\ \textit{subject to} \quad & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \end{aligned} \tag{3.7}$$

The parameter C controls the trade-off between minimizing the empirical error and the size of the hyperplane.

A more in-depth discussion of the SVM algorithm can be found in [57]. The SVM implementation used in this thesis is the LIBSVM [6] library, which is freely available online.

3.2 Word Recognition

In the previous section, we proposed to use an SVM classifier to assign articulatory feature labels to visual data observed at a particular moment in time. If the goal were phoneme recognition, the feature labels could be combined to obtain phonemic labels on a per-frame basis. However, although viseme recognition is an important problem, our ultimate goal is recognizing words and sentences. In this section, we describe our proposed approach to feature-based visual word recognition.

While SVMs are inherently static classifiers, speech is a dynamic process. Therefore, we need an additional statistical model to describe the evolution of articulatory features over time. To date, the most successful model for to speech recognition is the Hidden Markov Model (HMM.) The HMM combines acoustic, pronunciation and language modeling into a single framework. It describes the underlying process generating the observations as a single stream of hidden variables. The hidden state that we are interested in is articulatory gestures, such as the motions of the lips or tongue. Therefore, if we used an HMM, we could not model the independence of the articulators explicitly, as each hidden state would contain a specific combination of values of each feature. Allowing the feature streams to evolve independently is particularly important when combining audio and visual modalities. For example, the vibration of the vocal cords occurs independently from other gestures.

An alternative representation suggested in [31] uses Dynamic Bayesian Networks, which are a superset of HMMs in terms of modeling power. DBNs are capable of representing the evolution of several hidden streams of variables over time. This results in a more efficient representation than a single stream with a large state space.

Also, DBNs make it possible to allow different variables to evolve asynchronously over time. Next, we will provide a brief summary of DBNs and their applications to speech recognition, followed by an overall description of our system.

3.2.1 Dynamic Bayesian Networks

A Bayesian Network (BN) is a directed acyclic graph whose nodes correspond to random variables, $X_1, ..., X_n$, and whose edges point from parent to child nodes. Missing edges between nodes represent the conditional independence of the corresponding variables. The joint distribution for the graph is thus simplified to

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}), \tag{3.8}$$

where X_{π_i} are the parents of variable X_i. A Dynamic BN has a repeating structure of groups of nodes, with edges between the groups pointing in the direction of increasing time or space. DBNs are particularly useful for modeling dynamic processes such as speech.

Figure 3-1 shows the DBN structure commonly used in speech recognition applications [2]. This particular structure exhibits the explicit graphical representation approach, where the details of the book-keeping associated with speech recognition are represented explicitly in the graph. In comparison, an implicit representation, such as the one used by traditional HMMs, hides these details in the implementation, or in an expanded hidden state space [2]. In the graph, dashed edges represent true random dependencies, while solid edges represent deterministic dependencies. Colored nodes correspond to the observed variables, while the rest of the variables are hidden. The variables *Word* and *Word-Position* indicate the current word and the position within that word; together they determine the current value of *Phone*. Transitions between phones inside the word and words inside the utterance are also modeled explicitly by the *Word-Transition* and *Phone-Transition* variables. The *End-of-Utterance* variable enforces the constraint that the interpretation must end at the end of a word by having a probability distribution such that the observed value of 1 is only possible if

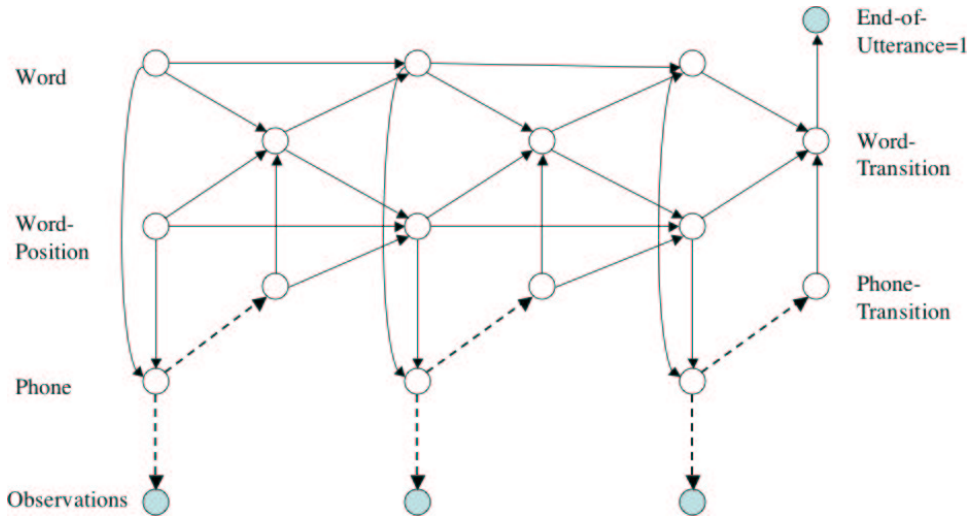


Figure 3-1: An explicit DBN structure for speech recognition.

the last frame has a word-transition value of 1.

The above phone-based model can be extended to include one hidden variable for each of the articulatory features, as shown in Figure 3-2. For simplicity, the book-keeping variables are omitted here. Each of the two frames shown in the figure contains hidden features $F_1, F_2, ..., F_N$. The features depend on both the current phone state P and their values in the previous frame. The visual observation variable O is conditioned only on the feature variables. An intuitive interpretation of this structure is that, while the articulators aim to reach their target positions for each phone, their actual state at a particular instant in time is influenced by continuity constraints and articulatory inertia.

3.3 System Architecture

The feature-based visual speech recognizer proposed in this thesis uses a lexical access model similar to the DBN structure used for audio feature-based pronunciation modeling in [30]. The main difference is that our task is automatic lipreading and, therefore, we use a smaller set of features. We also use a hybrid architecture, where the posterior probabilities of each observed articulatory feature are obtained from

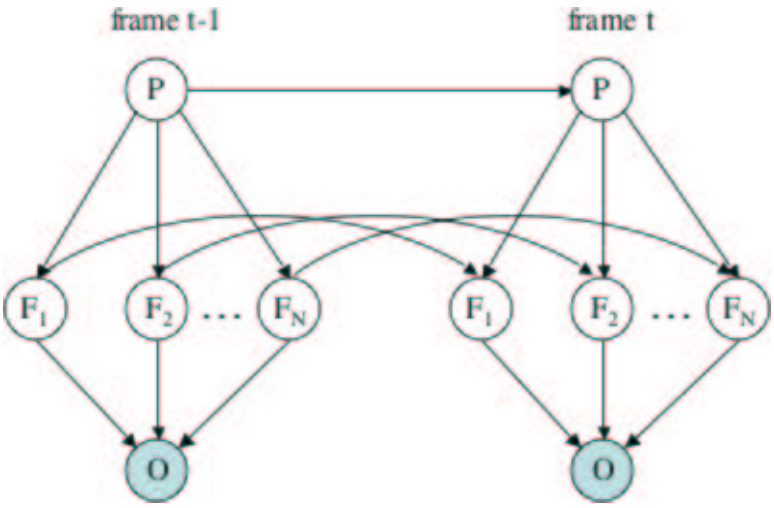


Figure 3-2: An AF-based DBN model.

SVM classifier outputs. The algorithm for converting SVM outputs to probabilities is described in [6].

Figure 3-3 shows one frame of the DBN model. We use three features: LIP-LOC (LL), LIP-OPEN (LO) and LAB-DENT (LD). The variables are:

- *Word* - the lexicon entry corresponding to the current word.
- *LL-Pos* - the position of the LIP-LOC feature in the underlying pronunciation. This variable has value 0 in the first frame, and in subsequent frames is conditioned on $Word_{t-1}$, $LL\text{-}Pos_{t-1}$ and $Word\text{-}Trans_{t-1}$.
- *LL* - the underlying value of the LIP-LOC feature. Its distribution is determined by the specific table defined for the current word.
- *LL-Obs* - the observed surface value of the LIP-LOC feature. $p(LL\text{-}Obs|LL)$ encodes the allowed feature substitutions.
- *Word-Trans* - a binary variable that indicates the last frame of a word.
- *LL-LO-Sync* - a binary variable that enforces a synchrony constraint between the LIP-LOC and LIP-OPEN variables. It is observed with value 1; its distribution is constructed such as to force its parent *LL-Pos* and *LO-Pos* variables obey the desired constraint.

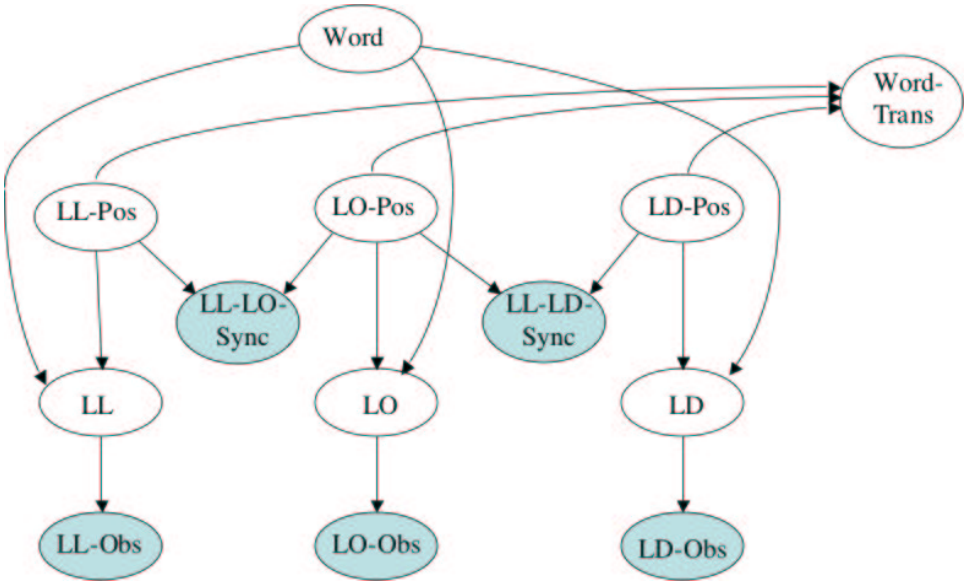


Figure 3-3: One frame of a DBN for a feature-based pronunciation model.

The rest of the LIP-OPEN and LAB-DENT variables are defined in a similar manner. Pronunciation variation and context effects are modeled by allowing the observed feature values to stray from the targets specified in the lexicon entry. This can happen in two ways: due to *substitution*, in which an articulator fails to reach its target; and *asynchrony*, in which different articulators proceed through their sequences of targets at different rates. Feature asynchrony can occur due to coarticulation. For example, if the underlying pronunciation of the word has a fricative followed by a rounded vowel, the lips may reach the protruded position before or during the labiodental gesture. We provide examples of both feature asynchrony and substitution in Chapter 5.

Although the DBN described above uses the three features LIP-LOC, LIP-OPEN and LAB-DENT, in general, any number and type of features can be used. In the next chapter, we will lay out the motivation behind selecting this particular feature set, and describe the features in more detail. Then, we will present several experiments using the proposed approach in Chapter 5.

Chapter 4

Articulatory Feature Design

In the previous chapter, we proposed a set of algorithms for performing word recognition using multiple streams of articulatory features extracted from the visual component of the speech signal. However, before we can build such a recognizer, we must answer the following fundamental questions:

- what is the optimal set of features?
- what values should each feature have?
- where do the feature labels come from, i.e. should they be based on acoustic data, visual data, or both?

In general, the set of features used in the development of a speech recognizer and the set of values they take on depends on the specific research goals. According to one review, about 40 different feature combinations have been used in various research studies [11]. An example feature set proposed in [31] is shown in Table 4.1.

Ideally, the chosen features should be easily distinguishable from the visual signal. Furthermore, features should exhibit themselves in a consistent manner not only in isolated phoneme examples, but also in continuous, co-articulated and spontaneous speech. Moreover, the same feature set should apply to multiple speakers.

One way of gaining some insight into what features might occur naturally is by performing a cluster analysis of visual speech data. Another approach to designing

a visual feature set is simply to choose one of the existing linguistically motivated feature sets, and agree upon that subset of it which applies to the visible portion of the vocal tract. However, it is not obvious which part of the articulatory process is captured by the camera, and which is not. For example, while it can be said with certainty that features such as nasality and voicing cannot be distinguished visually, the case is not as clear with frication or retroflection. Even if we restrict ourselves to features that specifically describe articulator positions, such as the ones in Table 4.1, we can still only eliminate the last two features as strictly non-visual. While we suspect that a subset of the tongue-related feature values can also be eliminated, it is not immediately clear what that subset should be. One approach would be to obtain labels for every possible feature, and train a machine learning algorithm on these labels. Then, we can keep only the features that the algorithm is able to classify reasonably well.

Once we have decided on a suitable set of articulator states to represent in our model, the next issue is how to obtain the labels for training the SVM classifier. To record the ground truth about articulatory gestures, subjects would either have to wear sensors in their mouth, or have their vocal tracts scanned using x-ray technology - both expensive and impractical solutions. Alternatively, feature labels can be derived by using phonetic audio transcriptions as the ground truth labels for the video. However, this assumes that the audio labels retain the complete information about the visible articulators. This is certainly true in some cases: for example, if the audio phoneme label is /uw/, it must correspond to the visible articulatory action of rounding the lips. However, due to coarticulation, a visible articulator can move into position before the target phone is produced. Furthermore, while this is happening, another, invisible, articulator can produce a phone, resulting in an acoustic label not related to the action of the first articulator. Therefore, it may not be always possible to tell from the acoustic labels precisely when a certain visible articulatory action begins and ends.

In the following sections, we set out to find the answers to some of the questions outlined above, with the ultimate goal of designing a feature set for use in the proposed

Table 4.1: Articulatory feature set proposed in [30].

Index	Feature Name	Values
0	LIP-LOC	protruded, labial, dental
1	LIP-OPEN	closed, critical, narrow, wide
2	TT-LOC	dental, alveolar, palato-alveolar, retroflex
3	TT-OPEN	closed, critical, narrow, mid-narrow, mid, wide
4	TB-LOC	palatal, velar, uvular, pharyngeal
5	TB-OPEN	closed, critical, narrow, mid-narrow, mid
6	VEL	closed, open
7	GLOT	closed, critical, wide

recognizer. Section 4.1 introduces the most commonly used visual speech unit - the viseme - and points out potential problems with viseme-based visual speech modeling. Section 4.2 describes the visual data representation used in the subsequent experiments. Section 4.3 analyzes the separability of phonetically labeled visual speech units into distinct feature classes using supervised clustering. Section 4.4 departs from acoustically-derived labels and performs unsupervised clustering of visual data in order to determine possible articulatory features. Finally, section 4.5 investigates the trade-offs between acoustically derived feature transcriptions and manual labeling of articulatory states from visual data.

4.1 The Existing Approach to Visual Unit Modeling

Traditionally, speech recognizers model speech as a sequence of basic units that are contiguous in time. These units can be derived using either linguistic knowledge or a statistical, data driven approach. Words, syllables and phonemes are examples of linguistically-derived units. HMM states are an example of statistically-derived units. In general, longer units such as words represent contextual variations more accurately than shorter units. It is possible to use words as the speech unit for small-vocabulary tasks, for example, digit recognition. However, in the case of general-

Table 4.2: An example viseme to phoneme mapping, using the TIMIT phone set.

Viseme	Phonemes
1	/ax/, /ih/, /iy/, /dx/
2	/ah/, /aa/
3	/ae/, /eh/, /ay/, /ey/, /hh/
4	/aw/, /uh/, /uw/, /ow/, /ao/, /w/, /oy/
5	/el/, /l/
6	/er/, /axr/, /r/
7	/y/
8	/b/, /p/
9	/bcl/, /pcl/, /m/, /em/
10	/s/, /z/, /epi/, /tcl/, /dcl/, /n/, /en/
11	/ch/, /jh/, /sh/, /zh/
12	/t/, /d/, /th/, /dh/, /g/, /k/
13	/f/, /v/
14	/gcl/, /kcl/, /ng/

purpose, large-vocabulary recognition, there is not enough data to train a separate model for each word. On the other hand, smaller units such as phonemes and syllables are vocabulary-independent and limited in number: there are only about 50 phonemes in the English language. Although context-independent phonemes generalize well, they are insufficient to capture the different realizations of a phoneme due to its surrounding context, i.e. its allophones. Therefore, in practice, context-dependent phonemes, such as biphones or triphones, are used to improve recognition accuracy.

A *phoneme* is defined as the minimal unit of speech sound that can distinguish one word from another. The term *phone* is generally used to denote a phoneme's acoustic realization. In order to model visual speech, researchers have defined the *viseme* to be the visual equivalent of a phoneme. Since not all phonemes are visually distinguishable (e.g. "mat" vs "pat"), several phonemes are usually mapped to one viseme. An example viseme-to-phoneme mapping is shown in Table 4.2, where the phoneme labels used are from the TIMIT phoneme set. With the exception of small vocabulary systems, where whole words are used as speech units, (context-dependent) visemes are the unit of choice for most visual speech recognition and synthesis applications.

Table 4.3: A commonly used mapping of consonants to visemes [17].

Viseme	Consonant phonemes
1	/f/, /v/
2	/th/, /dt/
3	/s/, /z/
4	/sh/, /zh/
5	/p/, /b/, /m/
6	/w/
7	/r/
8	/g/, /k/, /n/, /t/, /d/, /y/
9	/l/

Table 4.4: The 44 phoneme to 13 viseme mapping considered in [39], using the HTK phone set.

Viseme	Phonemes
Silence	/sil/, /sp/
Lip-rounding based vowels	/ao/, /ah/,/aa/, /er/, /oy/, /aw/, /hh/
	/uw/, /uh/, /ow/,
	/ae/, /eh/, /ey/, /ay/
	/ih/, /iy/, /ax/
Alveolar-semivowels	/l/, /el/, /r/, /y/
Alveolar-fricatives	/s/, /z/
Alveolar	/t/, /d/,/n/, /en/
Palato-alveolar	/sh/, /zh/, /ch/, /jh/
Bilabial	/p/, /b/,/m/
Dental	/th/, /dh/
Labio-dental	/f/, /v/
Velar	/ng/, /k/, /g/, /w/

A set of visemes is normally defined through one's knowledge of linguistics and the intuition of which phonemes might appear the same visually. For example, any pair of phonemes that differ only in the presence or absence of voicing are mapped to the same viseme (e.g. /t/ and /d/.) Place of articulation is another clue as to whether or not two phonemes belong to the same visual class (e.g. /m/ and /b/ are both bilabial.) However, in many cases, the mapping is not obvious. For instance, silence is very difficult to define as a single visual unit, because it is not tied to any particular configuration of the articulators. The phoneme /w/ is a less extreme example of an ambiguous mapping, with some researchers grouping it with /r/, some with /l/ and some putting it in a class of its own (see Tables 4.3 and 4.4 for examples.) In general, there is no agreement in the literature on a standard set of visemes, as there is in the case of phonemes. The mappings are somewhat ad-hoc and vary depending on the application.

Although visemes are the standard unit of recognition, we are interested in production-inspired visual units. One of the motivations for the existence of articulatory features is the fact that acoustic speech sounds form natural classes. In the following, we propose and evaluate an automatic method of defining natural visual speech classes. In the next two sections, we use a bottom-up agglomerative clustering technique in combination with a distance metric to come up with groupings of data. We also explore the effects of such factors as the signal representation, the image region, and the length of the time window on the resulting visual clusters. All experiments are conducted on the AVTIMIT corpus described in the appendix.

4.2 Visual Signal Representation

The visual speech signal consists of raw images of the face and must be represented as a lower dimensional set of features before further processing. In general, assuming the mouth and chin portion of the face has been located, the actual extracted *region of interest (ROI)* can vary in size and shape. The region can even be divided into several sub-windows with measurements extracted from each separately. In this work, we

experiment with three different sizes of rectangular regions. Once an N-by-M region of interest has been determined, it is normalized for lighting effects using histogram equalization. Then, a vector of measurements representing the region is extracted. This process is commonly referred to as *feature extraction*, and is described below. Finally, several consecutive frames are sometimes stacked together to obtain a more dynamic representation.

There are two main approaches to visual feature extraction for speech recognition. The first is an appearance-based, or bottom-up, approach, in which the raw image pixels are compressed, for example, using a linear transform, such as a discrete cosine transform (DCT), principle component analysis (PCA) projection, or a linear discriminant analysis (LDA) projection. The second is a model-based, or top-down, approach, in which a pre-determined model, such as the contour of the lips, is fitted to the data. Some approaches combine both appearance and model-based features. It has been found that, in general, bottom-up methods perform better than top-down methods, because the latter tend to be sensitive to model-fitting errors [39].

In this work, we use only appearance-based features. In particular, we experiment with two representations: raw images and DCT-compressed images. In the former, raw pixels are taken from the image. In the latter, the 16-by-16 subset of the 2-D DCT transform matrix containing the highest-frequency coefficients is used. In both cases, a PCA transform is applied and the top 32 coefficients retained to further reduce the dimensionality of the data vector.

The top 32 principal components of the raw pixel data are shown in Figure 4-1. In order to illustrate the contribution of each component, they are shown again in Figure 4-2, this time added to the mean mouth image. For example, the first principal component (accounting for the most variance) corresponds to the mouth being either more or less open. Of course, not all components describe variance due to speech-related movements.

We used 50 context-independent phonemes from our dataset, with a total of around 130,000 samples. Since visualizing the distribution of visual data corresponding to each phonemic label can help us analyze the results of clustering in the next

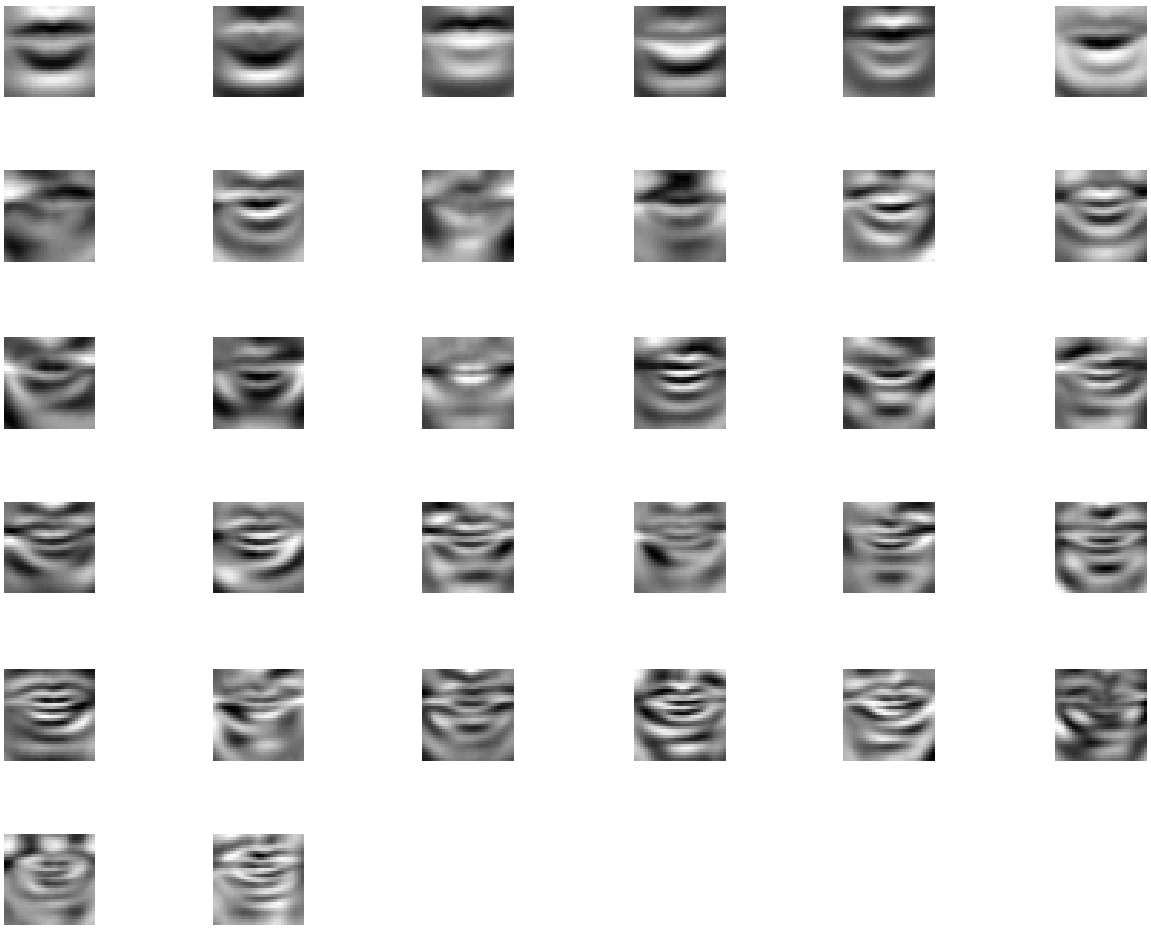


Figure 4-1: The first 32 principal components.

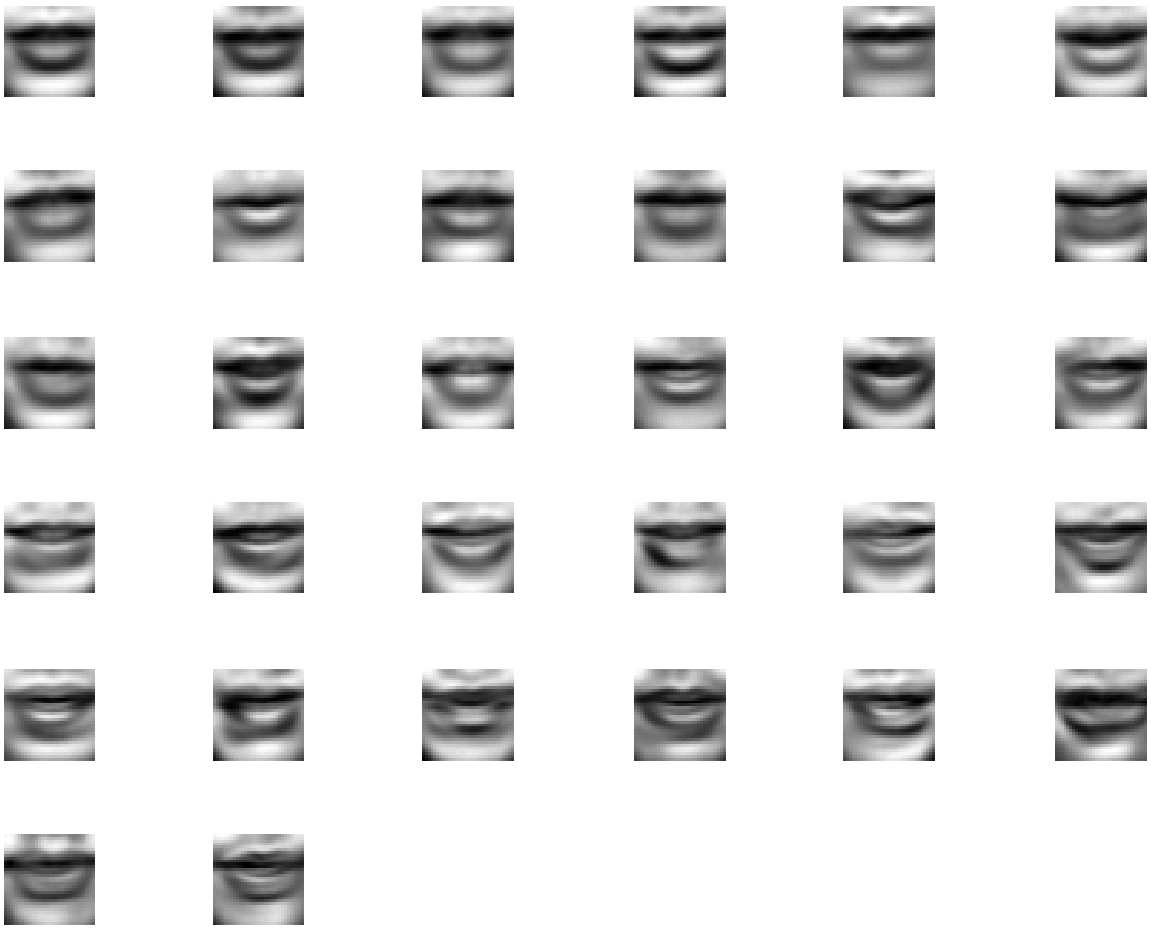


Figure 4-2: The first 32 principal components, added to the mean image.

Figure 4-3: Images reconstructed from the mean of 32-coefficient PCA vectors extracted from the middle frame of segments with the corresponding phonetic label.

section, we plot the mean of each distribution in Figure 4-3. The lip images are re-constructed from the means of the raw-pixel PCA coefficient distributions. Each distribution corresponds to the middle frame of one phoneme. However, it is difficult to see the differences between the images. Therefore, we also show the reconstruction of the means just from the 32 coefficients without the overall mean in Figure 4-4.

As for the DCT signal representation, we tried using two different ROIs: a smaller 16-by-32 ROI including just the lips, and a larger 32-by-32 ROI including the lips and the chin. The goal was to see whether varying the size of the region while keeping the dimensionality of the feature vector constant would influence the formation of clusters. The phoneme distributions for the smaller ROI are shown in Figure 4-5 and for the larger one in Figure 4-6. Once again, we do not add the mean vector to enable the reader to see the differences more clearly. Also note that the phoneme set is slightly different, with the diphthongs being split into two phonemes.

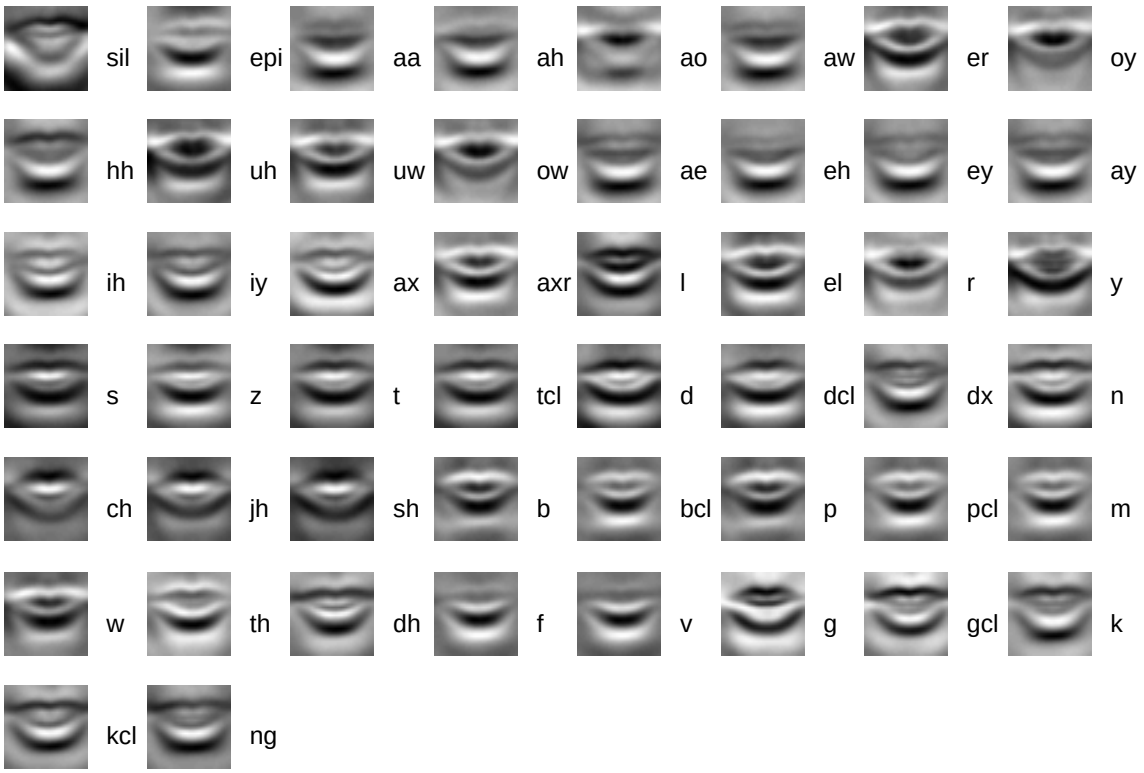


Figure 4-4: Mean phoneme images from 4-3, with the overall mean mouth image subtracted.

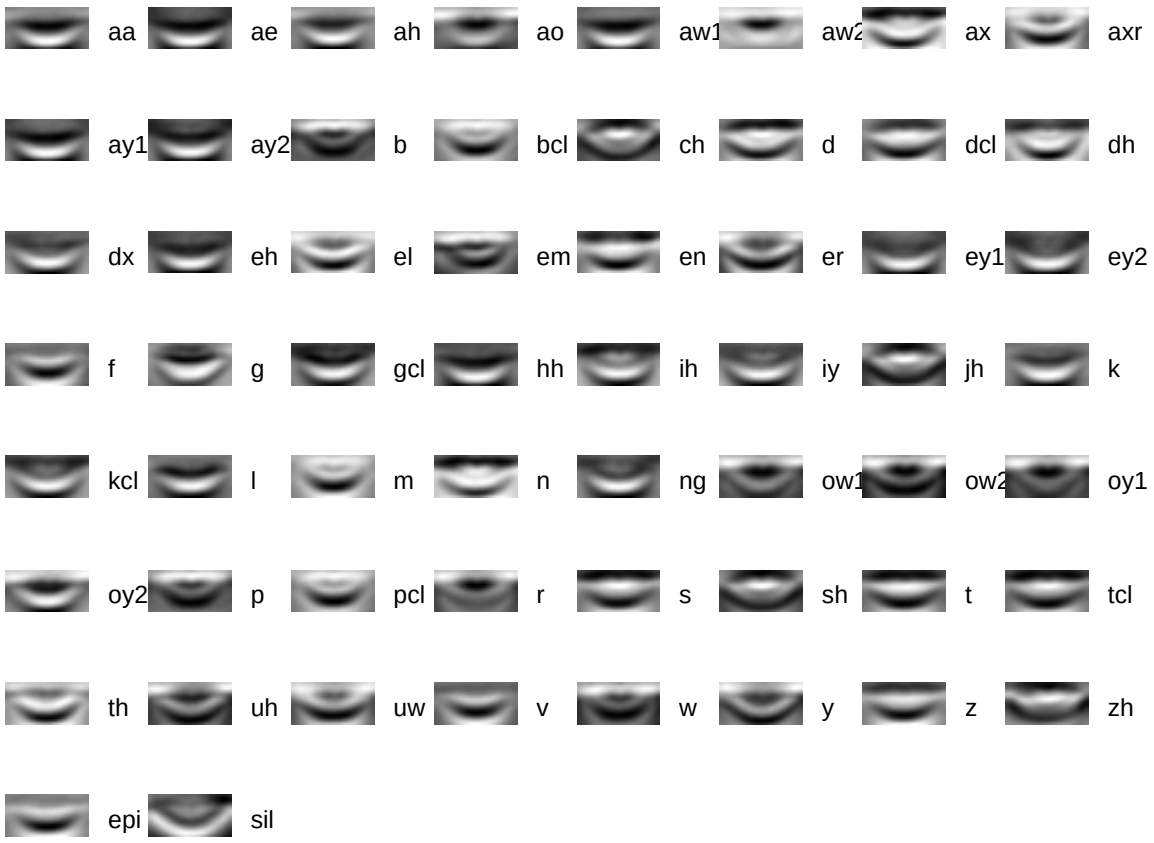


Figure 4-5: Images reconstructed from the mean 36 PCA coefficients, extracted from the 256 highest-frequency DCT coefficients.

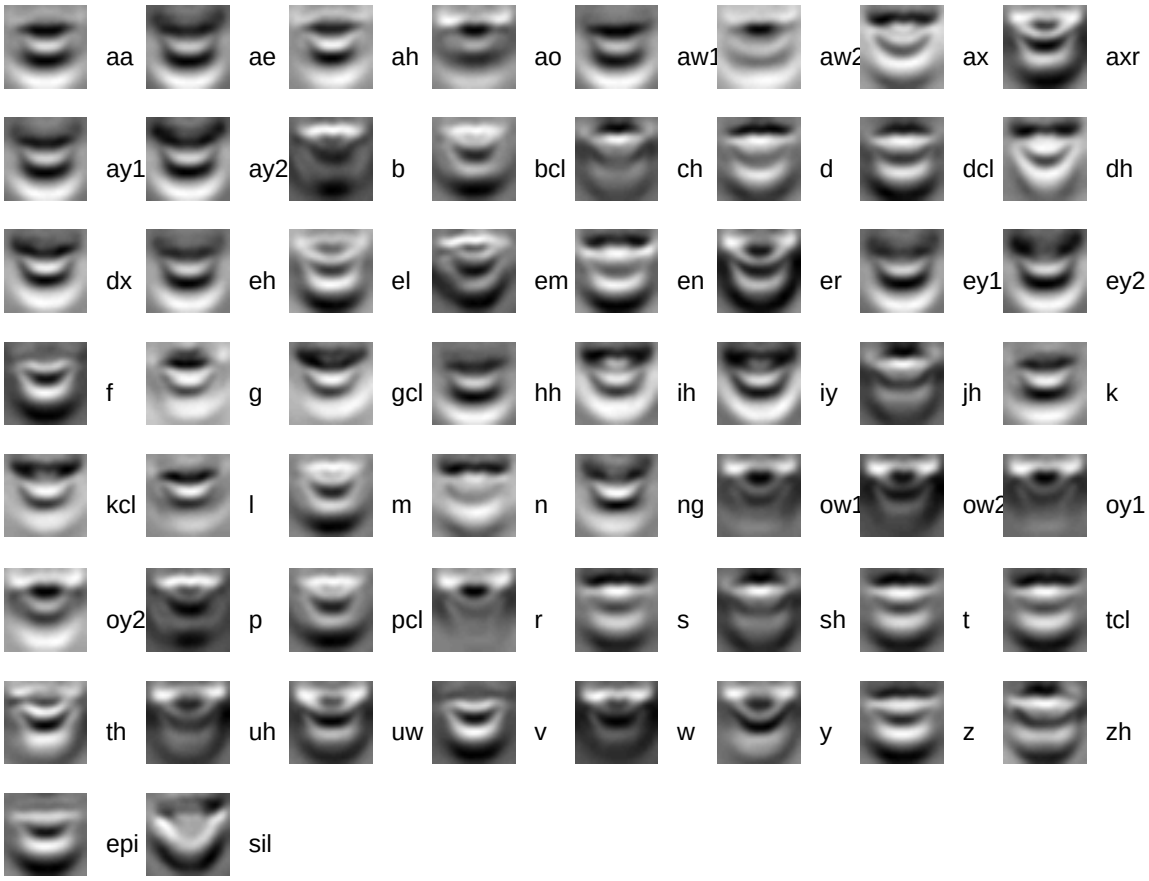


Figure 4-6: Image reconstructed from the mean 36 PCA coefficients, extracted from the 256 highest-frequency DCT coefficients.

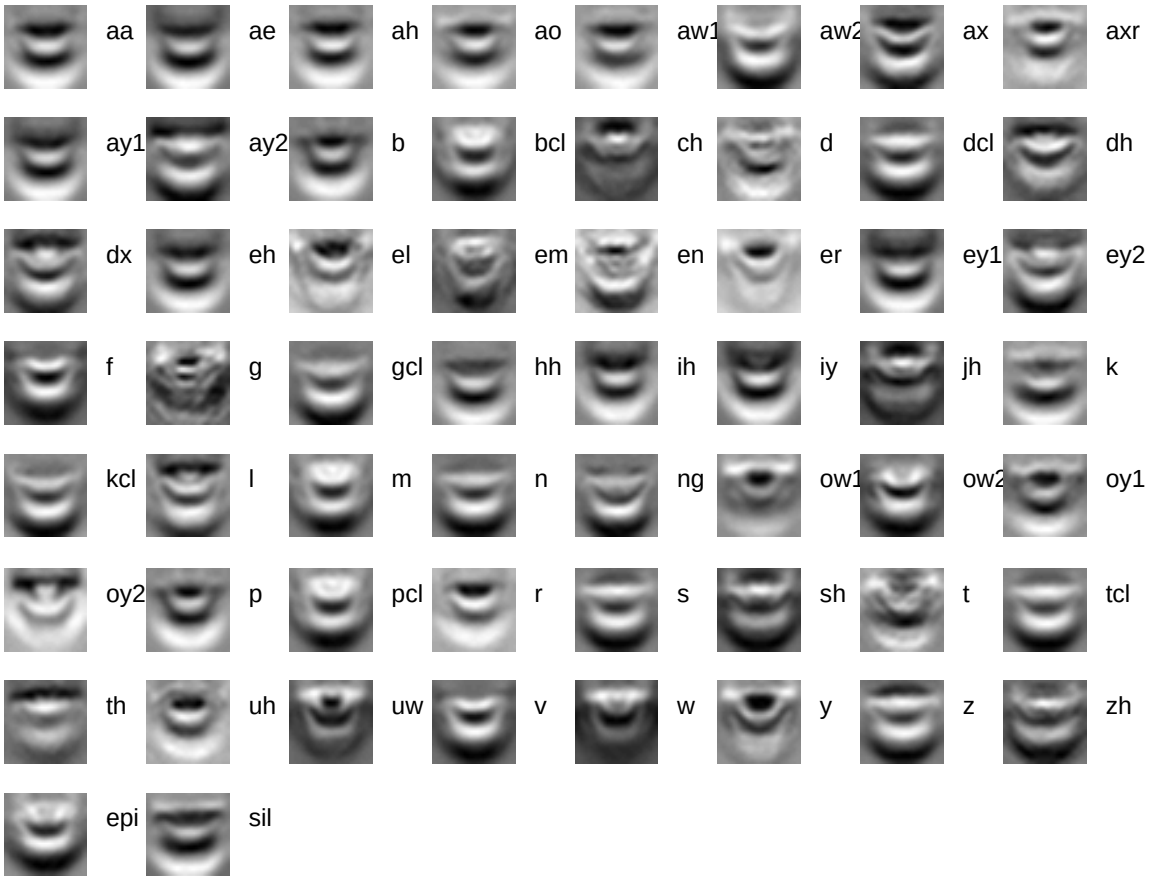


Figure 4-7: Images reconstructed from the mean 36 PCA coefficients, extracted from the 256 highest-frequency DCT coefficients taken from frame difference images.

Finally, we experiment with using differences between every other frame instead of static frames. Taking the difference removes some of the texture unique to the individual speaker, and also captures the dynamics of speech. In our case, the facial images were aligned prior to ROI extraction using correlation tracking of the bridge of the nose, to ensure that the motion between frames is mostly due to articulator movements. The corresponding distributions of the DCT data are shown in Figure 4-7. This time we did not remove the mean vector for presentation, since the frame differences are already quite visually distinct.

4.3 Clustering Using Phonetic Labels

One way to derive a set of articulatory feature units is to use a data-driven automatic clustering approach. This method has several advantages. First of all, since most phonetic recognizers use statistical models trained on data, it might be beneficial to automatically learn natural classes from the data. Another advantage is that, if a large amount of training samples is available, the data-driven algorithm can account for contextual variations and differences between speakers. This is particularly interesting because the knowledge-based mappings are usually made with canonical phonemes in mind, while recognition is done on continuous, co-articulated speech. Lastly, this approach enables us to explore the influence of the signal representation on the optimal visual units.

It is important to keep in mind that speech recognition consists of two closely related sub-problems: segmentation and classification. Although both are equally important, and neither has received enough attention in the context of visual speech recognition, in this work we will focus mainly on the latter. We believe that there is no simple one-to-one mapping between phonemic and visemic segment boundaries, however, for the time being, we will adopt the standard approach of using acoustically-derived phonemic boundaries to segment the visual signal into units. Therefore, we will be mainly concerned with grouping visual data synchronous with the acoustic segments into distinct, distinguishable classes.

In the following, we will use clustering to automatically discover speech classes from phonetically labeled visual data, using signal representations described in the previous section. First, we will describe the clustering algorithm.

4.3.1 Clustering Algorithm

We start with one cluster per phoneme, then use a standard agglomerative hierarchical clustering algorithm to successively merge clusters based on the maximum distance between them. We use the Battacharyya distance as the distance metric. The Battacharyya distance measures the similarity of two Gaussian distributions:

$$D_{bhat} = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \qquad (4.1)$$

where M_i is the mean vector and Σ_i is the covariance matrix of class ω_i, for i=1,2. The first term of Equation 4.1 gives the class separability due to the difference between class means, while the second term gives the class separability due to the difference between covariance matrices. The advantages of using the Bhattacharyya distance is that it is computationally simple and provides a "smoothed" distance between the two classes. The disadvantage is that it assumes that the data are normally distributed, which we do not believe to be the case. The clustering technique described above was used for phoneme clustering in [32].

4.3.2 Results and Discussion

We ran the algorithm on our data, using the following image encodings:

- Pixel-based PCA coefficients
  - single static frame, see results in Fig. 4-8
  - three consecutive static frames, see results in Fig. 4-9
- DCT-based PCA coefficients, large ROI
  - single static frame, see results in Fig. 4-10
  - single motion frame, see results in Fig. 4-11
  - three consecutive static frames, see results in Fig. 4-12
  - three consecutive motion frames, see results in Fig. 4-13
- DCT-based PCA coefficients, small ROI, Fig. 4-14

First, let us observe the general structure of the cluster trees. The bars are proportional in length to the distance between clusters. Overall, we see that many clusters at the lowest level are well-known visually confusable pairs, eg. {m,bcl},

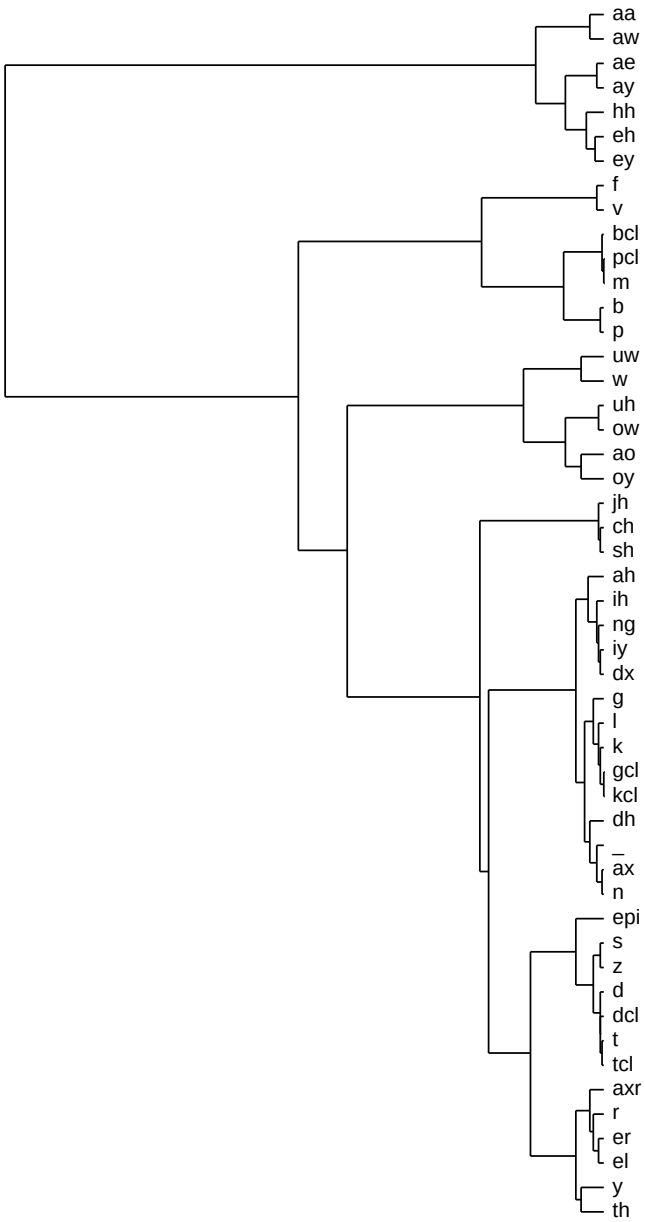


Figure 4-8: Cluster plot using Pixel PCA encoding of static frames.

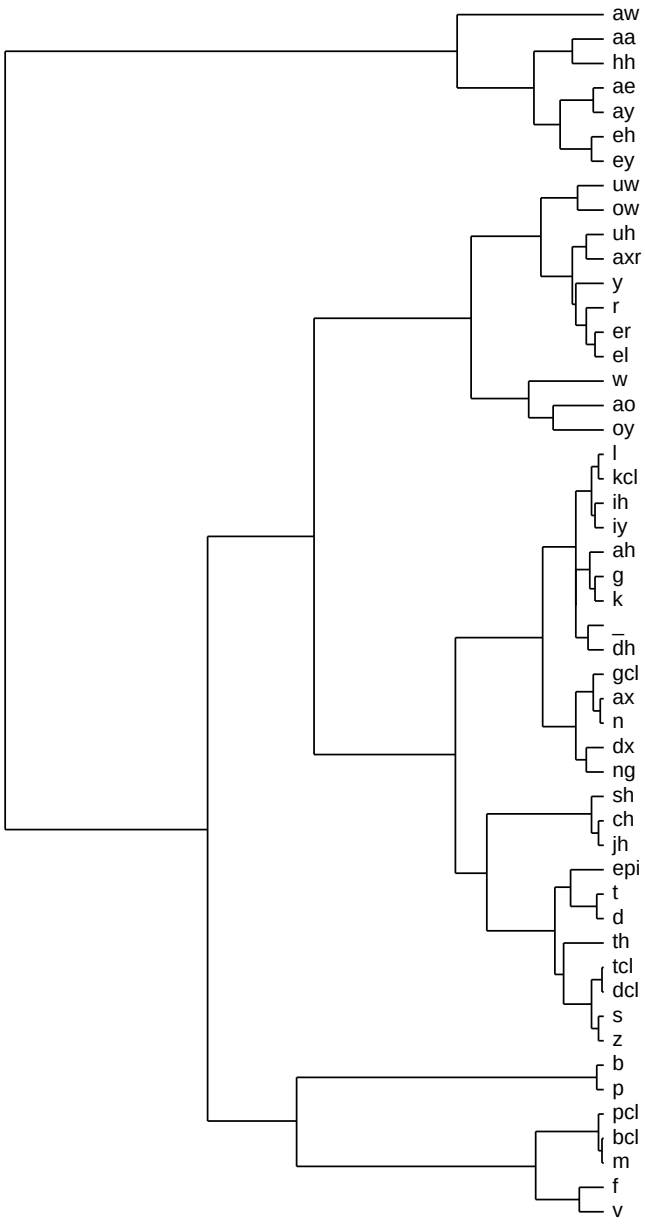


Figure 4-9: Cluster plot using a Pixel-PCA encoding of 3 stacked static frames.

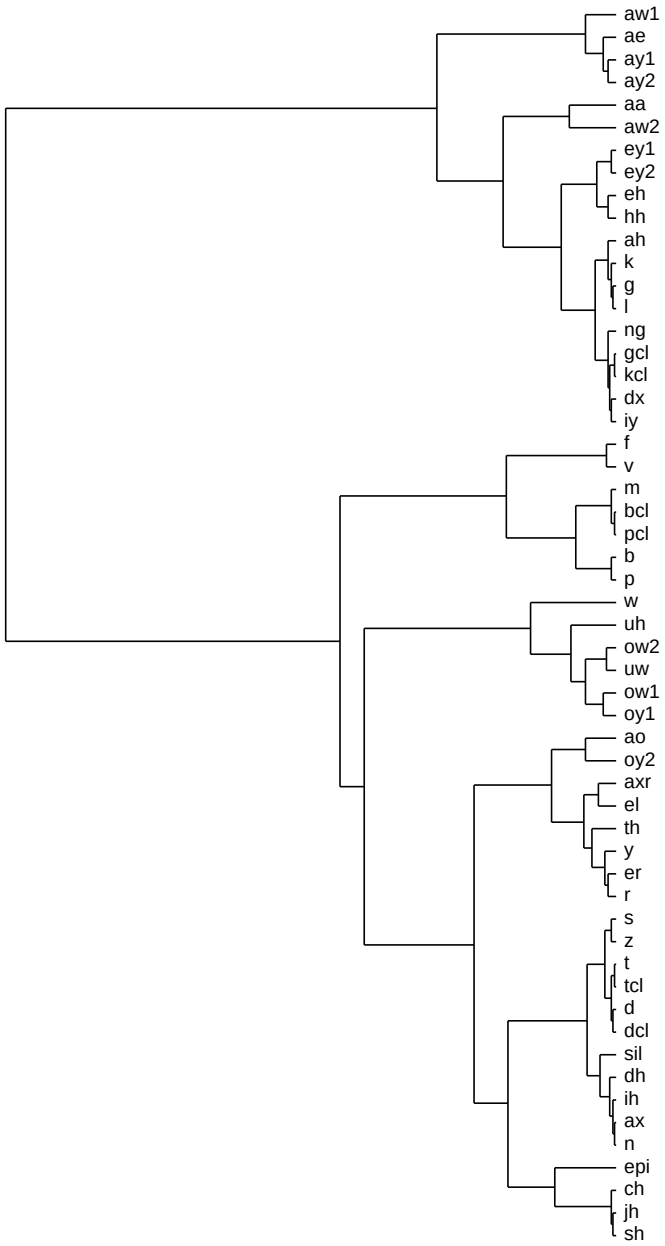


Figure 4-10: Cluster plot using a DCT-PCA encoding of static frames.

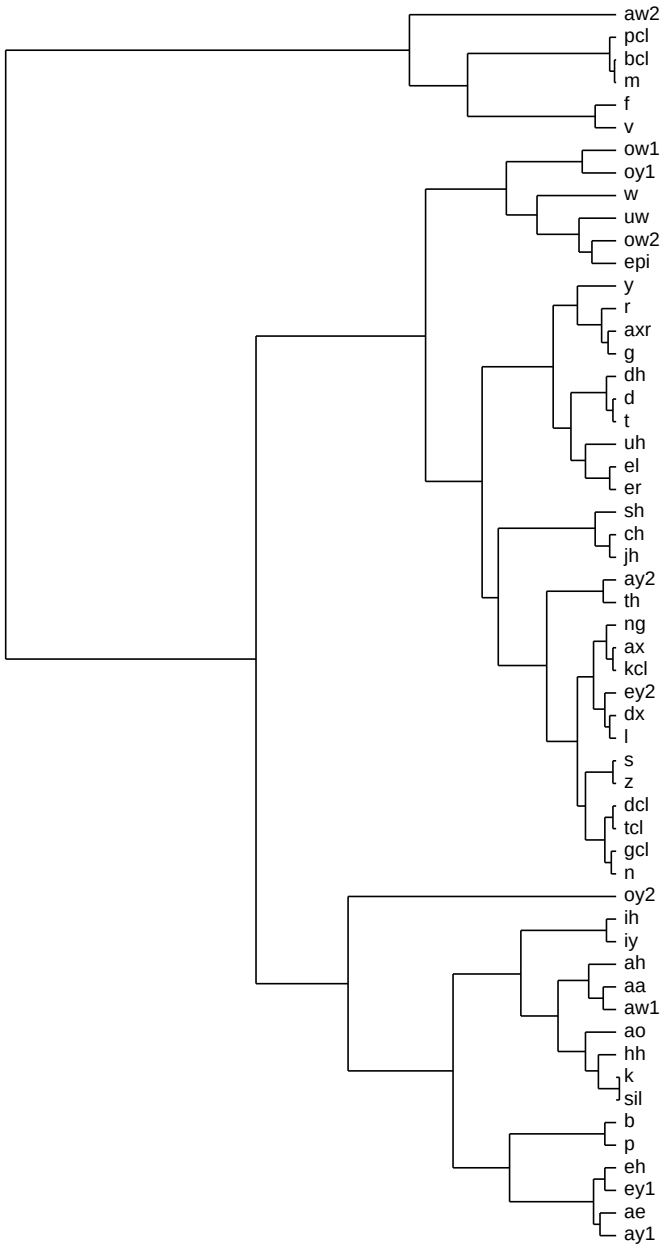


Figure 4-11: Cluster plot using a DCT-PCA encoding of motion frames.

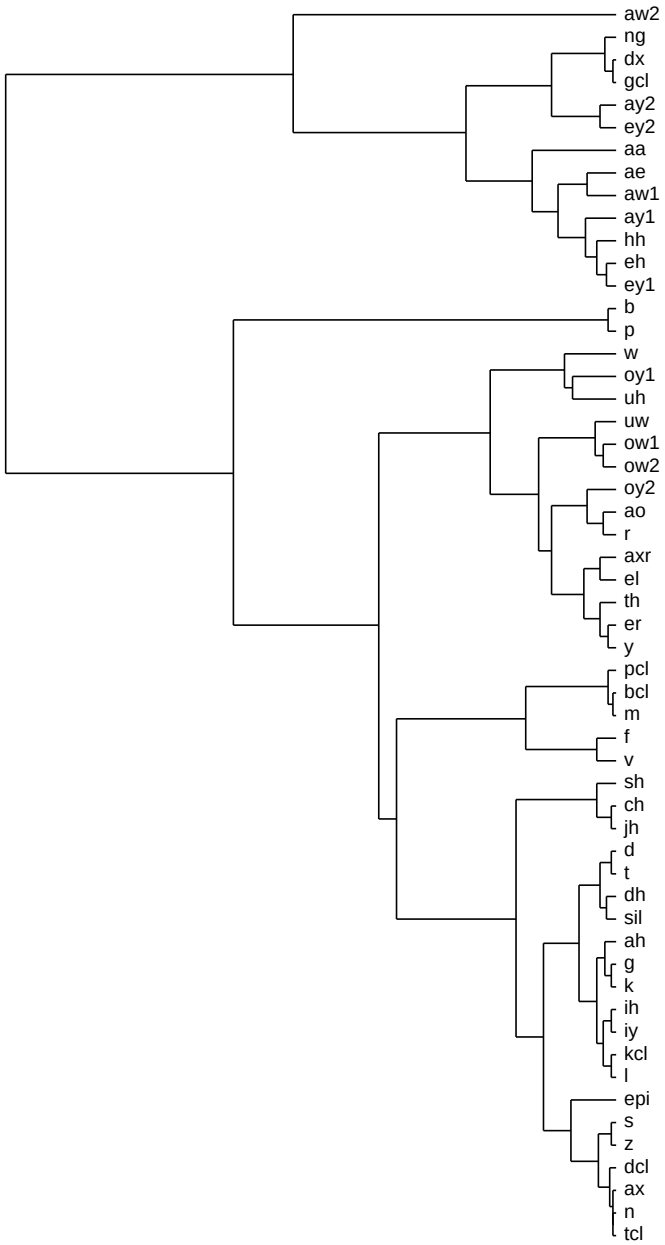


Figure 4-12: Cluster plot using a DCT-PCA encoding of 3 stacked static frames.

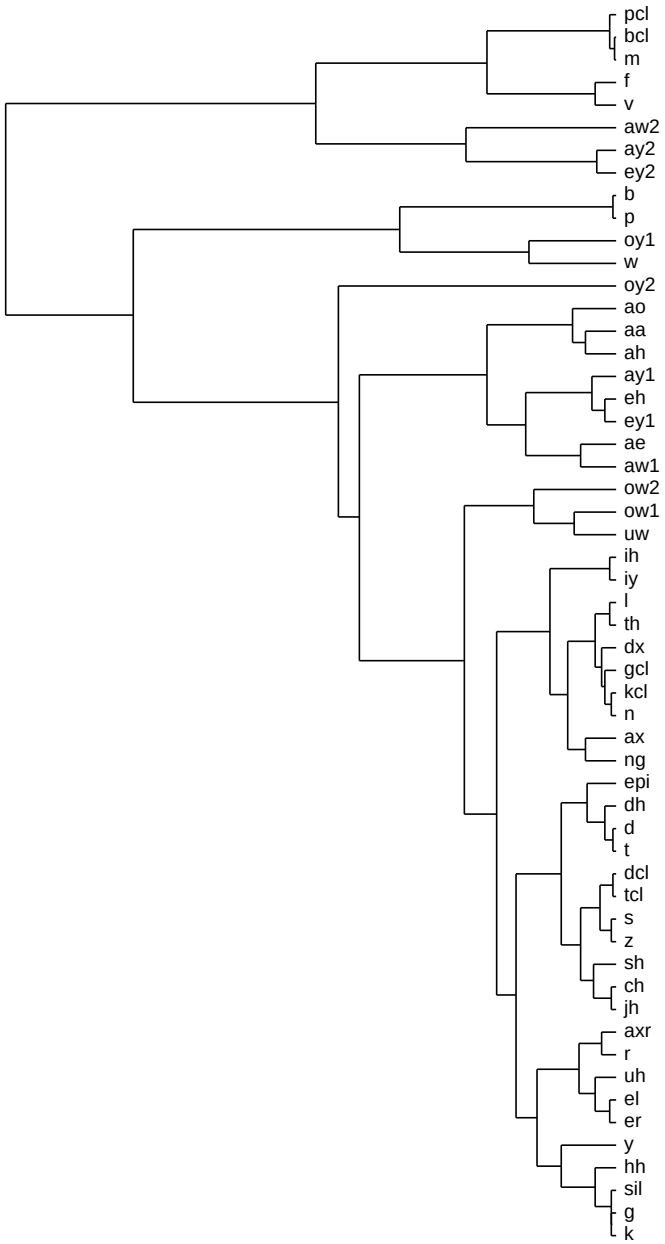


Figure 4-13: Cluster plot using a DCT-PCA encoding of 3 stacked motion frames.

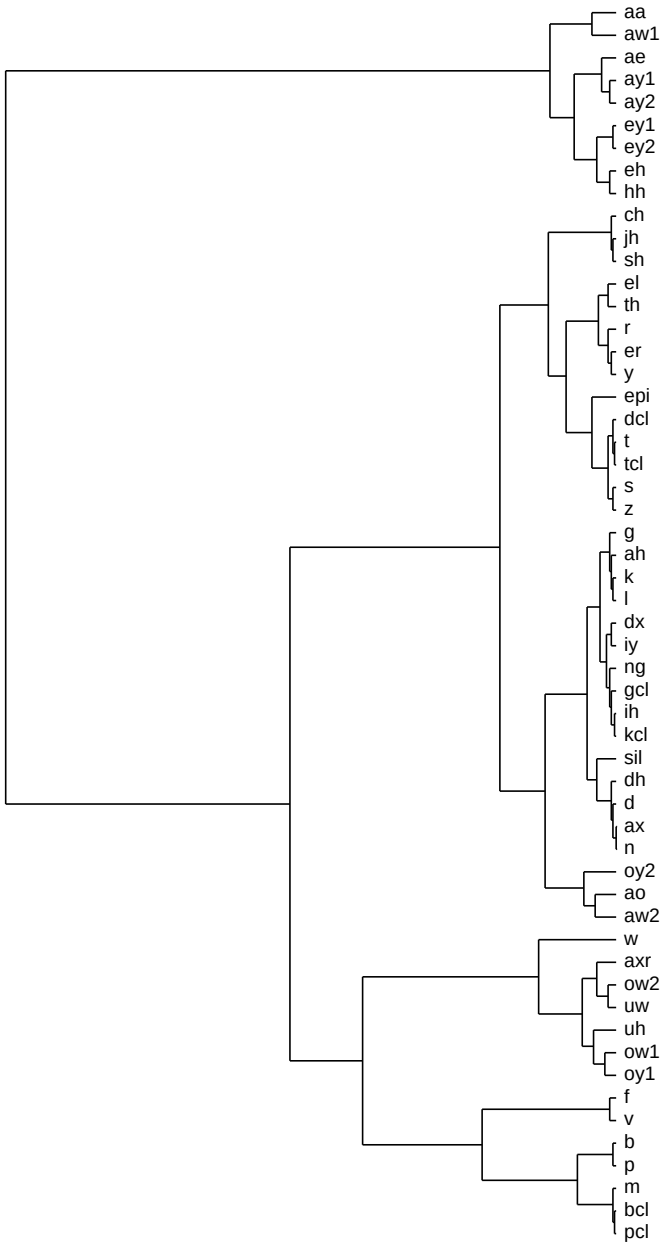


Figure 4-14: Cluster plot using a DCT-PCA encoding of a small ROI.

{t,d}, {f,v}, etc., although some of the immediate pairs, such as {n, ax}, are more difficult to justify. If we look at the cluster plot using pixel-based PCA features extracted from static frames (Fig. 4-8,) we can see that beyond the obvious confusable pairs, the larger granularity classes are as follows:

- mouth wide open
- mouth fully closed
- rounded lips
- everything else

It is interesting to observe the high level tree structure: by drawing an imaginary line a certain distance from the top of the tree, we can split it into arbitrarily high level clusters. For comparison, in an acoustic graph, the two top-level clusters correspond roughly to vowels and consonants. In the visual case, however, the top-level structure seems to change depending on the underlying signal representation. For example, if we look at the static pixel PCA features (Fig. 4-8), the two top-level clusters correspond to the "wide open lips" phonemes /aa/,/aw/,/ae/,/ay/,/hh/,/eh/,/ey/, and to the rest of the "less open or closed lips" phonemes. The situation is similar in the case of the DCT features derived from the small ROI (Fig. 4-14.) On the other hand, for the large ROI data (Fig. 4-10,) the two top-level clusters are more balanced, the more "open" cluster including not only the "wide open lips" group, but also the vowels /ah/,/dx/,/iy/, the liquid /l/ and the velar closures and bursts /gcl/, /kcl/, /ng/, /k/, /g/. One possible explanation for this is that the jaw position is now playing a role, so that lip position is not the only differentiating feature. In general, using a bigger ROI results in more distinct clusters that are farther apart.

We now study the effects of using dynamic information as opposed to the static frame. First, extracting the DCT coefficients from the motion frames instead of the full frames has a significant effect on the clustering tree (see Fig. 4-11.) Now that we are looking at the motion just preceding the center point of the phoneme, the two top level clusters have changed. The first cluster now contains the "closed lips"

phonemes /pcl/,/bcl/,/m/,/f/,/v/, which are characterized by the distinct motion of the lips coming together quickly. It also contains the seemingly unrelated second part of the /aw/ diphthong, /aw2/. However, noticing that /aw2/ is characterized by the fast motion of the closing jaw, as are the /pcl/,/bcl/,/m/,/f/,/v/ phonemes (see Fig.4-7), helps explain this fact. Further evidence that jaw movement is playing a role in distinguishing this phoneme from the others lies in the fact that it appears distinct in the larger ROI clustering plots (see also Fig. 4-12), but not in the smaller ROI (not including the chin) clustering plots (see Fig. 4-9.)

Another notable change when looking at motion frame clusters is that the second part of the /oy/ diphthong, /oy2/, is now very distinct from the other phonemes; in fact, it alone comprises one of the four top-level clusters, which also include the "lips closing" cluster described above and two other, more broad, clusters (Fig. 4-11.) This is perhaps not surprising, considering how distinct the mean difference frame for /oy2/ looks in Fig. 4-7, suggesting the action of the lip corners pulling apart and exposing the teeth.

While some phonemes become more distinct when motion is considered, others become less distinct and more confusable. For example, the velar bursts /g/ and /k/, previously closely grouped together, are now clustered almost randomly, the former with /axr/ and the latter with silence. A possible explanation is that the work in creating a glottal burst is being done in the back of the throat, leaving the visible articulators free to move in preparation of the next sound.

As we keep adding even more dynamic information by stacking static frames or stacking motion frames together, we continue to observe the above two trends (see Fig. 4-12 and Fig. 4-13.) On the one hand, certain phoneme clusters are becoming more distinct; on the other hand, the cluster of less distinguishable phonemes is growing larger. For example, in Fig. 4-13, we observe the following distinct groups:

- /pcl/, /bcl/, /m/, /f/, /v/
- /aw2/, /ay2/, /ey2/
- /b/, /p/

- /oy1/, /w/
- /oy2/
- /ao/, /aa/, /ah/, /ay1/, /eh/, /ey1/, /ae/, /aw1/

which can be loosely described by the following articulatory events:

- lower lip touching upper lip or upper teeth
- lower lip and jaw moving up from a wide open position
- lips coming apart quickly
- lips rounding
- mouth corners moving apart quickly
- lower lip and jaw moving down

The rest of the phonemes (a little more than half) are found in one large cluster. In particular, the velar bursts /g/,/k/ are now both closely grouped with silence. Since, in our database, silence corresponds half of the time to a closed mouth and half of the time to an open mouth, we conclude that /g/,/k/ must be just as variable in terms of their motion.

One of the reasons why the above clusters are so easy to differentiate with the stacked motion frame representation may be the speed of the articulator movements. For example, the bilabial burst happens very quickly, certainly within three frames. One suggestion for future work would be to study longer sequences in order to uncover slower distinctive articulatory events. Another reason may be that some of them are the second part of a diphthong, meaning that they always occur in the same context. Overall, these clusters show that, when it comes to the dynamic content of visual speech, the most distinctive visual units seem to be closely tied to the physical motion of the articulators.

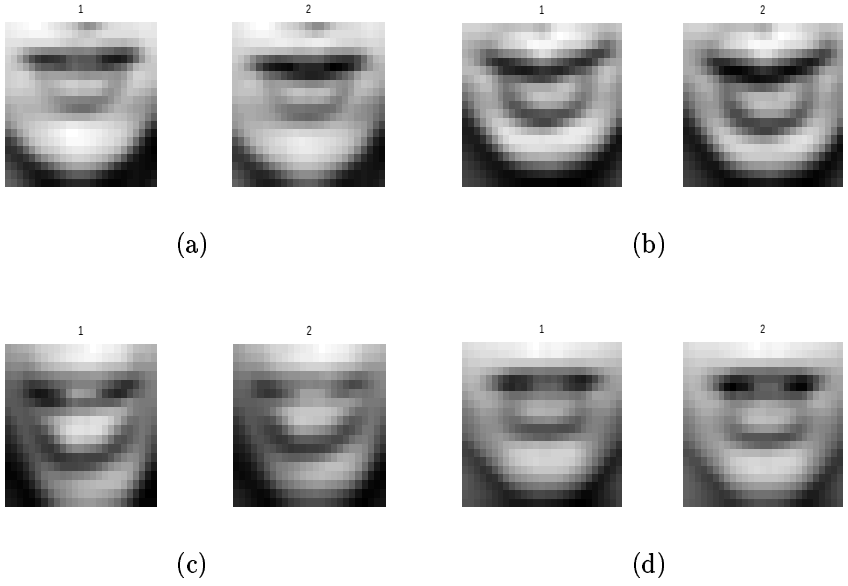


Figure 4-15: K-means clustering using two clusters.

4.4 Unsupervised Clustering

In the previous section, we used a supervised clustering technique to determine natural clusters of phonemes that might correspond to visible articulatory features. In this section, we depart from the standard technique of finding distinctive visual units through alignment with acoustically determined phoneme labels. The motivation behind this departure is that we would like to be able to recognize the underlying gestures associated with articulation. While these gestures are certainly correlated with the produced sounds, and, consequently, with the phonemes derived from either a manual or an automatic segmentation of the audio signal, some articulatory information may be lost in the process. For instance, the fact that the video and audio speech signals exhibit some amount of asynchrony indicates that an acoustically derived segmentation may not be appropriate. In addition, as we saw in the previous section, grouping together visual frames labeled with the same phoneme sometimes yields very variable clusters.

In general, our goal is to discover the natural building blocks of visual speech without necessarily restricting them to be synchronous with audio segments. As a

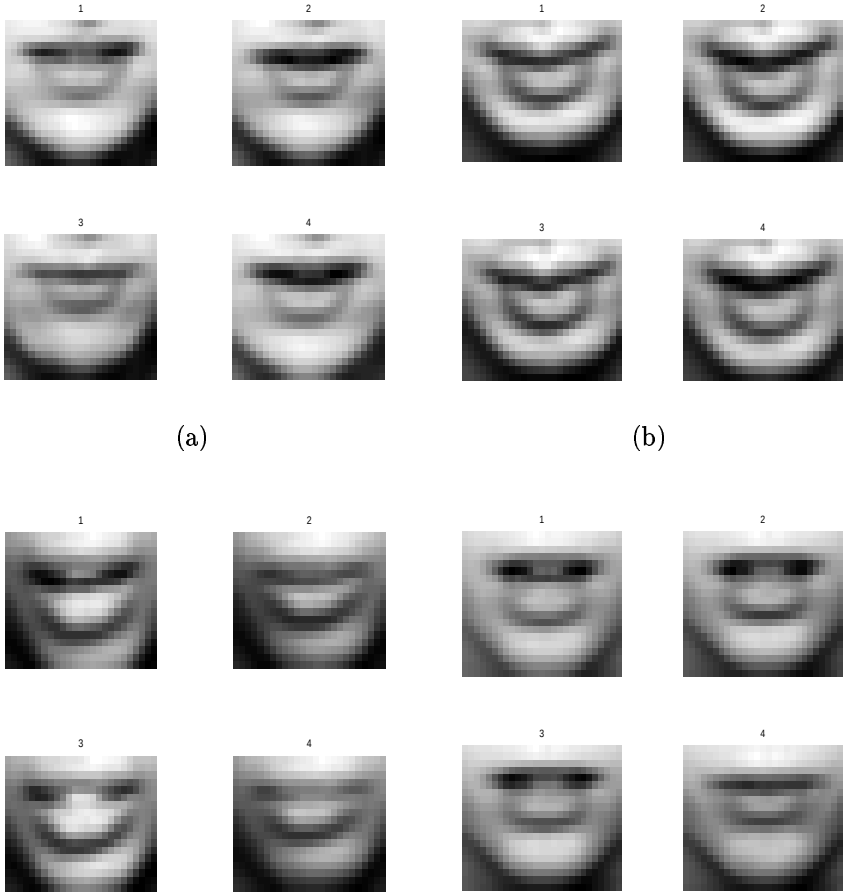


Figure 4-16: K-means clustering using four clusters.

preliminary investigation, we try an unsupervised clustering approach using the K-means algorithm. This is the simplest technique, and depends on having a fixed number of clusters. The results of K-means clustering of visual frames for each of four different speakers is shown in Figure 4-15. In this case, two clusters were used, and the means of each cluster are displayed. It is evident from the figures that the algorithm separates the mouth images according to the degree of opening. This leads to the observation that the lip/jaw opening and closing is the most distinguishable articulatory feature, and is consistent across different speakers. Note, however, that the actual appearance of cluster centroids varies from speaker to speaker.

Figure 4-16 shows the results of K-means clustering using four clusters. We can observe that the four most salient articulatory configurations are more of less consistent across different speakers. If we look at the distribution of phonetic labels for each cluster, they fall into the following lip-opening categories: 1) closed lips, 2) a narrow opening between the lips, 3) a medium opening, and 4) a wide-open mouth.

The obvious limitation of this approach is that the number of clusters must be specified in advance.

4.5 Manual Labeling of Articulatory Features

Ideally, the ground truth for the articulator trajectories should be obtained through accurate tracking each of the articulators. This can be achieved, for example, by attaching sensors to the speaker's vocal tract organs, as is done in Electromagnetic Articulography (EMA.) EMA provides two-dimensional kinematic data of most of the articulatory structures of interest, i.e. lips, jaw, tongue and velum, in readily analyzable form. Other monitoring devices include the x-ray microbeam system and MRI. Alternatively, since we do not have access to such devices, we could utilize computer vision techniques for tracking objects in video. For example, several methods for tracking lip contours have been proposed. The disadvantage is that such methods usually require the initialization and training of models, often requiring the user to click on mouth contour points; they also frequently suffer from tracking failures. Since

a manual initialization step seems inevitable, it may be useful to forego the tracking step altogether and instead label the articulatory features directly. This is, in fact, the approach we take in this section.

Manual labeling of speech data is the most commonly used technique for annotating acoustic speech corpora. Normally, transcription is done by trained linguists following a set of guidelines. However, as mentioned above, visual corpora are normally labeled using acoustic forced alignments, therefore, no analogous guidelines exist for visual speech transcription. Nevertheless, one might imagine a lipreading expert being able to assign either phoneme (viseme) or gesture labels to a sequence of facial images.

First, we must decide what set of labels to use in the transcription. Note that we choose to label the absolute state of the lips, jaw, and tongue at a particular instant in time, rather then their movements. Motion labels can then be inferred from the absolute labels. Also, note that articulators go through a continuous range of positions, while phoneme labels are discrete. We use a similar discretization of the space of all possible positions into a small set of feature values.

After visually inspecting recorded video sequences from the AVTIMIT database, we arrived at the feature set shown in Table 4.5. We used the analytically derived feature set shown in Table 4.1 as the basis, taking into account the following considerations: i) the features should describe all the articulations that the human labeler can distinguish from the video; ii) however, the number of features and their values should be small to avoid increased computational complexity. Notice that we did not use any tongue features. There are two reasons for this. The first one is that, since the tongue is partially hidden from view, its position is difficult to judge and must be inferred most of the time. The second reason is that we wanted to keep the number of features small in order to conduct the initial proof-of-concept experiments.

We have broken up the LIP-LOC feature into two binary features: LIP-LOC, which indicates lip protrusion, and LAB-DENT, which indicates the lower lip pressing up against the upper teeth. The reason for having two separate features is that it gives us more information in the case when the lips are in the labio-dental position

Table 4.5: The proposed articulatory feature set.

Index	Feature Name	Values
0	LIP-LOC	unrounded (U), rounded (R)
1	LIP-OPEN	closed (C), narrow (N), medium (M), wide (W)
2	LAB-DENT	non-labio-dental (N), labio-dental (Y)

and also protruded.

The third feature, LIP-OPEN, has four values, indicating four different degrees of openness of the lips: *labial*, or completely closed lips, *narrow*, or slight opening between the lips, *wide*, or very wide opening, and *medium*, or all other degrees between *narrow* and *wide*. We chose to partition the space of degree of opening into these particular four categories based on visual inspection, and also on the results of unsupervised clustering in the previous section. We could, of course, have more levels of opening, however, this would increase the parameter space.

We compare our manual transcriptions to labels generated automatically by mapping a phonetic transcription to feature values using Table 4.6. Figure 4-17 shows the alignment of the manual labels for each feature with the automatically generated mapped labels for fragments of the following three utterances: (a) "Don't ask me to carry an oily rag like that", (b) "Barb's gold bracelet was a graduation present", and (c) "A chosen few will become generals." The manual and mapped transcriptions differ: for example, in the automatic transcription for LIP-OPEN in (a), every visual frame for the initial /ae/ segment is labeled *wide*, while in the manual transcription the last three frames are labeled *medium*.

Although it is difficult to quantitatively compare manual and automatic transcriptions, we can at least observe how they affect the classifier. We trained two sets of SVM classifiers for LIP-OPEN and LIP-LOC, using mapped labels in the first set and manual labels in the second. The classifiers achieve 89-90% accuracy on training data using mapped labels, and 95-96% using manual labels, indicating that perhaps the manual labels are more consistent and therefore easier to recognize. We show a sample of the outputs of the classifiers for LIP-OPEN in Figure 4-18 and for LIP-LOC

Table 4.6: Mapping from phonemes to articulatory feature values.

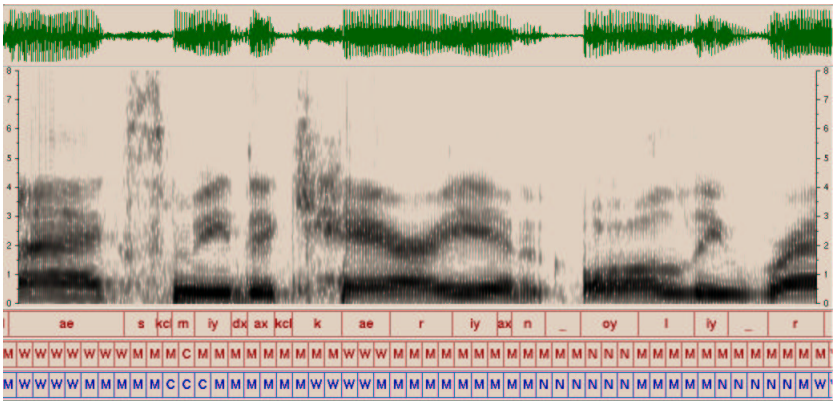
phone	LIP-LOC	LIP-OPEN	LAB-DENT
aa	U	M	N
ae	U	W	N
ah	U	M	N
ao	R	M	N
aw1	U	W	N
aw2	R	M	N
ax	U	M	N
axr	R	M	N
ay1	U	M	N
ay2	U	M	N
b	U	N	N
bcl	U	C	N
ch	U	M	N
d	U	M	N
dcl	U	M	N
dh	U	M	N
dx	U	M	N
eh	U	M	N
el	U	M	N
em	U	C	N
en	U	M	N
er	R	M	N
ey1	U	M	N
ey2	U	M	N
f	U	N	Y
g	U	M	N
gcl	U	M	N
hh	U	M	N
ih	U	M	N
iy	U	M	N
jh	U	M	N
k	U	M	N
kcl	U	M	N
l	U	M	N
m	U	C	N
n	U	M	N
ng	U	M	N
ow1	R	M	N
ow2	R	M	N
oy1	R	N	N
oy2	U	M	N
p	U	C	N
pcl	U	N	N
r	R	M	N
s	U	M	N
sh	U	M	N
t	U	M	N
tcl	U	M	N
th	U	M	N
uh	R	M	N
uw	R	N	N
v	U	N	Y
w	R	N	N
y	U	M	N
z	U	M	N
zh	U	M	N
epi	U	M	N
sil	U	M	N

in Figure 4-19. In each of the figures, the top part shows outputs of the mapped label classifier, and the bottom part the outputs of the manual label classifier. Upon visual inspection, the manual label classifiers seem to be assigning more correct labels.

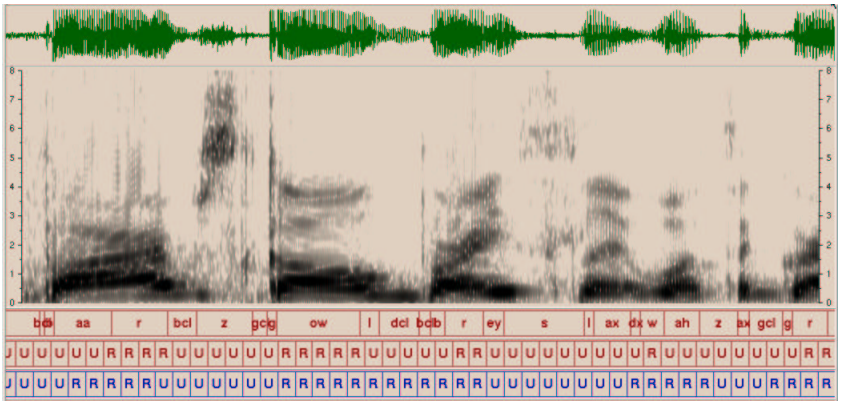
4.6 Conclusion

In the first half of this chapter, we analyzed visual data corresponding to the center of acoustically labeled phonemes, using an agglomerative clustering technique. The goal was to investigate the basic structure of continuous, multi-speaker visual speech. The results show that, although this method can be used to map phonemes to visual units, the optimal mapping depends on the region of interest, on whether one uses motion or static frames, and possibly on the length of the time window. This suggests that different signal representations can be used to provide complementary information. Furthermore, while some of the clusters corresponding to conventional visemes and are stable across representations (eg. {bcl,pcl,m}, or the bilabial viseme,) others are unstable and much less distinct. Overall, the results indicate that it may be better to think of visual speech units in terms of articulatory gestures, such as lips closing together, jaw moving up, and so on, rather than groups of visually similar phonemes.

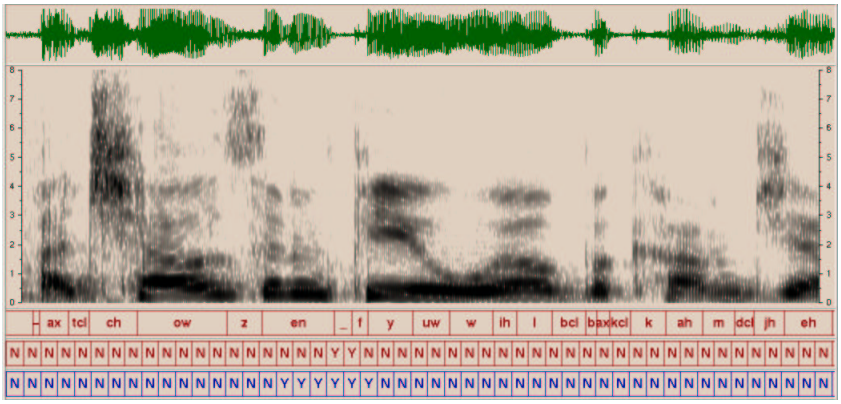
In the last half of this chapter, we proposed a set of articulatory visual features that will be used in our recognizer. We attempted to learn these features in an unsupervised manner from the visual database, however, more research is needed in this direction before an algorithm for visual data driven unit extraction can be proposed. Therefore, to facilitate the evaluation of our model in the next chapter, a manual transcription procedure was carried out.



(a) LIP-OPEN

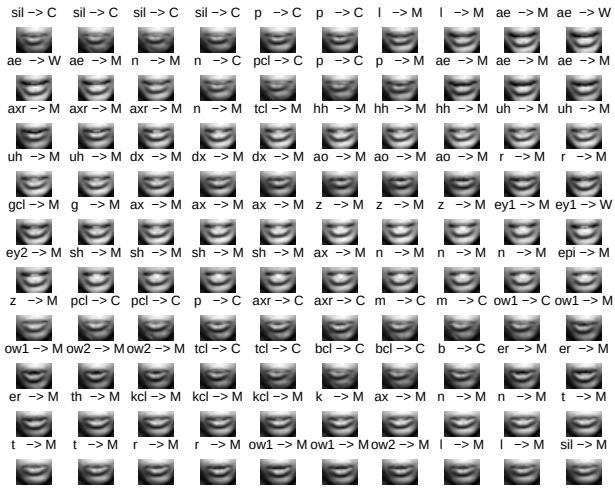


(b) LIP-LOC

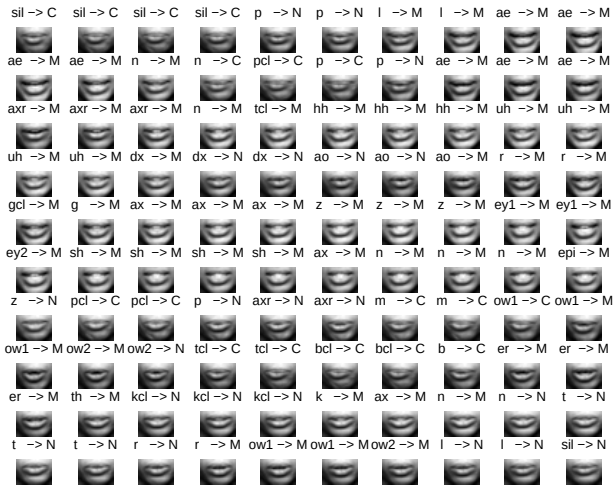


(c) LIP-FRIC

Figure 4-17: Alignment of manually labeled (top) and automatically labeled (bottom) features.

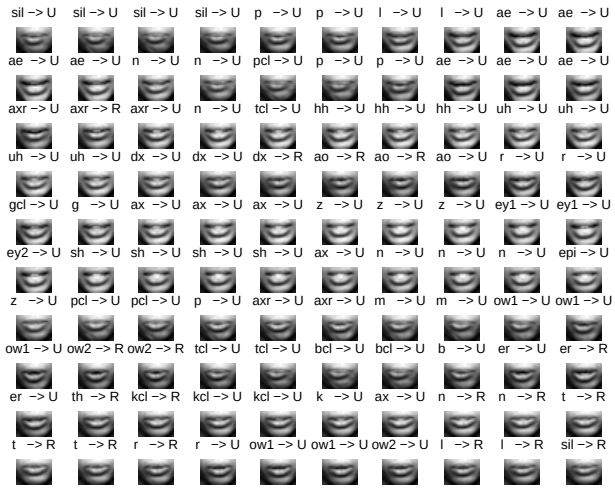


(a) Mapped

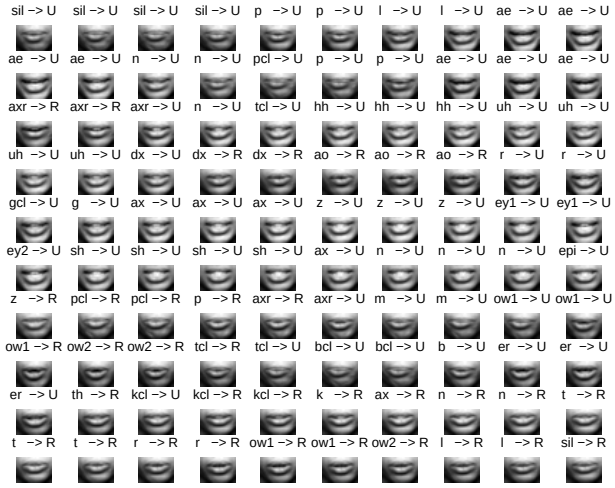


(b) Manual

Figure 4-18: Outputs of SVM classifiers for LIP-OPEN trained on (a) mapped labels and (b) manual labels.



(a) Mapped



(b) Manual

Figure 4-19: Outputs of SVM classifiers for LIP-LOC trained on (a) mapped labels and (b) manual labels.

Chapter 5

Experimental Evaluation

In this chapter, we compare the performance of AF-based visual speech recognition to conventional viseme-based recognition. We conduct two different types of experiments. In Section 5.1, we test an AF-based viseme classifier on visually noisy data, and show that it can lead to improved viseme classification rates in the presence of image noise. As this is still very much a work in progress, we have only limited initial experiments to report. Nevertheless, they indicate that our approach increases classification rates on a simple task, and, therefore, merits further investigation.

In Section 5.2, we evaluate our AF-based word recognizer on manually transcribed visual speech data for one speaker in the AVTIMIT database. We demonstrate how the recognizer handles context effects by allowing feature asynchrony and substitution.

5.1 Viseme Classification in the Presence of Visual Noise

5.1.1 Experimental Setup

In this section, we conduct our initial proof-of-concept experiments on a small two-speaker audio-visual speech corpus previously collected in our lab. The corpus consists of continuous repetitions of a nonsense utterance designed to provide a balanced coverage of English visemes. In order to facilitate the accurate extraction and tracking of

Table 5.1: Viseme to feature mapping.

Viseme	LIP-OPEN	LIP-ROUND
/ao/	Wide	Yes
/ae/	Wide	No
/uw/	Narrow	Yes
/dcl/	Narrow	No

the mouth region, the first speaker's lips were colored blue. A color histogram model was then used to segment the lip region of interest. The second speaker's lips were not colored, but rather segmented using correlation tracking, which resulted in imperfect ROI localization. Viseme labels were determined from an audio transcription, obtained automatically using an audio speech recognizer, via the mapping described in Table 5.1. Figure 5-1 shows some sample viseme images taken from the center of the corresponding phonetic segments. In this case, each viseme corresponded to a single phoneme.

Prior to classification, the original 120x160 sample image was scaled down to 10x14 pixels in size and then vectorized to form a 140-element data vector. The decision to use very simple image features (pixels) as input to the SVM was intentional. When applied to other pattern recognition tasks, SVMs have achieved very good results using only such simple input features. Furthermore, we wanted to allow the discriminative power of the SVM determine those parts of the image that are key to a particular feature without making any prior assumptions. We used a training set consisting of 200 samples per viseme, and a separate "visually clean" test set of 100 samples per viseme. The "visually noisy" test sets we created by either adding random Gaussian pixel noise to the down-sampled test images, or blurring the original images with a Gaussian filter to reduce their effective resolution.

As a start, we applied our approach to the task of viseme classification. For this experiment, we used only four visemes, corresponding to the phonemes /ao/, /ae/, /uw/ and /dcl/. We chose the viseme set so that it could be completely encoded by the cross product of two binary articulatory features, in this case, LIP-OPEN

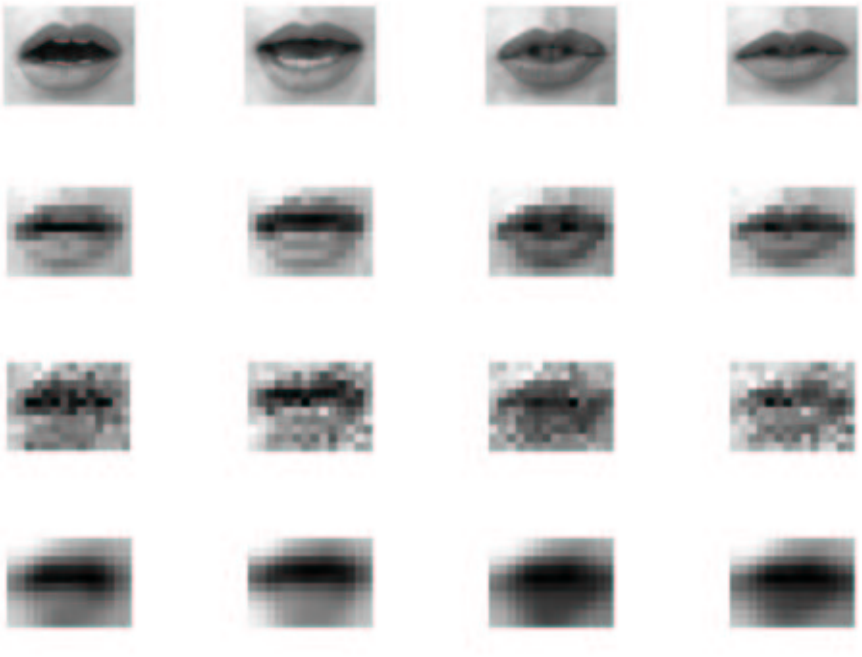


Figure 5-1: Sample viseme images for Speaker 1, from left to right: /ao/, /ae/, /uw/ and /dcl/. The original high-resolution images (top row); resized clean images used for training (2nd row); with added 50% pixel noise (3rd row); and blurred with Gaussian kernel of size 10 (bottom row).

and LIP-ROUND. Table 5.1 shows the mapping from the visemes to the articulatory feature values. In the general case, there would be on the order of a few dozen visemes, and so the number of features would necessarily increase. Note that we could have used more features, such as the visibility of teeth or the tongue position, making the feature set redundant.

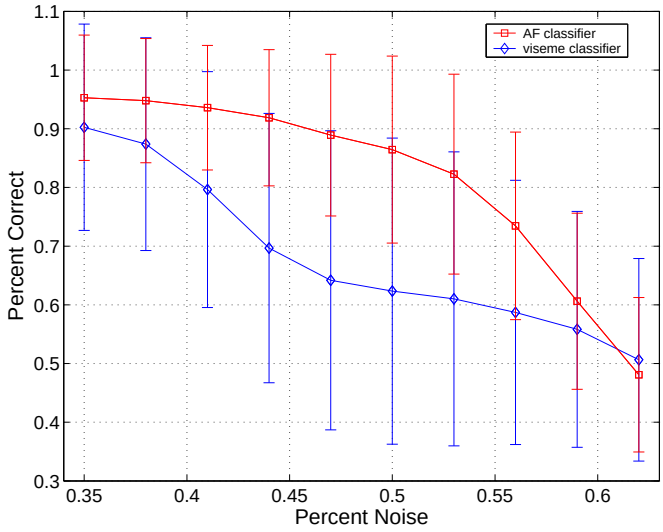
A separate SVM classifier was trained for each viseme, as well as for each of the two features, using LIBSVM software [6], which implements the "one-against-one" multi-class method. We used the radial basis function (RBF) kernel in all experiments, as we found it to give the best performance with the fewest free parameters. The RBF is defined as follows:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \quad \gamma > 0, \tag{5.1}$$

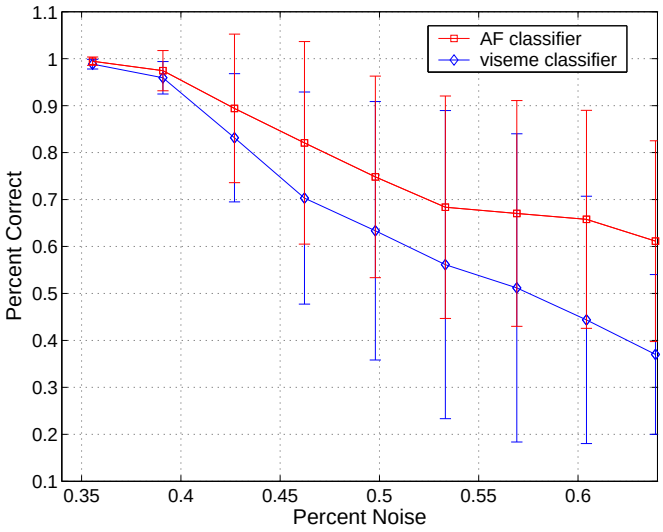
where x_i, x_j are training samples. Therefore, in addition to the penalty parameter of the error term, C, the RBF SVM has another free parameter, γ. To find the optimal values for these two parameters, we performed v-fold cross-validation on the training data. Figures 5.1.2 and 5.1.2 show the contour plots obtained from a grid search on C and γ for the 4-class "viseme" SVM and the two binary articulatory feature SVMs. The red star indicates the smallest parameter values at which the peak accuracy was achieved. Note that while the optimal parameters for the "viseme" and LIP-ROUND SVMs are similar, the optimal LIP-OPEN SVM parameters are lower, suggesting that it may have better generalization in the presence of noise, since a smaller value of g means a wider Gaussian. During classification, feature labels were converted to viseme labels using the mapping shown in Table 5.1. This is the simplest possible combination rule. Another alternative would have been to train a second-level viseme classifier that takes the concatenated probabilities of the two features obtained from the two first-level classifiers as input.

5.1.2 Results

Figure 5-2 shows the classification rates obtained by each classifier across several levels of random pixel noise, averaged over 20 training and testing runs. The horizontal



(a) Speaker 1



(b) Speaker 2

Figure 5-2: Comparison of viseme classification rates obtained by the AF-based and viseme-based classifiers on test data with added random pixel noise.

Table 5.2: Classification rates on pixel noise data for Speaker 1.

Noise Level	Viseme	OPEN	ROUND	Combined
None	99	100	99	99
30%	69	100	99	99
35%	50	100	98	98
40%	38	100	96	96
45%	27	100	84	84
50%	25	100	60	60
55%	25	94	51	48
60%	25	74	50	37

axis shows the percentage of Gaussian noise that was added to the test images. The vertical axis shows the correct viseme classification rate. Results for each speaker are shown on separate plots.

Tables 5.2 and 5.3 show the classification rates obtained by each classifier across several levels of random pixel noise in one particular run of training. The first column shows the percentage of Gaussian noise that was added to the test images. The second column shows the viseme classification rate using the viseme classifier, and the next two columns show the respective LIP-OPEN and LIP-ROUND feature classification rates. The last column shows the viseme classification rate obtained by combining the results of the individual feature classifiers. Table 5.4 shows the classification results on the low-resolution test data for Speaker 1. The first column shows the size of the Gaussian kernel used to blur the original high-resolution images. One interesting fact is the resilience of the SVM to significant amounts of noise and blurring. This could be attributed to the fact that the four chosen visemes can be distinguished using mostly low-frequency information. The same result may not hold for other visemes that can only be distinguished by high-frequency information, such as a small opening between the lips, etc. Overall, the results of our preliminary experiments clearly show the advantage of using articulatory feature modeling for viseme recognition from noisy images. While the viseme classifier's performance degrades significantly with increasing noise levels, the combined feature-based classifier retains a high recognition rate.

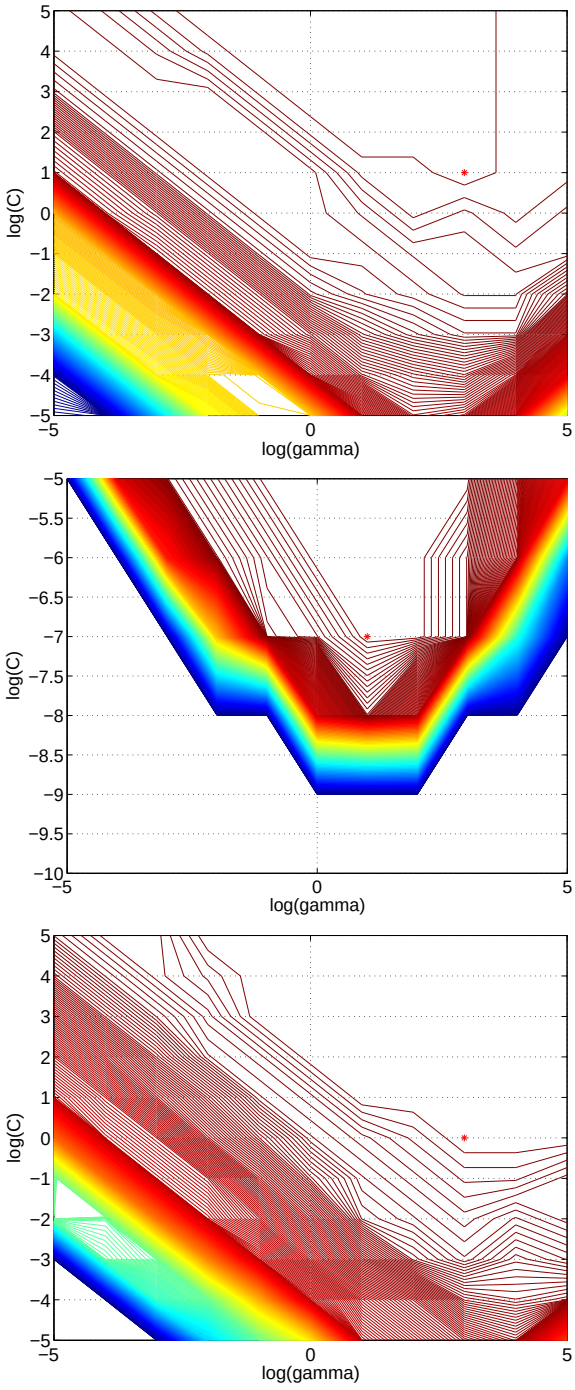


Figure 5-3: Contour plots of cross-validation accuracy as a function of the C and γ parameters for the "viseme" (top), "LIP-OPEN" (middle) and "LIP-ROUND" (bottom) SVMs for Speaker 1.

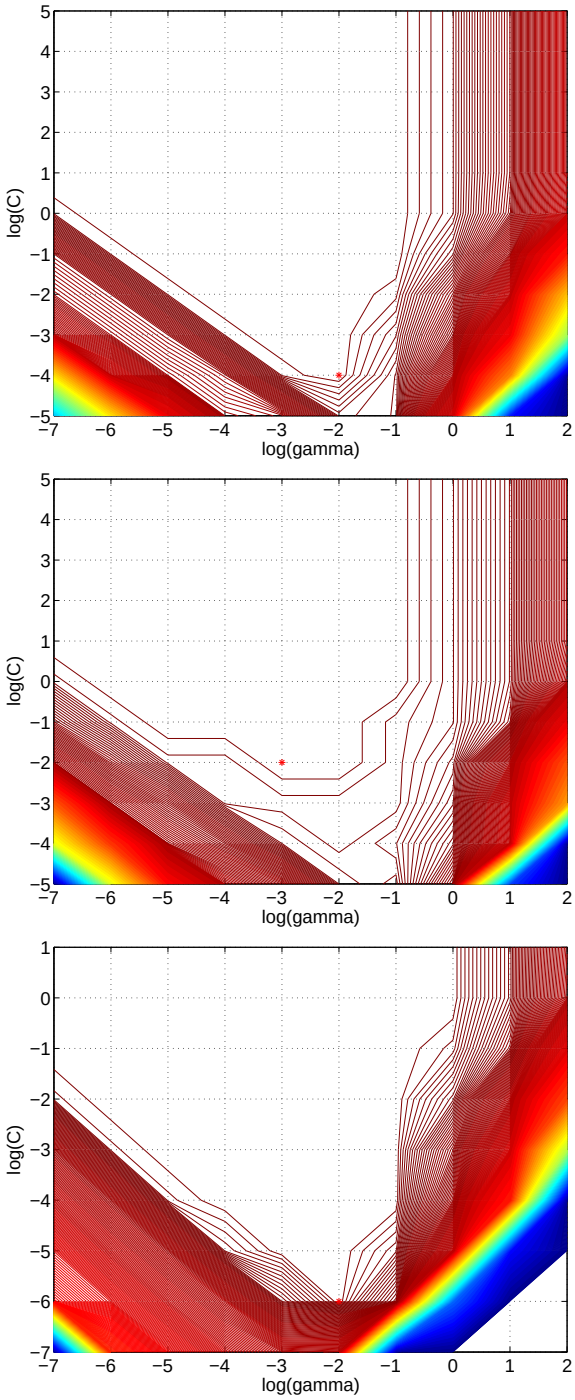


Figure 5-4: Contour plots of cross-validation accuracy as a function of the C and γ parameters for the "viseme" (top), "LIP-OPEN" (middle) and "LIP-ROUND" (bottom) SVMs for Speaker 2.

Table 5.3: Classification rates on pixel noise data for Speaker 2.

Noise Level	Viseme	OPEN	ROUND	Combined
None	100	100	100	100
30%	100	100	100	100
35%	99	100	100	100
40%	93	100	100	100
45%	59	100	100	100
50%	49	100	97	96
55%	37	99	82	81
60%	25	79	54	43

Table 5.4: Classification rates on low-resolution data for Speaker 1.

Kernel Size	Viseme	OPEN	ROUND	Combined
None	99	100	99	99
9	97	100	99	99
10	90	99	99	98

5.2 Word Recognition Using Manual Transcriptions

5.2.1 Experimental Setup

In this section, we describe our preliminary experiments using the proposed DBN model for articulatory feature-based word recognition (see Figure 3-3.) The goal of these experiments is two-fold. First of all, we would like to evaluate the feasibility of the proposed DBN structure. Although a similar model has been successfully used for feature-based lexical access in [30], the frame rate, the feature set and, of course, the modality are all different in this case. The second goal is to compare the performance of a multi-stream feature DBN to a viseme-based word recognizer. In our experiments, the latter is implemented by forcing the features to be completely synchronous and by not allowing any substitutions. Thus, the observed features are simply mapped to visemes, which are then used in lexical access. In the proposed

DBN, both feature asynchrony and substitution are allowed.

The recognizer takes as input the values of the observed feature variables *LL-Obs*, *LO-Obs* and *LD-Obs*. In this case, these values were obtained from manual feature transcriptions described in Section 4.5. In the future, manually transcribed labels will be replaced with SVM classifier outputs. Additionally, these hard decisions will be converted into posterior probabilities. However, for now, the main goal is to do a feasibility study of the general approach, so we use the transcriptions, keeping in mind that the actual classifier outputs will not be as accurate.

Eleven utterances from the AVTIMIT corpus, taken from a single speaker, were transcribed with the three feature values. Out of the resulting 162 words, the 139 words that remained after excluding words with fewer than 4 frames (a GMTK requirement) were used in the experiment. The total vocabulary size was approximately 1800 words.

5.2.2 Results

In general, we would not expect to achieve a very good word recognition rate from only visual observations on such a relatively large vocabulary. Also, we are only using three articulatory features in our model. Therefore, it is not surprising that the baseline model recognized only 9 words out of 139. The feature-based model recognized 13 words correctly, which is an improvement over the baseline. However, it may be more informative to look at the distribution of the ranks of the correct word, plotted in Figure 5-5. The cumulative distribution for the feature-based DBN is significantly higher than that for the baseline DBN. This means that, given the top N highest-ranked word hypotheses, the probability of having the correct word among that list has improved. In other words, where the baseline model may give the correct word a low probability score, our model is giving it a higher score, which is encouraging.

To illustrate the type of variability in the visual realization of a word that our DBN is modeling better than the baseline DBN, we show two sample feature stream alignments in Figure 5-6. The top part of each figure shows the spectrogram of the

utterance; the four bottom lines are aligned transcriptions of (from top to bottom): the phonemes, and the LIP-OPEN, LIP-ROUND and LAB-DENT features.

The top figure is an excerpt from the utterance "A chosen few will become generals." In the canonical representation, the words "few will", map to the phoneme sequence "/f/ /y/ /uw/ /w/ /ih/ /l/", which in turn maps to the LIP-ROUND feature values of "U U R R U U". However, in the actual realization of the sentence, we observe that the feature is has spread and has the value R for almost the entire length of the segment. This example falls under the category of context-dependent feature substitution, and is handled by the model by allowing the feature value to change from U to R. Also, note that in the word "generals," the LIP-ROUND feature continues to have the value R for one extra frame after the LIP-OPEN feature has already switched to the value M. This is an example of feature asynchrony and is handled by the DBN by allowing the features to proceed at different rates.

The bottom example is an excerpt from the utterance "Barb's gold bracelet was a graduation present." Notice that in the word "graduation," the LIP-OPEN feature has the value *closed* during the production of /n/. This is due to the effect of articulatory anticipation of /pcl/ in "present," which causes most of the velar /n/ to be produced with closed lips. This is another example of feature asynchrony, where the lips got a ahead of the velum.

One of the limitations of the current system is that it assumes that the features will synchronize at the beginning and at the end of each word. Therefore, while it can model cross-phone asynchrony, it cannot model cross-word asynchrony. However, the system will still be able to explain some instances of cross-word feature spreading as feature substitution. For example, in the word "few", it could model the rounding of the initial /f/ as the substitution of the value of the LIP-ROUND feature. This also applies to the "graduation" example above. This limitation will be remedied in a future version of the model.

Figure 5-7 shows a sample alignment for the word "supervision." It demonstrates how system aligns the underlying feature values, as determined by the dictionary and the phone-to-feature mapping tables, to the observed feature values.

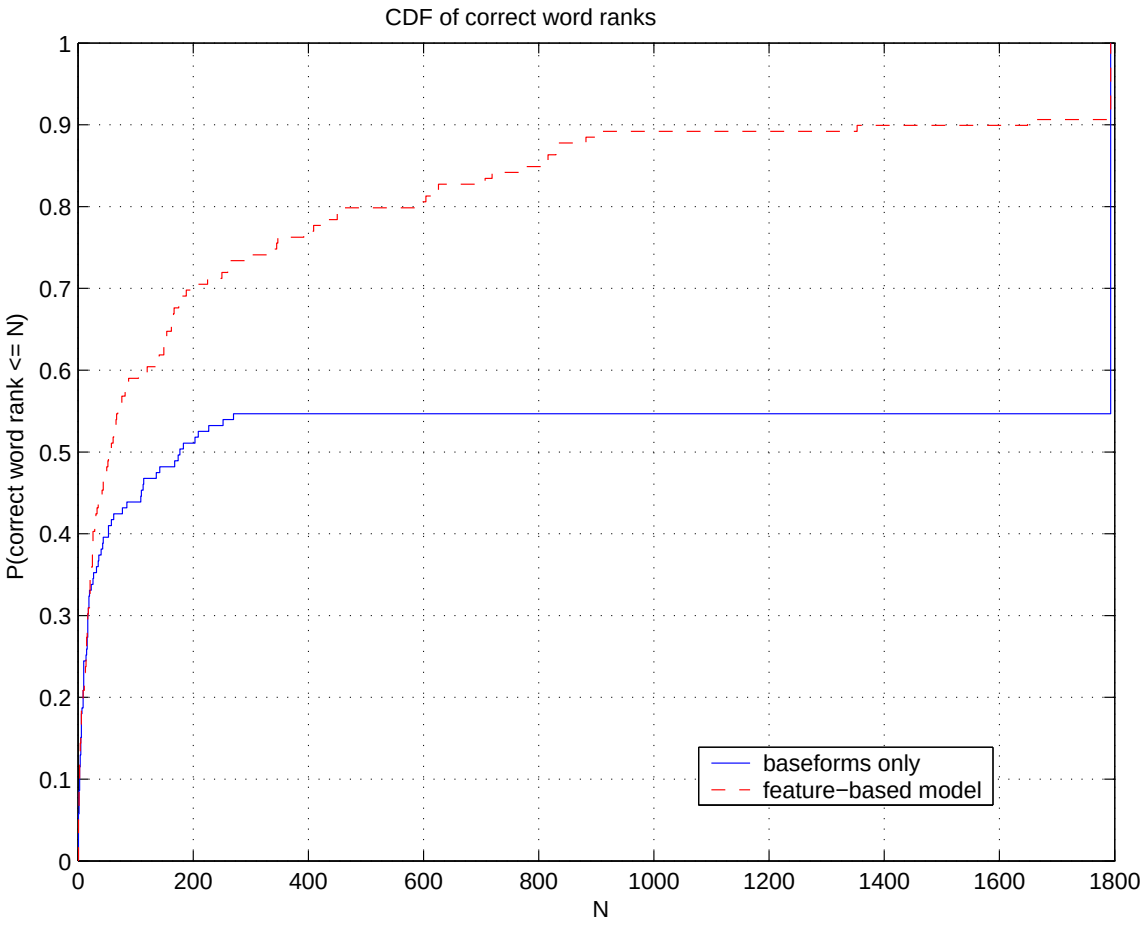
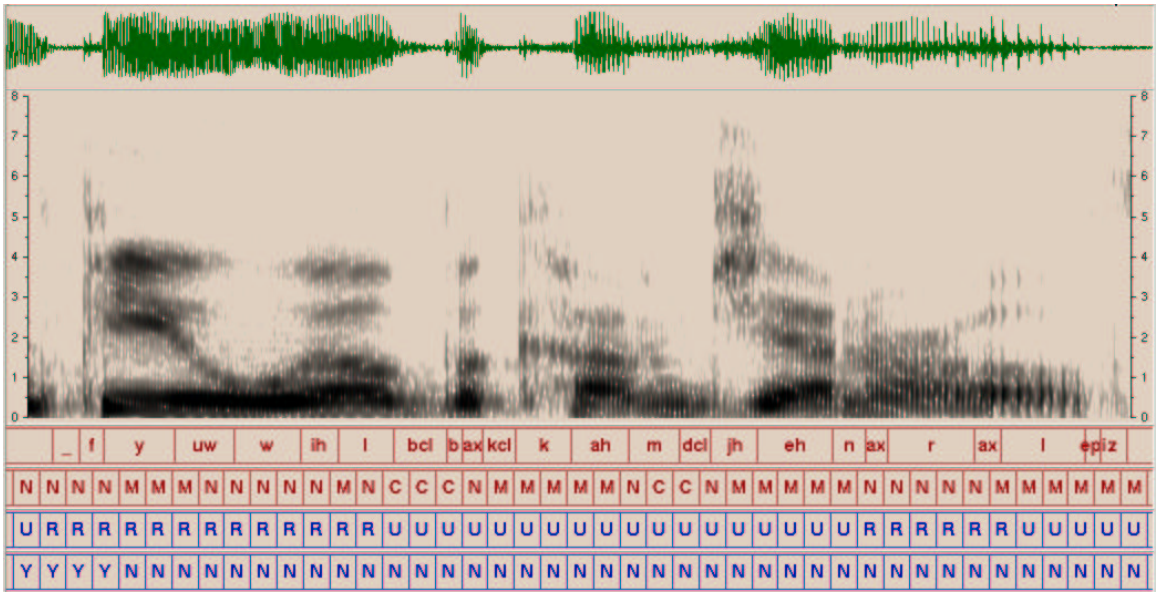
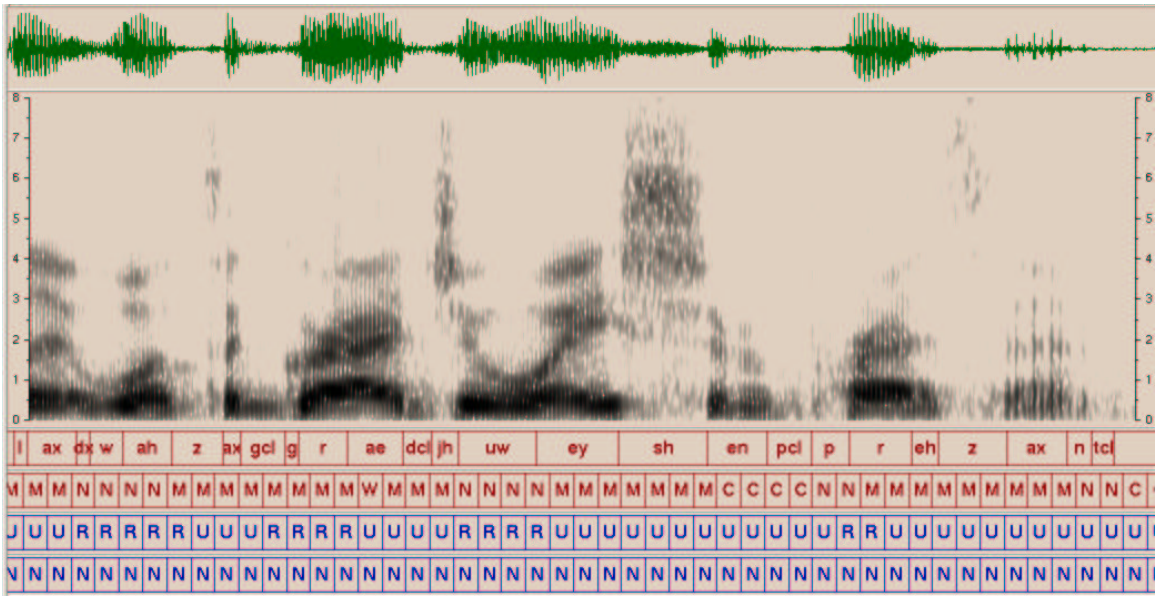


Figure 5-5: Rank of correct word, cumulative distribution.



(a)



(b)

Figure 5-6: Aligned feature transcriptions for two utterances.

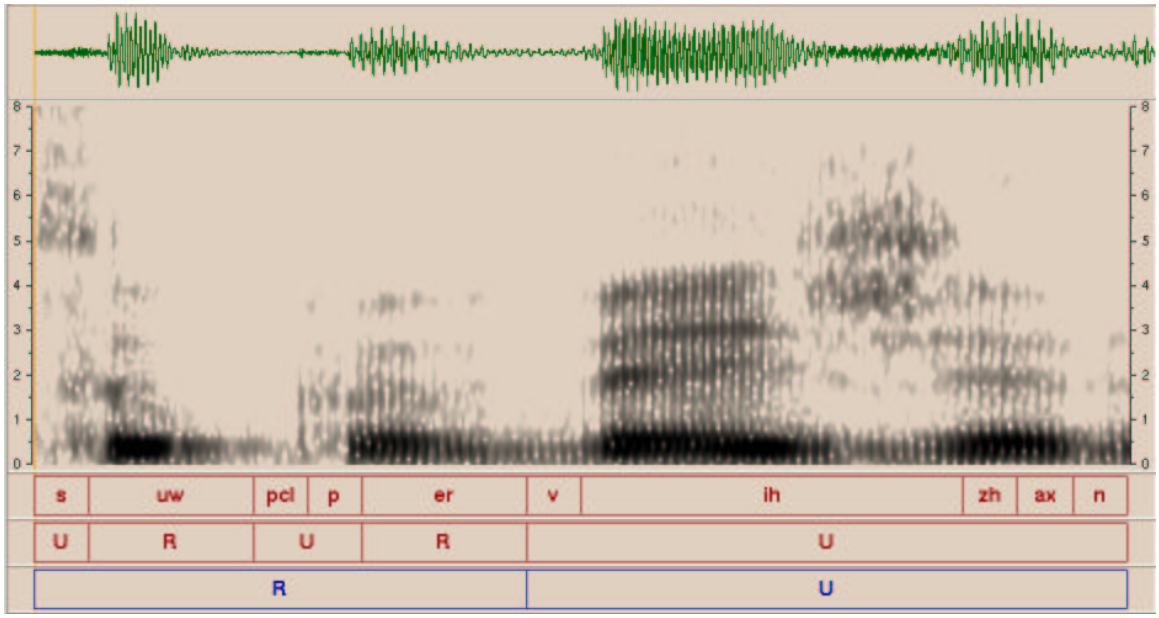


Figure 5-7: Sample alignment for the word "supervision".

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this work, we began laying the groundwork for feature-based visual speech modeling. We analyzed the natural clustering of visual phoneme segments and showed that, like their acoustic counterparts, visual units can be grouped according to articulatory features. We proposed a set of articulatory features for use in a feature-based visual speech recognizer, and conducted some initial experiments showing the robustness of feature-based classification at high levels of noise. We also trained and evaluated a feature-based word recognizer on manual feature transcriptions, demonstrating the feasibility of the visual DBN model.

6.2 Future Work

As this research is still in its early stages, there are many interesting open issues to pursue in the future. We would like to investigate automatic methods for labeling articulatory features, and to expand the number of features to cover the set of all possible visemes. Also, we used an SVM classifier for our experiments, however, we would like to explore whether other classifiers benefit from the articulatory feature modeling approach as well.

Since different feature detectors tend to have different levels of robustness, we

would like to be able to selectively ignore feature detectors that are unreliable in the process of lexical access. That way, recognition using a partial feature specification may be possible even in the face of severe degradation of the visual signal.

In addition, it has been noted that using articulatory features overlapping in time leads to advantages in context modeling over traditional multi-phone models [54]. Since the feature spreading property is particularly noticeable in the lip features, it would be interesting to evaluate this approach to context modeling in visual speech, and compare it to the traditional multi-phone approach. We would also like to extend the DBN to handle cross-word asynchrony.

Finally, the merits of the feature approach in an integrated audio-visual speech recognizer should be explored.

Appendix A

The AVTIMIT Corpus

A.1 Corpus Collection

To provide an initial corpus for our research in audio- visual speech recognition we collected a new corpus of video recordings called the Audio-Visual TIMIT (AV-TIMIT) cor- pus. It contains read speech and was recorded in a rela- tively quiet o ce with controlled lighting, background and audio noise level. The main design goals for this corpus were: 1) continuous, phonetically balanced speech, 2) mul- tiple speakers, 3) controlled o ce environment and 4) high resolution video. The following sections will describe each aspect of the data collection in detail.

A.1.1 Linguistic Content

Because of size and linguistic exibility requirements, we decided to create a corpus of phonetically rich and balanced speech. We used the 450 TIMIT-SX sentences originally designed to provide a good coverage of phonetic contexts of the English language in as few words as possible [17]. Each speaker was asked to read 20 sentences. The rst sentence was the same for all speakers, and is intended to allow them to become accustomed to the recording process. The other 19 sentences di ered for each round. In total, 23 di erent rounds of utterances were created that test subjects were rotated through. Each of the 23 rounds of utterances was spoken by at least nine di

erent speakers.

A.1.2 Recording Process

Recording was completed during the course of one week. The hardware setup included a desktop PC, a GN Netcom voice array microphone situated behind the keyboard, and a high-quality SONY DCR-VX2000 video camcorder. The camera was mounted on a tripod behind the computer dis- play to record a frontal view of each subject. A blue cur- tain was hung behind the chair to reduce image background noise; however, users were not told to restrict their move- ments. The audio quality was generally clean, but the micro- phone did pick up some noise from a computer fan. The average signal-to-noise ratio within individual utterances was approximately 25 dB, with a standard deviation of 4.5 dB. After being seated in front of the computer, the user was instructed to press and hold the "Record" button on the interface while reading each prompted utterance from the screen. Upon button release, the program echoed the recorded waveform back, so that the user could hear his/her own recording. To help ensure that the speech matched the orthographic transcription, an observer was present in the room to ask the user to re-record a sentence if necessary. For the last ve sentences, extra side lighting was added in order to simulate di erent lighting conditions (see Figure 1). Figure 2: Examples of tracked mouth regions from the AV-TIMIT corpus. The bottom row shows tracking failures.

A.1.3 Database Format

Full color video was stored in uncompressed digital video (DV) AVI format at 30 frames per second and 720x480 reso- lution. In addition to the audio track contained in the video les, the audio was also saved into separate WAV les, sam- pled at 16 KHz. The total database duration is approxi- mately 4 hours.

A.1.4 Demographics

The majority of volunteers came from our organization's community. The nal audio-visual TIMIT corpus contained 223 speakers, of which 117 were male and 106 were female. All but 12 of the subjects were native speakers of English. Different ages and ethnicities were represented, as well as people with/without beards, glasses and hats.

A.2 Annotation

A.2.1 Audio Processing

Time-aligned phonetic transcriptions of the data were cre- ated automatically using a word recognition system con- gured for forced-path alignment. This recognizer allowed multiple phonetic pronunciations of the spoken words. Al- ternate pronunciation paths could result either from a set of phonological variation rules or from alternate phonemic pro- nunciations speci ed in a lexical pronunciation dictionary. The acoustic models for the forced-path alignment pro- cess were seeded from models generated from the TIMIT corpus [17]. Because the noise level of the AV-TIMIT corpus was higher than that of TIMIT (which was recorded with a noise-canceling close-talking microphone), the initial time-aligned transcriptions were not as accurate as we had desired (as determined by expert visual inspection against spectrograms). To correct this, the acoustic models were it- eratively retrained on the AV-TIMIT corpus from the initial transcriptions. After two re-training iterations, a nal set of transcriptions were generated and deemed acceptable based on expert visual inspection. These transcriptions serve as the reference transcriptions used during the phonetic recognition evaluation presented in Section 4.

A.2.2 Video Processing

The video was annotated in two di erent ways. First, the face region was extracted using a face detector. In order to eliminate any translation of the speaker's head, the face sequence was stabilized using correlation tracking of the nose region. However,

since we also needed a reliable way to locate the speaker's mouth, we then used a mouth tracker to extract the mouth region from the video. The mouth tracker is part of the visual front end of the open source AVCSR toolkit available from Intel [6].

Although the front end algorithms were trained on di er- ent corpora than our own, they performed relatively well on the AV-TIMIT corpus. The mouth tracker uses two clas- si ers (one for mouth and the other for mouth-with-beard) to detect the mouth within the lower region of the face. If the mouth was detected successfully in several consecutive frames, the system entered the tracking state, in which the detector was applied to a small region around the previously detected mouth. Finally, the mouth locations over time were smoothed and outliers were rejected using a median lter. For more details about the algorithm, see [8]. The system performed well on most speakers; however, for some it pro- duced unacceptable tracking results (see Figure 2). Two possible reasons for such failures are side lighting and ro- tation of the speaker's head, both of which the system had di culty handling. Facial expressions, e.g. smiling, also seemed to have a negative e ect on tracking. Another pos- sibility is that the xed parameters used in the search did not generalize well to some speakers' facial structure. To obtain better tracking in such cases, the search area for the mouth in the lower region of the face was adjusted manually, as was the relative size of the mouth rectangle.

With these measures, most of the remaining tracking fail- ures were in the rst few frames of the recording, before the speaker started reading the sentence. The nal tracking re- sults, consisting of a 100x70 pixel rectangle centered on the mouth in each frame, were saved to a separate le in raw AVI format.

Bibliography

[1] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in HMM-based ASR," in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, pp. 461-471, 1996.

[2] J. Bilmes, et. al. "Discriminatively Structured Graphical Models for Speech Recognition," Johns Hopkins University, CSLP 2001 Summer Workshop Final Report.

[3] S. Boll, "Speech enhancement in the 1980s: noise suppression with pattern matching," In Advances in Speech Signal Processing, pp. 309-325, Dekker, 1992.

[4] C. Bregler and Y. Konig, "Eigenlips for Robust Speech Recognition," In Proc. ICASSP, 1994.

[5] M. Chan, Y. Zhang, and T. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in Proc. Works. Multimedia Signal Processing, pp. 65-70, Redondo Beach, CA, 1998.

[6] C. Chang and C. Lin, LIBSVM: A Library For Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[7] N. Chomsky and M. Halle, The Sound Pattern of English, Harper and Row, New York, 1968.

[8] S. Chu and T. Huang, "Bimodal speech recognition using coupled hidden Markov models," In Proc. Int. Conf. Spoken Lang. Processing, vol. II, Beijing, China, pp. 747-750, 2000.

[9] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," In Proc. Europ. Conf. Computer Vision, Germany, pp. 484-498, 1998.

[10] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and application," Computer Vision Image Understanding, vol. 61, no. 1, pp. 38-59, 1995.

[11] J. Danhauer and S. Singh, "Multidimensional Speech Perception by the Hearing Impared," University Park Press, Baltimore, MD, pp. 38-44, 1975.

[12] L. Deng, "Speech Recognition Using Autosegmental Representation of Phonological Units with Interface to the Trended HMM," Speech Communication, 23, 3, pp. 211-222, 1997.

[13] L. Deng and D. Sun, "A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features," Journal of the Acoustical Society of America, 95, 5, May 1994.

[14] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Trans. Multimedia, vol. 2, no. 3, pp. 141-151, 2000.

[15] G. Fant, Acoustic Theory of Speech Production, Netherlands: Mouton and Co., 1960.

[16] E. Fosler-Lussier, S. Greenberg, and N. Morgan, "Incorporating contextual phonetics into automatic speech recognition," Proc. Int. Congress of Phonetic Sciences, San Francisco, CA, 1999.

[17] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine based dynamic network for visual speech recognition applications," EURASIP J. Appl. Signal Processing, vol. 2002, no. 11, pp. 1248-1259, 2002.

[18] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," Proc. Human Language Technology Conference, San Diego, 2002.

[19] S. Gurbuz, Z. Tufekci, E. Patterson, and J. Gowdy, "Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition," in Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 177-180, Salt Lake City, UT, 2001.

[20] IBM Research - Audio Visual Speech Technologies: ViaVoice Data Collection. Available at http://www.research.ibm.com/AVSTG/data.html.

[21] R. Jakobson, C. Fant and M. Halle, "Preliminaries to speech analysis. The distinctive features and their correlates," Acoustic Laboratory, MIT, Technical Report No. 13, 1952.

[22] M. Jones and T. Poggio, "Multidimensional morphable models: A framework for representing and matching object classes," In Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 1998.

[23] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," Proceedings of International Joint Conference on Neural Networks, Portland, Oregan, 2003.

[24] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," Int. J. Computer Vision, vol. 1, no. 4, pp. 321-331, 1988.

[25] S. Keyser and K. Stevens. "Feature geometry and the vocal tract," Phonology 11, no 2, pp. 207-236, 1994.

[26] S. King, T. Stephenson, S. Isard, P. Taylor and A. Strachan, "Speech recognition via phonetically featured syllables," In Proc. ICSLP, Sydney, 1998.

[27] K. Kirchhoff, G. Fink and G. Sagerer, "Combining Acoustic and Articulatory-feature Information for Robust Speech Recognition," In Proc. ICSLP, pp. 891-894, Sydney, 1998.

[28] J. Koreman, B. Andreeva, and W. Barry, "Do Phonetic Features Help to Improve Consonant Identification in ASR?" In Proceedings of ICSLP, Sydney, Australia, December 1998, vol. 3, pp. 1035-1038.

[29] G. Krone, B. Talle, A. Wichert, and G. Palm, "Neural architectures for sensor fusion in speech recognition," In Proc. Europ. Tut. Works. Audio-Visual Speech Processing, pp. 57-60, Greece, 1997.

[30] K. Livescu and J. Glass, "Feature-based Pronunciation Modeling for Speech Recognition," In Proc. HLT/NAACL, Boston, 2004.

[31] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature modeling for speech recognition using dynamic Bayesian networks." Proc. EUROSPEECH, Geneva, Switzerland, August-September 2003.

[32] B. Mak and E. Barnard, "Phone clustering using the bhattacharyya distance." In ICSLP96, The Fourth International Conference on Spoken Language Processing, volume 4, pages 20052008, 1996.

[33] K. Mase and A. Pentland, "Automatic Lipreading by optical flow analysis," Systems and Computers in Japan, vol. 22, no. 6, pp. 67-76, 1991.

[34] D. Massaro and M. Cohen, "Perceiving asynchronous bimodal speech in consonant vowel and vowel syllables, Speech Commun., vol. 13, pp. 127134, 1993.

[35] I. Matthews, J. A. Bangham, and S. Cox, "Audio-visual speech recognition using multiscale nonlinear image decomposition," Proc. Int. Conf. Spoken Lang. Processing, Philadelphia, PA, pp. 38-41, 1996.

[36] I. Matthews, T. Cootes, A. Bangham, S. Cox and R. Harvey, "Extraction of Visual Features for Lipreading," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, 2002.

[37] F. Metze and A. Waibel, "A Flexible Stream Architecture for ASR Using Articulatory Features," In Proc. ICSLP, Denver, 2002.

[38] G. Miller and P. Nicely, "An Analysis of Perceptual Confusions among some English Consonants," J. Acoustical Society America, vol. 27, no. 2, pp. 338-352, 1955.

[39] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop," In Proc. Works. Signal Processing, pp. 619-624, Cannes, France, 2001.

[40] L. Ng, G. Burnett, J. Holzrichter, and T. Gable, "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing," In Proc. ICASSP, Istanbul, Turkey, 2000.

[41] P. Niyogi, "Towards a Computational Model of Human Speech Perception," in Proc. From Sound to Sense: 50+ Years of Discoveries in Speech Communication, Boston, 2004.

[42] P. Niyogi, E. Petajan, and J. Zhong, "Feature Based Representation for Audio-Visual Speech Recognition", Proceedings of the Audio Visual Speech Conference, Santa Cruz, CA, 1999.

[43] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in Proc. Int. Conf. Acoust., Speech, Signal Processing, Orlando, FL, pp. 2017-2020, 2002.

[44] E. Petajan, "Automatic lipreading to enhance speech recognition," In Proc. Global Telecomm. Conf., pp. 265-272, Atlanta, GA, 1984.

[45] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," Audio- and Video-based Biometric Person Authentication, Germany, 1997.

[46] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," In Proc. Eur. Conf. Speech Comm. Tech., pp. 1293-1296, Geneva, 2003.

[47] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", In Proc. IEEE, 2003.

[48] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A Cascade Image Transform for Speaker-Independent Automatic Speechreading," In Proc. ICME, volume II, pp. 1097-1100, New York, 2000.

[49] C. Sanderson, "The VidTIMIT Database," IDIAP Communication 02-06, Martigny, Switzerland, 2002.

[50] C. Sanderson, "Automatic Person Verification Using Speech and Face Information," PhD Thesis, Griffith University, Brisbane, Australia, 2002.

[51] P. Smeele et al., "Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position, in Proc. Int. Conf. Spoken Language Processing, Alberta, Canada, pp. 6568, 1992.

[52] K. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," in Journal of Acoustical Society of America, vol. 111, no. 4, pp. 1872-1891, 2002.

[53] W. Sumby, and I. Pollack, "Visual contribution to speech intelligibility in noise," J. Acoustical Society America, vol. 26, no. 2, pp. 212-215, 1954.

[54] J. Sun and L. Deng, "An Overlapping-Feature Based Phonological Model Incorporating Linguistic Constraints: Applications to Speech Recognition", J. Acoustic Society of America, vol. 111, No. 2, pp. 1086-1101, 2002.

[55] M. Tang, S. Seneff, and V. Zue, "Modeling Linguistic Features in Speech Recognition," Proc. Eurospeech, Geneva, Switzerland, 2003.

[56] P. Teissier, J. Robert-Ribes, and J. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," IEEE Trans. Speech Audio Processing, vol. 7, no. 6, pp. 629-642, 1999.

[57] V. Vapnik, Statistical Learning Theory, J. Wiley, New York, 1998.

[58] X. Zhang, R. Mersereau, M. Clements and C. Broun, "Visual Speech Feature Extraction for Improved Speech Recognition," In Proc. ICASSP, Orlando, May 2002.