TANDEM APPROACH FOR INFORMATION FUSION IN AUDIO VISUAL SPEECH RECOGNITION

by

HARUN KARABALKAN

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
February 2009

TANDEM APPROACH FOR INFORMATION FUSION IN AUDIO VISUAL SPEECH RECOGNITION

APPROVED BY

Assist. Prof. Dr. Hakan ERDOĞAN .....
(Thesis Supervisor)

Prof. Dr. Aytül ERÇİL .....

Assoc. Prof. Dr. Berrin YANIKOĞLU .....

Assist. Prof. Dr. Müjdat ÇETİN .....

Assist. Prof. Dr. Murat SARAÇLAR .....

DATE OF APPROVAL: .....

©Harun Karabalkan 2009
All Rights Reserved

to my family

Acknowledgements

I would like to express my gratitude to my thesis supervisor Hakan Erdoğan for his invaluable guidance, support and encouragement throughout my thesis.
I would like to thank TÜBİTAK for providing the necessary financial support for my masters education.
I am also very grateful to the members of Vision and Pattern Analysis Laboratory of Sabanci University for their friendship. Last but not the least, I would like to thank my thesis jury members Aytül Erçil, Berrin Yanıkoğlu, Müjdat Çetin and Murat Saraçlar.

TANDEM APPROACH FOR INFORMATION FUSION IN AUDIO VISUAL SPEECH RECOGNITION

HARUN KARABALKAN

EE, M.Sc. Thesis, 2009

Thesis Supervisor: Hakan Erdoğan

Keywords: speech recognition, audiovisual, multimodality

Abstract

Speech is the most frequently preferred medium for humans to interact with their environment making it an ideal instrument for human-computer interfaces. However, for the speech recognition systems to be more prevalent in real life applications, high recognition accuracy together with speaker independency and robustness to hostile conditions is necessary.

One of the main preoccupation for speech recognition systems is acoustic noise. Audio Visual Speech Recognition systems intend to overcome the noise problem utilizing visual speech information generally extracted from the face or in particular the lip region. Visual speech information is known to be a complementary source for speech perception and is not impacted by acoustic noise. This advantage brings in two additional issues into the task which are visual feature extraction and information fusion.

There is extensive research on both issues but an admissable level of success has not been reached yet. This work concentrates on the issue of information fusion and proposes a novel methodology. The aim of the proposed technique is to deploy a preliminary decision stage at frame level as an initial stage and feed the Hidden Markov Model with the output posterior probabilities as in tandem HMM approach. First, classification is performed for each modality separately. Sequentially, the individual classifiers of each modality are combined to obtain posterior probability

vectors corresponding to each speech frame. The purpose of using a preliminary stage is to integrate acoustic and visual data for maximum class separability. Hidden Markov Models are employed as the second stage of modelling because of their ability to handle temporal evolutions of data.

The proposed approach is investigated in a speaker independent scenario for digit recognition with the existence of diverse levels of car noise. The method is compared with a principal information fusion framework in audio visual speech recognition which is Multiple Stream Hidden Markov Models (MSHMM). The results on M2VTS database show that the novel method achieves resembling performance with less processing time as compared to MSHMM.

GÖRSEL-İŞİTSEL KONUŞMA TANIMA'DA ARDIŞIK VERİ KAYNAŞTIRMA YAKLAŞIMI

HARUN KARABALKAN

EE, Yüksek Lisans Tezi, 2009

Tez Danışmanı: Hakan Erdoğan

Anahtar Kelimeler: konuşma tanıma, görsel-işitsel, çok kiplilik

Özet

İnsanların çevresiyle etkileşiminde en çok tercih ettiği araçların başında ses ve konuşma gelir. Bu durum, konuşma tanıma sistemlerini gelecekteki insan-bilgisayar arayüzlerinin vazgeçilmez bir parçası haline getirmektedir. Ancak, konuşma tanıma sistemlerinin gerçek hayatta uygulanabilir olması için çevresel gürültüden etkilenmeden yüksek tanıma oranlarına ulaşabilir olması gerekmektedir. Görsel-İşitsel Konuşma Tanıma Sistemleri, işitsel gürültünün olumsuz etkilerini en aza indirgemek için dudak hareketlerinden elde edilen görsel konuşma bilgisini kullanmaktadır. Görsel bilginin sisteme dahil edilmesinin sebebi, konuşma tanımada görsel bilginin işitsel bilgiyi bütünleyici bir bilgi kaynağı olması ve işitsel gürültüden etkilenmemesidir. Bu avantaj ile birlikte sistem tasarımı açısından iki yeni husus ortaya çıkmaktadır. Hususlardan ilki, görsel öznitelik çıkarımı, diğeri ise görsel ve işitsel bilginin kaynaştırılmasıdır. Bu çalışma, görsel ve işitsel bilginin kaynaştırılması problemine odaklanmakta ve özgün bir görsel-işitsel konuşma tanıma sistemi önermektedir.

Önerilen yöntemde, her iki bilgi akımı için ayrı ayrı sınıflandırıcılar eğitilmekte ve daha sonra bu sınıflandırıcılar bir birleştirici sınıflandırıcısı ile birleştirilmektedir. Böylece, görsel ve işitsel bilgi kaynaştırılmış olmaktadır. Birleştirici sınıflandırıcısının çıktısı olan sonsal olasılık vektörleri ise Saklı Markov Modelleri için gözlem vektörleri olarak kullanılmaktadır.

Önerilen yaklaşım ile tasarlanan kişiden bağımsız rakam tanıma sistemi, değişen seviyelerde araba gürültüsünün mevcut olduğu koşullarda test edilmektedir. Yeni yöntem, şu ana dek önerilmiş en başarılı görsel-işitsel konuşma tanıma sistemlerinden biri olarak kabul edilen Çok Akımlı Saklı Markov Modeli (ÇASMM) ile tanıma oranı ve hız açısından karşılaştırılmaktadır. Deneysel sonuçlar göstermektedir ki, yeni yöntem daha az işlem yükuÿle ÇASMM yöntemine yakın tanıma oranlarına ulaşmaktadır.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Speech is the most frequently preferred medium for humans to interact with their environment. Hence, speech recognition systems are promising candidates for future human-computer interfaces. However, for a real life application, speech recognition technology must offer high recognition accuracy as well as high degree of robustness against all kinds of degrading circumstances. The main difficulty for a prosperous speech recognition system is the acoustic noise which is almost always present in real life applications.

Although, some speech recognition systems exhibit high recognition rates in situations where there is no acoustic noise, their performances degrade dramatically with decreasing Signal-to-Noise Ratio (SNR). A solution to the problem lies in the psychophysics of human perception of speech. It is demonstrated by McGurk that humans integrate visual speech information generally obtained from the lip region with the acoustic information in order to recognize speech [1]. McGurk's work also claims that visual speech information is not a secondary source for speech perception, instead it is complementary to acoustic information. This phenomenon gives the motivation to include visual information in speech recognition systems especially when the recognition systems are impacted by environmental noise.

On the other hand, the idea of using visual speech information brings in two additional issues to speech recognition. First issue is to discover the most appropriate visual feature extraction scheme. Second issue is to determine the visual information integration procedure. To date, there is no such visual feature that has found common acceptance as the most appropriate feature set but there are

some techniques that most of the work is concentrated on. Beyond that, fusion of the information from the two modalities remains as the main focus of improvement. Researchers intuitively propose statistical information fusion methodologies for audio visual speech recognition but their performances have not yet reached an admissable level. This work intends to contribute to such progress of information fusion for audio visual speech recognition systems proposing a novel methodology which is fast and easy to implement.

Information fusion methods in audio visual speech recognition can be categorized in three main groups. The first group is *Feature Fusion* or *Early Fusion* in which the features from the two modalities are concatenated to form a combined feature vector and the combined feature vector is fed into a Hidden Markov Model (HMM) as an observation. The second group is *Decision Fusion* or *Late Fusion* in which the features from different streams are separately modelled with HMMs and a final decision is made by combining the decisions according to a designated rule. The third group is *Model Fusion* in which the features from the two modalities are modelled in a parallel structure with HMM. The primary model fusion technique is *Multiple Stream Hidden Markov Model (MSHMM)* with more advanced versions such as *Product HMM*, *Factorial HMM* and *Coupled HMM*. A novel framework is proposed in this thesis as the fourth category of audio visual information fusion techniques, named *Tandem Fusion*. The novel approach has grounds both in Feature Fusion and Decision Fusion and is based on employing a preliminary decision stage before HMM training.

1.2 Literature Review

The benefit of visual information for speech recognition is first investigated by Sumby and Pollack who conducted speech intelligibility tests with and without visual information and compared the two cases [2]. McGurk demonstrated that incorrect visual information can cause humans to perceive the true utterance wrong and concluded that visual information is in fact complementary to the acoustic information [1]. This phenomenon is called the *McGurk Effect*.

McGurk effect has been the primary motivation for audio visual speech recognition research introducing the visual feature extraction and the information fusion

issues into the problem. The information fusion problem is addressed in many works, this thesis being one. Petajan was the first to create an audio visual speech recognition system [3]. In that first system, visual information is used to select one of the best two candidates from the audio based recognizer to give the final decision for the spoken word. Tomlinson et. al. [4] focused on the problem of information fusion in terms of feature concatenation where the feature vectors from the two streams are concatenated to train a single HMM. Tomlinson observed improved performance compared to the audio-only speech recognition systems [4]. Since the concatenation of the two feature vectors results in a high dimensional feature vector, Potamianos et. al. [5] applied Linear Discriminant Analysis to the combined feature vectors for dimensionality reduction before feeding the feature vectors to the HMMs.

Adjoudani and Benoit et. al. [6] and Teissier et. al. [7] compared the decision fusion and the feature fusion techniques to conclude that decision fusion achieves higher recognition accuracy. Both Adjoudani and Teissier trained audio-only and video-only HMMs and then linearly combined the log-likelihoods of the two streams adjusting weights for each.

Multiple Stream HMM, in which the two streams are independently modelled, is investigated by Dupont for audio visual speech recognition. This paper reports the superior performance of MSHMM compared to both feature fusion and decision techniques [8]. MSHMM is accepted as one of the most successful audio visual information fusion methodologies [8, 9]. A drawback of MSHMM is the restriction of the audio and visual streams to be state synchronous so that a transition from a state to another takes place at the same time. This is not a desirable situation since the visual information can sometimes precede the acoustic information, i.e., the lip movement can occur before the speech is produced. Product HMM (PHMM), which is an extension of MSHMM, allows state asynchrony between the two streams forcing the streams to be synchronous at the phoneme boundaries [10]. There are also more advanced HMMs utilized in audio visual speech recognition which include Factorial HMM (FHMM) and the Coupled HMM (CHMM). In FHMM, the audio and visual states are independent of each other, but they jointly model the likelihood of the audiovisual observation vector, and hence become correlated indirectly [9]. In CHMM, the likelihoods of the audio and visual observation vectors are modeled independent

of each other, but each of the audio and visual states are conditioned jointly by the previous set of audio and visual states [11]. The performances of MSHMM, PHMM, FHMM and CHMM are compared by Nefian [9]. The results in that work showed that PHMM and FHMM do not improve the recognition rate compared to MSHMM and CHMM outperforms MSHMM by absolute 2% approximately.

A novel information fusion framework is proposed in this work which is based on *tandem feature extraction* method of Hermansky [12]. The idea of *tandem feature extraction* is driven from the idea of *Hybrid HMM*. The conventional HMMs generate observations from a Gaussian Mixture distribution but there is some work that replaces Gaussian Mixture Model (GMM) by a more discriminative model taking the name *Hybrid HMM*. To date, Neural Networks (NN) and Support Vector Machines (SVM) are used in hybrid HMM structures. A hybrid NN/HMM system showed superior performance compared to a conventional HMM system in [13]. However, it is stated by Bourlard, Morgan and their partners in Wernicke Project, that hybrid approaches employing neural networks are computationally very expensive and traning neural network parameters for speech recognition in standard workstations is very impractical, nearly impossible [14]. Similar to NN/HMM hybrid system, SVM/HMM hybrid architecture is proposed and analysed for several acoustic speech recognition tasks by Ganapathiraju in a series of papers [15, 16, 17] reporting improved performance of the hybrid system compared to a conventional HMM system.

Garcia-Moral compared the performance of a neural network based hybrid system with an SVM based hybrid system and concluded that they exhibit resembling performance [18]. Gordan et. al. [19] and Krüger et. al. [20] implemented an SVM/HMM hybrid method for visual speech recognition referencing Ganapathiraju's work. The idea of hybrid SVM/HMM is also applied to audio visual speech recognition by Gurban et. al. [21] where one-versus-rest SVMs are trained for each modality and the outputs of the two modalities are combined with the product rule.

Deficiency of GMM in hybrid approaches led Hermansky et. al. [12] to combine NN processing with GMM modelling. Hermansky used NN to obtain posterior probabilities and the posterior probabilities are fed into a conventional HMM to report 50% improvement compared to the baseline system. The idea is that the

classifier posteriors are more discriminative features as compared to regular features for HMMs. The approach employing a classification stage before the HMM stage is named as the *Tandem Approach*. Hagen and Morris made a comprehensive analysis of the tandem approach by testing it on multistream audio data and declared that the tandem approach performs better than the conventional HMM systems. In their work, a stream corresponded to a different audio feature set [22]. The posterior probabilities from each stream are concatenated by means of Principal Component Analysis (PCA) and the concatenated posterior probability vectors are used as observations for the GMM based HMM system.

1.3 Contributions

This work extends the idea of tandem approach in single modality tasks to audio visual speech recognition task, proposing a novel methodology for information fusion to improve recognition performance. The new method is investigated in a speaker independent scenario for digit recognition with the existence of diverse levels of car noise. Its performance is compared with the performance of Multiple Stream Hidden Markov Models in terms of accuracy and processing speed to conclude that the new approach achieves a resembling performance with less processing time.

1.4 Outline

This thesis is organized in five chapters including the Introduction chapter. In Chapter 2, audio and visual feature extraction techniques and Hidden Markov Modelling are discussed. The proposed audio visual information fusion framework and the conventional information fusion methods are described in Chapter 3. The experimental results are investigated in Chapter 4. Finally, the conclusions and future work are expressed in Chapter 5.

Chapter 2

Background

Automatic Speech Recognition (ASR) systems with single data stream consist of two main stages diagrammed in Figure 2.1. First stage is the signal analysis stage in which the input signal is converted to a sequence of feature vectors. The input signal can either be acoustic or visual. There are various audio and visual feature extraction techniques proposed in the literature. The most common methods will be discussed and their performances will be compared in this work. The second stage of ASR systems is the modelling stage in which a model is trained for each specified class using the feature vectors in the training dataset. Hidden Markov Models (HMM) have been the primary tool for speech modelling since their first application to speech recognition by Baker [23] and Jelinek [24]. A class in HMM can be a word or a phoneme which is the basic structural unit that distinguishes meaning. The words in the dataset used in this study, digits in French from zero to nine, and their phonetic contents are given in Table 2.1 (phonetic contents of French digits are provided by Guillaume Gravier).

The organization of the chapter is as follows: In section 2.1.2, audio feature extraction procedure and the most common audio feature extraction techniques are introduced. Visual feature extraction procedure and the most common visual feature extraction techniques are described in section 2.2. Section 2.3 is the final section

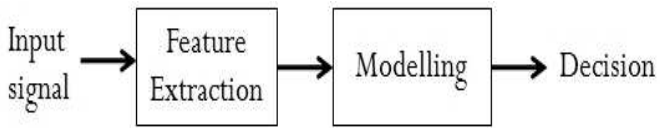


Figure 2.1: Single Stream ASR Framework

Word	Phonemes			
zero	z	e	R	0
un	U			
deux	d	2		
trois	t	R	w	a
quatre	k	a	t	R
cinq	s	U	k	
six	s	i	s	
sept	s	E	t	
huit	H	i	t	
neuf	n	9	f	

Table 2.1: The Phonetic Contents of the Words in the Dataset (%)

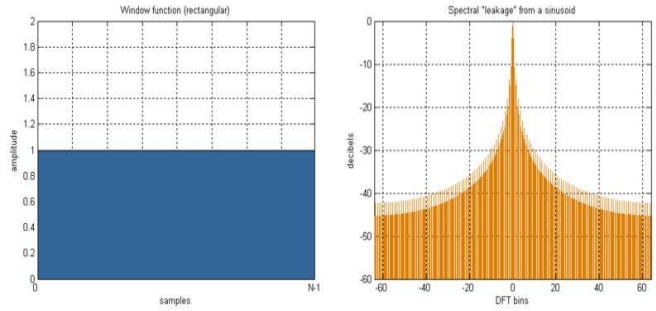
covering a brief introduction to Hidden Markov Models.

2.1 Audio Feature Extraction

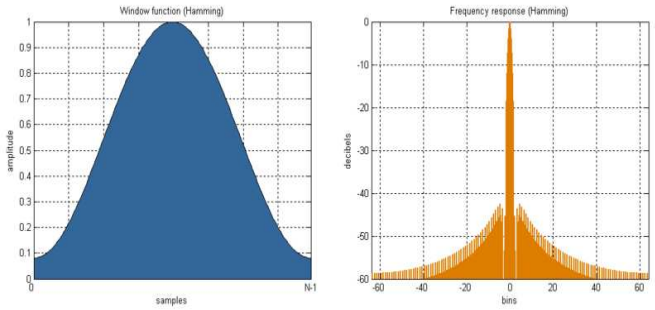
The first step in a statistical speech recognition system is to convert speech waveform into a stream of feature vectors. Feature vectors are parametric representations of speech to classify different acoustic units. Audio feature extraction can be analysed in three stages. First stage is the windowing stage in which the speech signal is divided into short time segments called *frames* to carry out short-time analysis of speech. In the second stage, the static features are extracted from each speech frame. In the last stage, dynamic features are extracted using the static features of the consecutive frames to model the dynamic nature of the speech signal.

2.1.1 Windowing

Since speech is a dynamic signal, the analysis is carried on short time segments called *frames*. The frame duration has to be chosen such that the set of parameters representing that segment are almost constant throughout the segment. The typical frame length is 25ms and overlapping frames are extracted at a frame rate of 10ms. Windows are overlapped to deal with window artifacts. Extracting a short time



(a)



(b)

Figure 2.2: Rectangular window vs. Hamming window

segment is equivalent to applying sharp rectangular window to the signal but since the Fourier Transform of a rectangular signal is *sinc function*, its spectrum has a curved main lobe and large amount of ripple in the stop band which introduces spectral distortion. Hamming window is used instead in almost all recognition systems because it has a flatter pass band, and less ripple in the stop band compared to the rectangular window as can be seen in Figure 2.2. The Hamming window function is

$$w(n) \;=\; 0.53836 - 0.46164\cos(\frac{2\pi\, n}{N-1}), \tag{2.1}$$

where N is the total number of samples in a window and $0 \leq\ n \leq\ N-1$.

2.1.2 Audio Feature Extraction Methods

Three of the most commonly preferred audio feature types in the literature are Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Predictive (PLP) Coefficients and Advanced Front End (AFE). MFCC, which is issued as standard

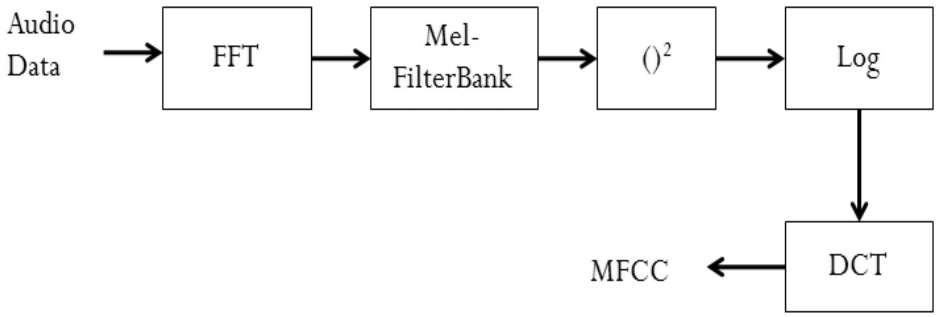


Figure 2.3: MFCC Extraction Scheme

audio feature for speech processing by European Telecommunications Standard Institute (ETSI) in 2000 [25], is the most popular among the three. AFE is also an ETSI standard which is an extended version of MFCC employing a noise reduction scheme [26]. PLP, although not issued as a standard feature type, is a competing feature against MFCC and can perform better depending on the application. All three feature types are described and experimantally analysed in this work. The best performing feature type on our database is selected to be used as the audio feature in the audio visual scenario.

Mel Frequency Cepstral Coefficients

Cepstral analysis is a way of representing the spectral envelope of a speech frame by performing a transform to the logarithm of the power spectrum. This concept was first introduced by Bogert et al. [27] in 1963 and it provides the information to discriminate between different phonetic units. Mel Frequency Cepstral Coefficients (MFCC) are derived by cepstral analysis. The MFCC feature extraction scheme is diagrammed in Figure 2.3.

First of all, the spectrum of a speech frame is obtained by applying a Fourier Transform to the input signal. Secondly, the spectrum is segmented into critical bands by applying a Mel-filterbank to the spectrum. The Mel-filterbank consists of overlapping triangular filters with center frequencies in Mel-scale determined by equation (2.2).

$$f_{mel} = 2595 \times \log(1 + f/700) \tag{2.2}$$

Figure 2.4 shows the mapping of original frequency to Mel scale. Mapping of the original frequency axis to the Mel-scale is essential to model the nonlinear spectral

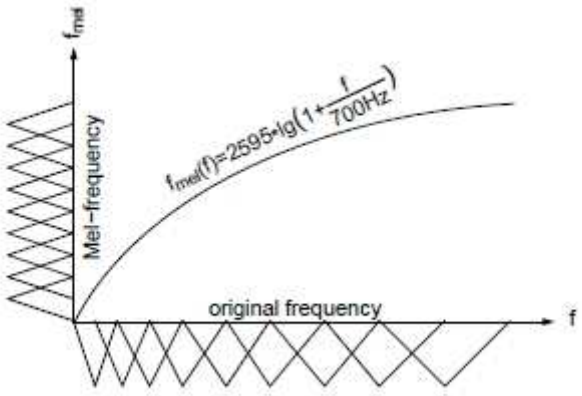


Figure 2.4: Mel frequency scale

resolution of human auditory system along the frequency axis. For instance, humans can easily discriminate between tones of 200Hz and 250Hz, however they can not discriminate between tones of 2000Hz and 2050Hz.

As the next step following the Mel-filterbank, the energy in each of the triangular filter is calculated and consequently logarithm is applied to theses energy terms. Finally, Discrete Cosine Transform (DCT) is performed on the log-energy terms as if they are the samples of a time domain signal. The resulting DCT coefficients are the MFCCs. Generally, 12 lowest order coefficients are used together with the energy coefficient which add up to 13 static features for each speech frame. The number of MFCCs used determines the precision to represent the spectra.

Perceptual Linear Predictive Coefficients

The perceptual linear predictive (PLP) coefficients, proposed by Hermansky et al. [28], are derived by linear predictive analysis of a specially modified, short-term speech spectrum. In PLP analysis, the speech spectrum is modified by a set of transformations that are based on models of the human auditory system. To be more precise, the following three concepts from the psychophysics of hearing are applied to derive an auditory spectrum estimate:

- The critical-band spectral resolution

- The equal-loudness curve
- The intensity-loudness power law.

The spectral resolution of human hearing is roughly linear up to 800Hz - 1000Hz but it decreases with increasing frequency above this frequency range. PLP remaps the frequency axis to the Bark scale [29] by the equation

$$\Omega(\omega) = 6\ln\{w/1200\pi + \sqrt{(w/1200)^2 + 1}\}, \tag{2.3}$$

where w is the original frequency and $\Omega(w)$ is the corresponding frequency in the Bark-scale. The Bark-scaled spectrum is convolved with the power spectrum of the critical band filter given in equation (2.4) to find the critical band spectrum approximation.

$$\Phi(\omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 < \Omega < -0.5 \\ 1 & -0.5 < \Omega < 0.5 \\ 10^{-1(\Omega-0.5)} & 0.5 < \Omega < 2.5 \\ 0 & \Omega > 2.5 \end{cases} \tag{2.4}$$

Also, at conversational speech levels, human hearing is more sensitive to the middle frequency range of the audible spectrum. PLP incorporates the effect of this phenomenon by multiplying the critical-band spectrum by an equal loudness curve defined by equation (2.5), that suppresses both the low and high frequency regions relative to the midrange from 400 to 1200 Hz.

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)\omega^4}{(\omega^2 + 6.3 \times 10^6)(\omega^2 + 0.38 \times 10^9)(\omega^6 + 9.58 \times 10^{26})} \tag{2.5}$$

In addition, there is a nonlinear relationship between the intensity of sound and the perceived loudness. PLP approximates the power law of hearing by using a cube-root amplitude compression of the loudness equalized, critical band spectrum estimate using the equation.

$$L(\omega) = I(\omega)^{1/3}, \tag{2.6}$$

where $L(\omega)$ is the perceived loudness and $I(\omega)$ is the intensity of the sound.

Once auditory-like power spectrum is estimated after the three transformations stated above, Inverse Discrete Fourier Transform (IDFT) is applied to the power

spectrum. The outputs of the IDFT are used as inputs to a Linear Prediction routine. Linear Prediction or Linear Predictive Coding (LPC) is a discrete time signal analysis tool that estimates a future sample value of a signal by a linear combination of the previous samples, mathematically defined by

$$\widehat{x} = \sum_{i=1}^{p} a_i x[n-i], \tag{2.7}$$

where $\widehat{x}$ is the estimate of the sample $x[n]$ at time n, a_i is i'th LPC coefficient and p is the order of LPC, i.e., the number of previous samples used to estimate a future sample value. The LPC coefficients are estimated to minimize the sum of the squared error in a finite length speech frame mathematically expressed as

$$\begin{aligned} E &= \textstyle\sum_{i=1}^{N-1} e[n]^2 \\ &= \textstyle\sum_{i=1}^{N-1} (x[n] - \sum_{i=1}^{p} a_i x[n-i])^2, \end{aligned} \tag{2.8}$$

where N is total number of samples in the frame. LPC features can be used to obtain an envelope to the spectrum of the input signal. The estimated LPC coefficients are the PLP features. Optionally, LPC coefficients can be converted to cepstral coefficients through cepstral analysis as explained in section 2.1.2 to get the PLP features.

PLP coefficients are said to be more robust against the differences between training and testing data and they also seem to be more stable in terms of parametrization settings against MFCC [30]. On the other hand MFCCs are considered to be more effective for clean conditions.

Advanced Front End (AFE)

Advanced Front End (AFE) is an extended version of MFCC extraction which is issued as an ETSI standard in 2002 [26]. In AFE, MFCC extraction is preceded by a two-stage Wiener filtering for noise reduction which provides improved recognition performance but also brings three times more computational load [31]. Noise reduction is an extensive research area in speech recognition and is beyond the scope of this thesis. The details of two-stage Wiener filtering can be found in [32].

2.1.3 Dynamic Information

Trajectory of parameters along the consecutive frames carry essential information about the speech to be recognised. Thus, first and second derivatives in time which are called delta coefficients and acceleration coefficients respectively, are extracted from the static features. The delta coefficients are computed using the formula given in equation 2.9 where d_t is the delta coefficient at time t and $c_{t+\theta}$ and $c_{t-\theta}$ are the corresponding static coefficients.

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.9}$$

The value Θ is the number of consecutive frames over which the derivation is applied with a reasonable value ranging from 2 to 5. Accelaration coefficients are computed similarly by applying equation (2.9) to delta coefficients.

2.2 Visual Feature Extraction

As stated in Chapter 1, visual information is complementary to the acoustic information for speech recognition and it is not impacted by acoustic noise. Hence, it has the potential to boost the recognition performance of ASR systems. The idea of utilizing the visual information brings in the visual feature extraction issue into the speech recognition problem. Although there are no standardized techniques for visual feature extraction as in the case of audio feature extraction, there are particular methods that most researchers concentrate on. Mainly, we can classify visual feature extraction methods into three categories:

- Region (or appearance) based visual features
- Lip contour based visual features
- Combination of region and contour based visual features

Lip contour based visual features can further be divided into two categories:

- Geometric visual features
- Lip shape model visual features

Geometric visual features are features giving information about the aperture of the mouth such as width and height of the mouth, the aperture angle or the area of the aperture. Visual features based on lip shape models are the parameters of the parametric or statistical model of the lip contour. Both geometric and shape model based visual features substantially rely on a preprocess which is the tracking of lip movements. Unfortunately, only a minor deviation in tracking could result in a major inaccuracy in recognition. On the other hand, appearance based visual features do not necessitate such precision for recognition accuracy. This makes the appearance based visual features preferrable in most audio-visual speech recognition architectures.

Appearance based visual features rely on the pixel values, either grayscale or colored, of the region of interest. However, the dimensionality constitutes a problem in statistical analysis. Therefore, various transformations are used to obtain visual features of admissible dimension. Dimensionality reduction does not only offer efficient computation but also helps to reduce speaker dependency of the recognition system due to the nature of these transformations. The transformations which will be analysed in this work are Discrete Cosine Transform (DCT), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

2.2.1 Region of Interest (ROI) Extraction

Prior to feature extraction, a region of interest (ROI) has to be obtained which directly affects the performance of the overall system. The ROI is typically a rectangle enclosing the mouth including the nose tip and the chin. In this study, lip region is assumed to be in the lower 40% of the face vertically and central 50% of it horizontally. The face is detected using Viola and Jones's method of visual object detection [33]. The correlation between the consecutive frames is used to fix the central point of the mouth and suppress the interframe vibrations. An example face image and the lip region extracted from that face image can be seen in Figure 2.5.

2.2.2 Visual Feature Extraction Methods

Three most commonly preferred appearance based visual feature extraction methods are analysed in this work which are Discrete Cosine Transform (DCT), Principal

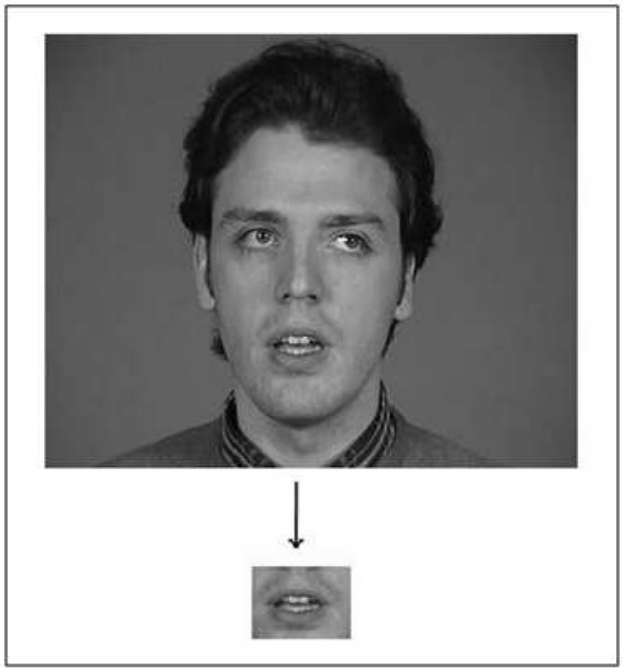


Figure 2.5: Region of Interest Extraction

Component Analysis (PCA) and Linear Discriminant Analysis (LDA). All three methods apply dimensionality reduction on the grayscale ROI to reduce the the original dimension M (number of pixels) to L where $L < M$.

Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT) is widely used in visual feature extraction as well as image compression. Potamianos et al. [34] was the first to use DCT in visual speech recognition and concluded that DCT outperforms the lip contour based visual features. DCT's coherence is also analysed in other related work [35, 36] and its popularity depends on three facts. First, DCT has a strong *energy compaction* property so that most of the signal information is concentrated in a few low frequency components. Second, it has a fast implementation which is an advantage in real time processing. Third, it requires no training data. In this work, we perform two dimensional DCT on the lip region image and pick L low frequency components and use them as visual features.

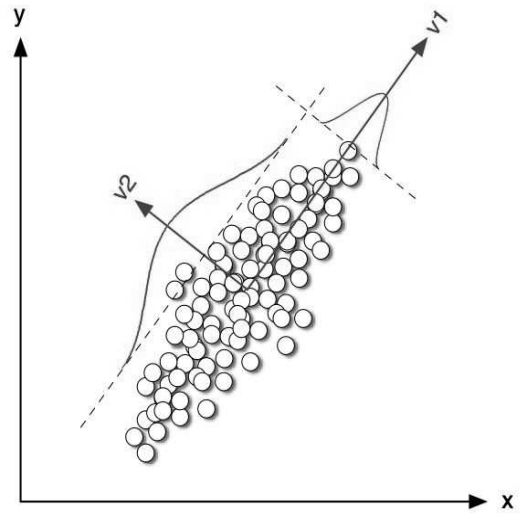


Figure 2.6: Principal Components for 2-dimensional Feature Set

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate called the first principal component, the second greatest variance on the second coordinate, and so on. Figure 2.6 shows the principal components $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ for a two-dimensional feature set.

PCA can be used for dimensionality reduction by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the most important aspects of the data.

Supposing the mouth ROI contains M number of pixels and there are N number of video frames in the training set, a mean subtracted data matrix

$$\boldsymbol{X} = [\boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_N] - \boldsymbol{u}\boldsymbol{h}, \quad \boldsymbol{h} = [111 \cdots 1]_{1\times N} \tag{2.10}$$

is created where $\boldsymbol{x}_i$ is the $NM \times 1$ dimensional column vector with the grayscale pixel values of mouth ROI in a frame and $\boldsymbol{u}$ is the $NM \times 1$ dimensional mean vector of the whole data. Next, the covariance matrix of mean subtracted data is calculated by

$$\boldsymbol{C} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T. \tag{2.11}$$

An eigenvalue decomposition is applied to the covariance matrix by the formula

$$\boldsymbol{V}^{-1}\boldsymbol{C}\boldsymbol{V} = \boldsymbol{D}, \tag{2.12}$$

where $\boldsymbol{V}$ is the $M \times M$ square matrix with an eigenvector in each column and $\boldsymbol{D}$ is a diagonal matrix containing the corresponding eigenvalues of eigenvectors. Eigenvectors form an orthogonal basis for the data and the eigenvectors with higher eigenvalues are the most informative. Dimensionality reduction is realized when you keep the eigenvectors with high eigenvalues while ignoring the ones with low eigenvalues. Say, the dimension is to be reduced to L where $1 \leq L \leq M$, then L eigenvectors with the highest corresponding eigenvalues are placed in columns of the transformation matrix $\boldsymbol{W}$ of size $M \times L$. Once the transformation matrix is obtained, M dimensional feature vector $\boldsymbol{x}_i$ can be re-expressed as an L dimensional feature vector $\boldsymbol{y}_i$ by the formula

$$\boldsymbol{y}_i = \boldsymbol{W}^T\boldsymbol{x}_i. \tag{2.13}$$

The elements of $\boldsymbol{y}_i$ are the coefficients of the orthogonal basis vectors.

Linear Discriminant Analysis (LDA)

PCA is an unsupervised technique to describe the data but it is not optimized for class separability and there is no guarantee that the directions of maximum variance will contain good features for dicrimination of classes. Linear Discriminant Analysis (LDA), on the other hand, is a supervised technique seeking solutions for the following questions:

- Which set of features best represent the class association?
- What is the best linear rule for class separation?

In the case of dimension reduction, LDA is not utilized for seeking a class separation rule but for selecting a feature set that best discriminates the data. This is done by searching for basis vectors in the underlying feature space that are most discriminant among classes.

Suppose an M dimensional feature space is to be projected onto an L dimensional feature space where $L << M$ through a projection matrix $\boldsymbol{W}$ of size $M \times L$. Using the labeled training set, two measures are defined which are

- within class scatter matrix given by

$$\boldsymbol{S}_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\boldsymbol{x}_i^j - \boldsymbol{\mu}_j)(\boldsymbol{x}_i^j - \boldsymbol{\mu}_j)^T, \tag{2.14}$$

  where $\boldsymbol{x}_i^j$ is the i'th sample of class j, $\boldsymbol{\mu}_j$ is the mean of class j, c is the number of classes and N_j is the number of samples in class j, and

- between class scatter matrix given by

$$\boldsymbol{S}_b = \sum_{j=1}^{c} (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T, \tag{2.15}$$

  where $\boldsymbol{\mu}$ is the mean of all samples.

The aim is to minimize the within class scatter, $\boldsymbol{S}_w$, and maximize the between class scatter, $\boldsymbol{S}_b$ in the projected space, i.e., maximize the ratio

$$\frac{det(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W})}{det(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W})}. \tag{2.16}$$

If $\boldsymbol{S}_w$ is a nonsingular matrix, the ratio in equation 2.16 is maximized when column vectors of the projection matrix, $\boldsymbol{W}$, are the eigenvectors of

$$\boldsymbol{S}_w^{-1} \boldsymbol{S}_b. \tag{2.17}$$

The eigenvectors of the expression in equation (2.17) are obtained by eigenvalue decomposition which is mathematically defined by the formula

$$\boldsymbol{V}^{-1} \boldsymbol{C} \boldsymbol{V} = \boldsymbol{D}, \tag{2.18}$$

where $\boldsymbol{V}$ is the square matrix with an eigenvector in each column, $\boldsymbol{D}$ is a diagonal matrix containing the corresponding eigenvalues of eigenvectors and $\boldsymbol{C} = \boldsymbol{S}_w^{-1} \cdot \boldsymbol{S}_b$. The eigenvectors are sorted in order of decreasing eigenvalue and L number of eigenvectors are collected as columns of the projection matrix $\boldsymbol{W}$. Once the projection matrix is obtained, every M dimensional feature vector $\boldsymbol{x}$ can be projected onto an L dimensional feature vector $\boldsymbol{y}$ according to the equation

$$\boldsymbol{y} = \boldsymbol{W}^T \boldsymbol{x}. \tag{2.19}$$

There are two issues to consider in implementation of LDA on a M dimensional feature space with c classes:

- there are at most $c-1$ nonzero eigenvectors, so the reduced dimension can be maximum $c-1$ and
- at least M samples are required for each class to guarantee that $\boldsymbol{S}_w$ does not become singular.

Thus, usually other dimension reduction algorithms such as DCT or PCA are applied prior to LDA in order to handle the restrictions stated above.

2.2.3 Dynamic Information and Synchronization

The dymanic information is extracted by means of delta and acceleration coefficients as in the case of audio feature extraction in section 2.1.3. However, there is an additional step to take in visual feature extraction which differs from audio feature extraction.

In visual feature extraction, a feature vector is generated for each video frame. Considering that the videos used in this work are 25fps, the visual features are extracted at a frequency of 25Hz. If the task is to train a visual-only speech recognition system, feature vectors at 25Hz can be used both for training and testing. On the other hand, for an audio visual speech recognition architecture, the sychronization of the audio and visual feature vectors might be required depending on the audio visual information fusion methodology. Therefore, after the calculation of the delta and accelaration coefficients on 25Hz data, visual feature vectors are upsampled to the frequency of audio feature vectors which is 100Hz by linear interpolation.

2.3 Hidden Markov Models

Once the speech signal is analysed and feature vectors are extracted, the next step is to model the speech using the feature vectors. Hidden Markov Models (HMM) with the ability to handle temporal evolutions in data have been the core framework for speech modelling since their first application to speech recognition [24]. An HMM is trained for each possible class using the feature vectors as observations. The terms *feature vector* and *observation* can be used interchangeably in an HMM context. A model corresponds to either a word or a phoneme depending on the application. If a model is defined to be a word then *word-HMM*s are trained and

if a model is defined to be a phoneme then *phoneme-HMM*s are trained. In this section, the theory of Hidden Markov Models (HMM) will be introduced in the context of isolated word recognition based on Rabiner's tutorial on HMM's [37] and the HTK Book [38]. Isolated word recognition is the task to recognise a single word from a set of possible words. Continuous speech recognition systems are established by embedding word HMMs in a finite state word network.

2.3.1 Objective of Isolated Word Recognition

Before introducing the details of HMMs, it would be helpful to illustrate the scope of modelling. The main objective of isolated word recognition is to find the most probable word spoken according to the observations given. In mathematical terms, given the observation sequence

$$\boldsymbol{O} = \boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T, \tag{2.20}$$

with T number of observation vectors, the aim is to find

$$\text{argmax}_i\{P(w_i|\boldsymbol{O})\}, \tag{2.21}$$

where w_i is the i'th vocabulary word. Since this probability is not directly computable, using Bayes' Rule it can be re-expressed as

$$P(w_i|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|w_i)P(w_i)}{P(\boldsymbol{O})}. \tag{2.22}$$

Then, the problem is reduced to finding the likelihood $P(\boldsymbol{O}|w_i)$ given the prior probabilities of each word. Considering that a model M_i is built corresponding to each word w_i, this likelihood can also be stated as $P(\boldsymbol{O}|M_i)$.

Equation 2.22 clearly points out that making a decision is a matter of likelihood calculation for each possible model.

2.3.2 Hidden Markov Models in Speech Recognition

HMM is a stochastic finite state machine which changes its state from state i to state j once every time unit with a transition probability of a_{ij} and at each time t that a state j is entered, an observation is generated by j'th state from the probability distribution $b_j(\boldsymbol{o}_t)$. The following parameters define an HMM:

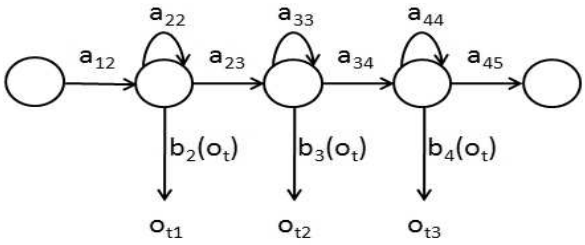


Figure 2.7: 5-state HMM with Non-emitting Entry and Exit States

- N : Number of states,
- $\boldsymbol{A} = \{a_{ij}\}$: Set of state transition probabilities from state i to state j,
- $\boldsymbol{B} = \{b_j(\boldsymbol{o}_t)\}$: Set of observation probability distributions in state j,
- $\boldsymbol{\Pi} = \{\pi_i\}$: Initial state distribution, i.e., the set of probabilities of state i being the initial state.

In the context of speech recognition, a special type of HMM named left-to-right HMM is favored with the following specifications:

- Only, transitions from a state to itself or to the following state is possible
- The entry and exit states are both unique and non-emmiting, i.e., they do not generate any observations.

A five-state, left-to-right HMM topology is given in Figure 2.7.

The parameter N has to be determined a priori which is a kind of a regularization parameter for HMMs. There is a tradeoff between too few states and too many states. Too few states will be inadequate to model the structure of the data and too many states will model the noise too.

Every emitting state corresponds to a segment of speech utterance. Usually 3-5 emitting states are used for phoneme-HMMs and 10-15 emitting states are used for word-HMMs. If phoneme-HMMs are built, then a word-HMM can be constructed by concatenation of appropriate phoneme-HMMs. Sequentially, continuous speech recognizers can be established by concatenation of word-HMMs. The entry and exit states serve for joining the HMMs.

The output distribution can be variant depending on the application but for speech recognition, generally Gaussian mixture densities are preferred. The mathematical representation for a Gaussian mixture density is

$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_{jm}|}} \exp(-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})\boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})), \tag{2.23}$$

where M is the number of mixtures, n is the dimension of the observation vector and $\boldsymbol{\mu}_{jm}$, $\boldsymbol{\Sigma}_{jm}$, c_{jm} are the mean, the covariance and the weight of mixture m of state j.

2.3.3 Training Hidden Markov Models

Training an HMM is determining the parameter set $\boldsymbol{\lambda} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}\}$. For left-to-right HMMs, the parameter $\boldsymbol{\Pi}$ is not relevant since the initial state is known to be the non-emitting entry state. The parameter $\boldsymbol{B}$ for Gaussian mixture distribution is equivalent to the parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, c\}$ which are the means, covariances and weights of the mixtures. The parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ are estimated recursively by the Baum-Welch Algorithm, also known as the Forward-Backward Algorithm. Estimation procedure is based on two newly defined probabilities which are forward and backward probabilities.

The forward probability $\alpha_j(t)$, defined as the probability of observing first t observation vectors and being in state j, can be recursively computed by

$$\alpha_j(t) = \sum_{i=2}^{N-1} [\alpha_i(t-1)a_{ij}]b_j(\boldsymbol{o}_t) \tag{2.24}$$

with initial conditions

$$\alpha_1(1) = 1, \tag{2.25}$$

$$\alpha_j(1) = a_{1j}b_j(\boldsymbol{o}_1), \tag{2.26}$$

for $1 < j < N$ and the final condition

$$\alpha_N(T) = \sum_{i=2}^{N-1} [\alpha_i(T)a_{iN}]. \tag{2.27}$$

Notice here that from the definition of $\alpha_j(t)$

$$P(\boldsymbol{O}|M) = \alpha_N(T). \tag{2.28}$$

Similarly, the backward probability $\beta_i(t)$, defined as the probability of observing the observation vectors from $t+1$ to T and being in state i, can be recursively computed by

$$\beta_i(t) = \sum_{j=2}^{N-1} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_j(t+1) \tag{2.29}$$

with initial conditions

$$\beta_i(T) = a_{iN}, \tag{2.30}$$

for $1 < i < N$ and the final condition

$$\beta_1(1) = \sum_{j=2}^{N-1} a_{1j} b_j(\boldsymbol{o}_1) \beta_j(1). \tag{2.31}$$

Multiplying the forward and backward probabilities, the probability of being in state i at time t and in state j at time $t+1$, $\xi_t(i,j)$, is derived as

$$\xi_t(i,j) = \frac{\alpha_i(t) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_j(t+1)}{\sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_i(t) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_j(t+1)}, \tag{2.32}$$

and the probability of being in state i at time t, $\gamma_i(t)$, is derived as

$$\gamma_i(t) = \sum_{j=1}^{N} \xi_t(i,j). \tag{2.33}$$

Given the above definitions, the re-estimation formulae can be expressed as

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_i(t)}, \tag{2.34}$$

$$\boldsymbol{\mu}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_{jm}(t)}, \tag{2.35}$$

$$\boldsymbol{\Sigma}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) (\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^{T} \gamma_{jm}(t)}, \tag{2.36}$$

$$c_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t)}{\sum_{t=1}^{T} \gamma_j(t)}. \tag{2.37}$$

Needless to say, the parameters to be estimated have to be initialized before the re-estimation procedure. The initial estimates can be chosen such that

$$a_{ij} = 0.5 \qquad 1 < i < N-1, \quad 2 < j < N, \tag{2.38}$$

$$\boldsymbol{\mu}_{jm} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{o}_t, \tag{2.39}$$

$$\boldsymbol{\Sigma}_{jm} = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})', \tag{2.40}$$

$$c_{jm} = \frac{1}{M}. \tag{2.41}$$

2.3.4 Recognition with the Viterbi Algorithm

Once the models are established for each word, recognition can be performed based on the model likelihoods. The likelihoods $P(\boldsymbol{O}|M)$ are calculated for each model over the most likely state sequence. The most likely state sequence can be identified using the Viterbi Algorithm.

In Viterbi Algorithm, for a given model M, $\phi_j(t)$ representing the maximum likelihood of observing first t observation vectors and being in state j is recursively computed by

$$\phi_j(t) = \max_i\{\phi_i(t-1)a_{ij}b_j(\boldsymbol{o}_t)\}, \tag{2.42}$$

where

$$\phi_1(1) = 1, \tag{2.43}$$

$$\phi_j(1) = a_{1j}b_j(\boldsymbol{o}_1), \tag{2.44}$$

for $1 < j < N$ which gives the best state sequence. Eventually, the likelihood $P(\boldsymbol{O}|M)$ can be evaluated by

$$P(\boldsymbol{O}|M) = \phi_N(T) = \max_i\{\phi_i(T)a_{iN}\} \tag{2.45}$$

and the model with the highest likelihood is decided to be the word spoken.

As stated earlier, a continuous speech recognizer can be established by embedding word HMMs in a finite state word network derived from a task grammar. The task grammar specifies the possible sequence of words. The grammar used for digit recognition in this work states that any digit can follow any other digit through the sequence and there are 10 digits to recognize in total. The word network resulting from the task grammar is diagrammed in Figure 2.8.

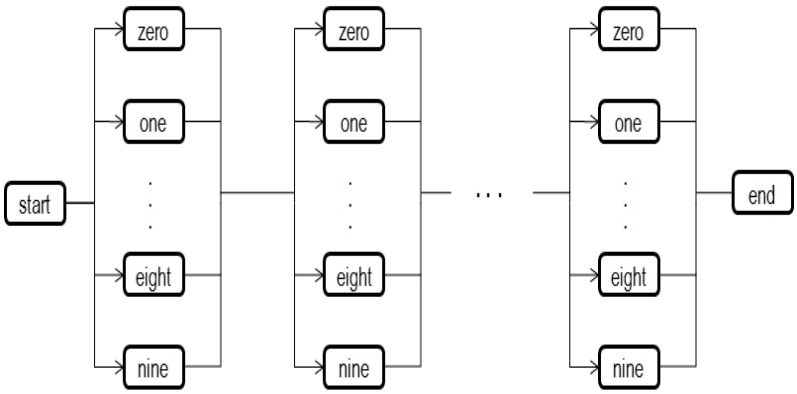


Figure 2.8: Digit Recognition Word Network

Chapter 3

Audio Visual Information Fusion

As mentioned in Chapter 1, humans integrate visual speech information extracted from the lip region with the acoustic information to recognize speech. McGurk was the first to conduct experiments to analyse the bimodality of speech perception and based on his experiments he concluded that visual information is not a secondary source but a complementary one [1]. Besides, visual information is not affected by the acoustic noise. All these lead to a theory that if discriminative visual information can be acquired and properly combined with the acoustic information, the performance of speech recognizers can be boosted especially in situations where there is acoustic noise. This is the main inspiration behind the Audio-Visual Speech Recognition research. There are two additional subjects to consider in audio visual speech recognition relative to audio speech recognition. First one is the visual feature extraction which is covered in section 2.2. The second one is the audio visual information fusion. Researchers intuitively propose statistical information fusion methodologies but their performances have not yet reached an admissable level. This work intends to contribute to such progress of information fusion for audio visual speech recognition systems.

In this chapter, the conventional information fusion techniques are presented in section 3.1 and a novel appraoch to information fusion is proposed in section 3.2.

3.1 Conventional Information Fusion Techniques

Audio Visual Information fusion algorithms proposed to date can be classified into three main groups:

- Feature Fusion (Early Fusion)

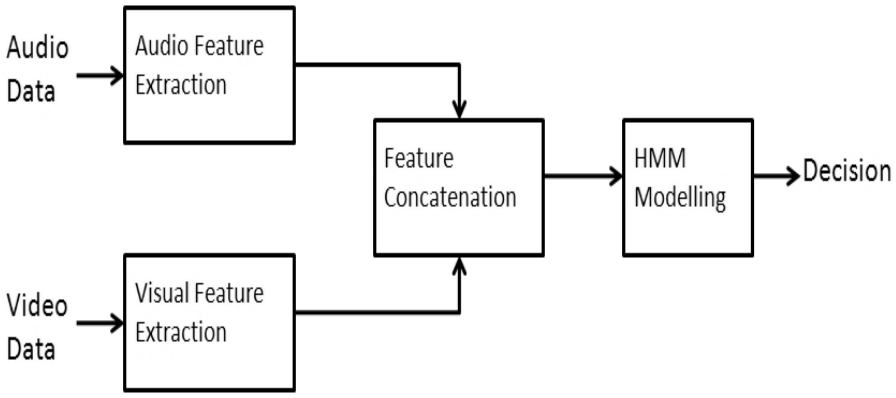


Figure 3.1: Feature Fusion Architecture

- Decision Fusion (Late Fusion)
- Model Fusion.

3.1.1 Feature Fusion (Early Fusion)

Feature level fusion, also named Early Fusion, is perhaps the most primitive approach to information fusion for audio visual speech recognition (AVSR) in which feature vectors from multiple streams are concatenated to form a combined feature vector and this combined feature vector is fed into an HMM as an observation resulting in a single model for each word. Dimensionality reduction techniques of which LDA is the most popular can be applied if the combined feature vector is oversized. LDA as a feature reduction technique is analyzed in Chapter 2, hence it will not be repeated here. The AVSR system architecture with feature concatenation is given in Figure 3.1.

3.1.2 Decision Fusion (Late Fusion)

In decision fusion (or late fusion), observations from each data source are separately modelled attaining posterior probabilities for each data stream. Subsequently, posterior probabilities of each stream are combined to come up with a final decision. Decision fusion architecture is given in Figure 3.2.

As explained in section 2.3, the likelihood of an observation sequence $\boldsymbol{O}$ extracted from an utterance to be generated by a word model M_i, $P(\boldsymbol{O}|M_i)$, is evaluated for each word model and the word with the highest posterior probability is determined

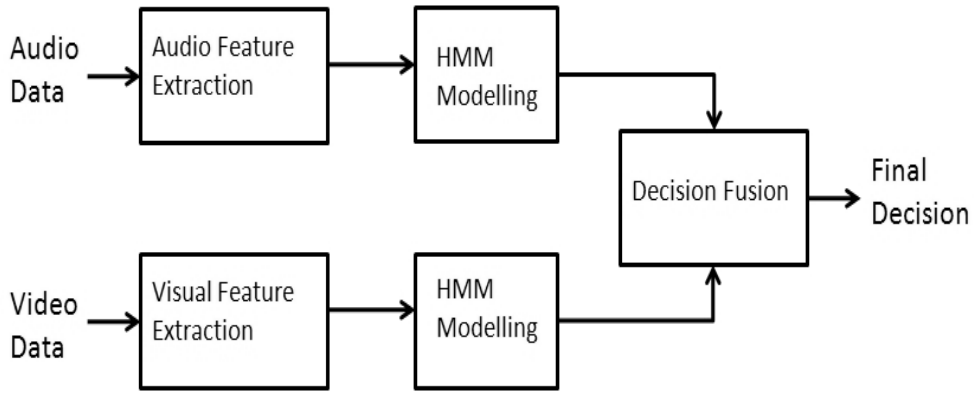


Figure 3.2: Decision Fusion Architecture

to be the word spoken. This decision methodology is valid if there is only one observation sequence and one model for a particular word. However, in a multimodal task, two observation sequences are extracted from an utterance and two models are built corresponding to one word. An acoustic model M_a is trained with acoustic feature vectors $\boldsymbol{O}_a$ and a visual model M_v is trained with visual feature vectors $\boldsymbol{O}_v$. Decision fusion aims to combine $P(\boldsymbol{O}_a|M_{ia})$ and $P(\boldsymbol{O}_v|M_{iv})$ to make the final decision.

The likelihoods can be combined with some simple techniques such as multiplying the likelihoods of each stream, summing them or taking the maximum. The list can be extended but these simple techniques do not offer weighting of the two modalities for different noise levels. A commonly used scheme which provides weighting of the streams is

$$\hat{W}_i = \text{argmax}_{i=1:N}\{\gamma_a \cdot log(P(\boldsymbol{O}_a|M_{ia})) + \gamma_v \cdot log(P(\boldsymbol{O}_v|M_{iv}))\}, \tag{3.1}$$

where $\hat{W}_i$ is the most likely word, N is the number of possible words and γ_a and γ_v are weights of acoustic and visual weights respectively. The weights are adjusted depending on the conditions.

Above, decision fusion is described for isolated word recognition. For continuous word recognition, decision fusion may require enumerating all possible word sequences which is not easy.

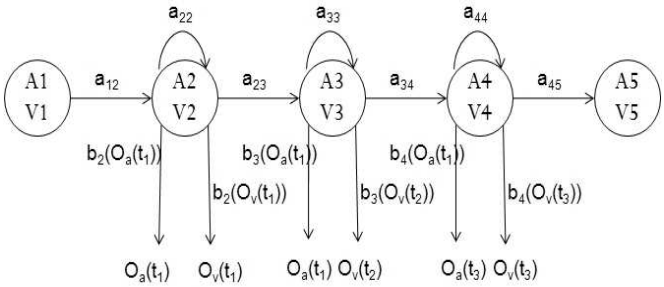


Figure 3.3: Multiple Stream HMM Topology

3.1.3 Model Fusion

Model fusion algorithms integrate the information from the two streams during the model building procedure. The principal model fusion architecture is Multiple Stream Hidden Markov Models (MSHMM). MSHMMs model more than one stream of observations in a parallel structure allowing independent likelihood calculation for each stream. Its topology for a five state phone model is pictured in Figure 3.3.

The states of MSHMM are tied states which means that the same states are shared between the two streams. Therefore, state transition probabilities are the same for both streams. Mathematically, MSHMMs differ from regular HMMs only in observation probability distribution given by

$$b_j(\boldsymbol{o}_t) = \prod_{s=\{a,v\}} \left[\sum_{m=1}^{M} c_{jsm} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_{jsm}|}} \exp(-\frac{1}{2}(\boldsymbol{o}_{st} - \boldsymbol{\mu}_{jm})\boldsymbol{\Sigma}_{jsm}^{-1}(\boldsymbol{o}_{st} - \boldsymbol{\mu}_{jm})) \right]^{\gamma_s} \tag{3.2}$$

for the two stream case where $s = \{a, v\}$ represents the audio and visual streams respectively and γ_s is the weight of the stream s. The rest of the parameters are the same as the parameters of equation 2.23. In this work a facility called single-pass retraining is utilized to train the MSHMM. Single-pass retraining is a mechanism for mapping a set of models trained using one parametrisation into another set based on a different parametrisation. This is done by computing the forward and backward probabilities using the original models together with the original training data, but then switching to the new training data to compute the parameter estimates for the new set of models. Since the audio models are more reliable for clean data in audio visual speech recognition; first, audio models are generated from the audio stream. The visual models trained using single-pass retraining perform better than

the visual models trained using only the visual observations.

The MSHMM restricts the streams to be state synchronous so that a transition from a state to another takes place at the same time. This is not a desirable situation since the visual information can sometimes precede the acoustic information, i.e., the lip movement can occur before the speech is produced. Product HMM (PHMM), which is an extension of MSHMM, allows state asynchrony between the two streams forcing the streams to be synchronous at the model boundaries [10]. There are also more advanced HMMs utilized in audio visual speech recognition which include Factorial HMM (FHMM) and the Coupled HMM (CHMM). In FHMM, the audio and visual states are independent of each other, but they jointly model the likelihood of the audiovisual observation vector, and hence become correlated indirectly [9]. In CHMM, the likelihoods of the audio and visual observation vectors are modeled independent of each other, but each of the audio and visual states are conditioned jointly by the previous set of audio and visual states [11].

The performances of MSHMM, PHMM, FHMM and CHMM are analysed by Nefian [9]. The results in that work showed that PHMM and FHMM do not improve the recognition rate compared to MSHMM and CHMM outperforms MSHMM by 1-2% . Investigating Nefian's results; PHMM, FHMM and CHMM are not considered in this study due to their implementation complexity.

3.2 Proposed Framework : Tandem Fusion

The *Tandem Fusion Approach* to information fusion in audio visual speech recognition proposed in this work is founded on the *tandem* framework for audio speech recognition first presented by Hermansky in 2000 [12]. In Hermansky's system, a neural network is trained to estimate the posterior probabilities of each frame for each possible class where a class corresponded to a phoneme. The inputs to the neural network were the MFCC features and the outputs were vectors of posterior probabilities, with one element for each phoneme. The posterior probability vectors were used as observations for a Gaussian-mixture-based HMM system. The results demonstrated that the novel tandem approach improved the recognition accuracy compared to the conventional HMM system where the MFCC features are directly used as observations. In this study, Hermansky's tandem approach is exploited to

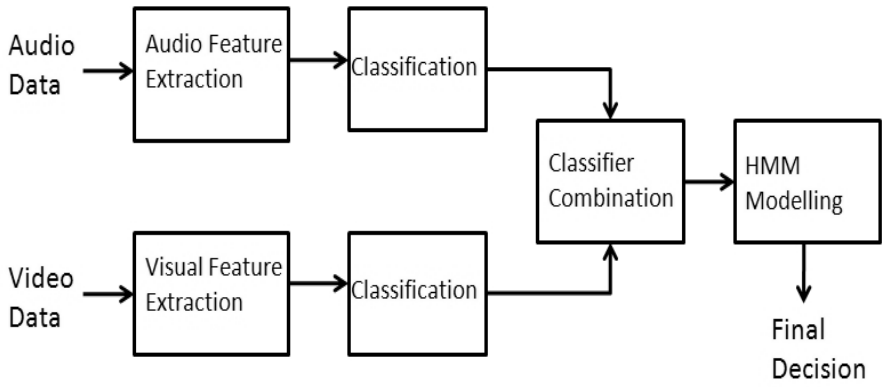


Figure 3.4: Tandem Fusion Architecture

propose a novel information fusion framework for audio visual speech recognition.

Tandem fusion framework with the block diagram in Figure 3.4, has grounds both in feature fusion and decision fusion. It appears to be a kind of feature fusion scheme since the information from the two modalities are fused prior to Hidden Markov Modelling. On the other hand, it can be associated with decision fusion techniques because it employs a preliminary decision stage for each modality separately before information fusion. The intention of this approach is to provide more discriminative observations for HMM utilizing both modalities maximally in changing conditions.

The tandem fusion framework can be divided into four main stages. The first stage is feature extraction stage, the second stage is separate classification of each stream at frame level, the third stage is classifier combining and the last stage is modelling.

3.2.1 Training the System

As the first step of training, audio and visual feature vectors are obtained for each speech frame on clean data with the techniques described in sections 2.1 and 2.2.

Training the First Level Classifiers

The next step following the feature extraction step is the training of individual classifiers for each stream. In this study, Gaussian Mixture Models (GMM) are utilized as individual classifiers. A GMM with 12 mixtures is trained for each possible class in each stream where a class corresponded to a word in this case. Assuming there are C number of classes in the dataset, C number of GMMs are

trained for audio stream using audio feature vectors as inputs and C number of GMMs are trained for visual stream using visual features as inputs.

A labelled training dataset is needed to train a GMM for each class but the dataset used in this work is not labelled. The labels of the feature vectors are obtained according to the results of an audio speech recognition system with MFCC features since the audio only system achieves a recognition accuracy of 100% on noise-free data. To assign a label to each speech vector, exact word boundaries has to be known and these boundaries are determined by the Viterbi alignment procedure. The Viterbi alignment procedure is a constrained Viterbi decoding process where the correct word labels are known.

The probability distribution formula for GMM is given by the equation

$$p(x) = \sum_{m=1}^{M} c_m \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_m|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_m)\}, \qquad (3.3)$$

where $\boldsymbol{x}$ is the d dimensional feature vector, M is the total number of mixtures which is 12 in this case, c_m is the m'th mixture weight, $\boldsymbol{\mu}_m$ is the mean of mixture m and $\boldsymbol{\Sigma}_m$ is the covariance matrix of mixture m. Different classifiers can be used instead of GMM as individual classifiers, most popular examples being neural networks and support vector machines. GMMs are preferred to others in this work for computational restrictions and for their common success in speech modelling.

Training the Combining Classifier

GMM training stage is followed by the classifier combining stage where the integration of the information from the two modalities is established. In this work, Linear Discriminant Classifier (LDC) is chosen as the combining classifier. Support Vector Machines and Neural Networks are also thought as alternatives but could not be implemented due to computational insufficiencies. The results showed that LDC fulfills the needs though. An LDC is trained for each noise level. The variation of LDCs for different noise levels is obtained by using noisy data as the audio input.

The input vectors of the LDC are the output posterior probability vectors of the GMM stage. The posterior probability of a feature vector for a given class (a class corresponds to a word) is calculated by

$$p(\boldsymbol{x}|C_i) = \sum_{m=1}^{M} c_{im} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_{im}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{im})^T \boldsymbol{\Sigma}_{im}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{im})\}, \quad (3.4)$$

where C_i is the i'th class, $\boldsymbol{x}$ is the d dimensional feature vector, M is the total number of mixtures which is 12 in this case, c_{im} is the m'th mixture weight of class i, $\boldsymbol{\mu}_{im}$ is the mean of mixture m of class i and $\boldsymbol{\Sigma}_{im}$ is the covariance matrix of mixture m of class i.

The values of $p(\boldsymbol{x}|C_i)$ for each class in a stream are gathered to form a C dimensional posterior probability vector for each speech frame. C dimensional posterior probability vector from the audio stream and the C dimensional posterior probability vector from the visual stream are concatenated to be the input for the LDC. Note that the training dataset used for LDC training is different from the training dataset used for GMM training. This separate training data is called held-out or validation data in some studies.

LDC assumes that each class has a multivariate Gaussian distribution and all classes share the same covariance matrix. Training the LDC is equivalent to finding means for each class and the common covariance matrix. The common covariance matrix is calculated the same way as within class scatter matrix is calculated for LDA in section 2.2.2.

Once the means for every class and the common covariance matrix are acquired, the next step is to extract observation vectors for HMMs which are the outputs of the LDC stage. The dataset used for HMM training is the combination of the GMM training dataset and the LDC training dataset. LDC's discriminant function given in equation (3.5) is evaluated for every speech frame and for every class to generate a C dimensional observation vector for HMM.

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(x - \boldsymbol{\mu}_i)^T \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln(P_i). \quad (3.5)$$

In equation (3.5), $g_i(\boldsymbol{x})$ is the discriminant function giving a scalar value, $\boldsymbol{x}$ is the $2C$ dimensional posterior probability vector from the GMM stage, $\boldsymbol{\mu}_i$ is the mean of class i, $\boldsymbol{\Sigma}$ is the common covariance matrix and P_i is the prior probability of class i which is calculated as the ratio of the number of training examples belonging to class i, to the number of total training examples.

Training HMM

C dimensional observation vectors for every speech frame are used to train a regular single stream word HMM as the final stage of training. GMM-LDC sequential stage serves for extracting more discriminative features compared to the unprocessed audio features by taking advantage of both the idea of tandem approach and the integration of visual information to the system. The GMM-LDC sequential stage is followed by an HMM stage because the combining classifier generates posterior probabilities at frame level and frame level decision is vulnerable to abrupt interferences whereas HMM as a state machine tolerates such interferences and handles the temporal evolution of data well.

3.2.2 Testing Process

The main concept in the tandem fusion approach is extracting the most adequate observations for class separability in HMMs. Every audio and visual feature vector is processed through the GMM-LDC sequential stage to create audio-visual observation vectors for a regular HMM. First, the posterior probability of an audio feature vector for a given class is calculated using equation (3.4). Repeating this for each class and collecting the posterior probability values in a vector, a posterior probability vector is formed corresponding to the audio feature vector. A similar procedure is applied for the visual feature vector to generate a posterior probability vector for the visual stream. The posterior probability vectors from the two streams are concatenated to create the combined posterior probability vector which is the input vector for LDC. Remember that there is an LDC trained for each noise level. Using the appropriate LDC for the present noise condition, the discriminant function in equation (3.5) is evaluated on the combined posterior probability vector for a given class. Repeating this for each of C classes, a C dimensional audio-visual observation vector is generated as the input to a single stream HMM. The rest is the standard HMM procedure explained in Chapter 2.

3.3 Computational Time Comparison of Tandem Fusion and MSHMM

The MSHMM and the tandem fusion approach will be compared in terms of recognition accuracy and run time in Chapter 4. Additionally, the two frameworks are compared in terms of computation time in this section.

3.3.1 Computation Time of Tandem Fusion

Tandem fusion approach has three stages to consider for computational load calculation; GMM stage, LDC stage and HMM stage. Consider a single speech frame with d_A dimensional audio feature vector and d_V dimensional visual feature vector. In the GMM stage, the posterior probability of the audio feature vector being in class i is calculated by equation (3.4). The computational load of the term

$$c_{im}\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_{im}|^{1/2}} \tag{3.6}$$

in equation (3.4) can be neglected since the terms are independent of the input vector, hence are not repetitively calculated for each speech frame. The main computational load comes from the calculation of the term

$$\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{im})^T\boldsymbol{\Sigma}_{im}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{im})\}, \tag{3.7}$$

where the covariance matrix $\boldsymbol{\Sigma}_{im}$ is diagonal. Assuming it takes T time units to process the term in equation 3.7 for a scalar x, it would take d_AT time units for a d_A dimensional $\boldsymbol{x}$. This is repeated for M number of mixtures and C number classes making a total of

$$CMd_AT \tag{3.8}$$

time units to calculate the posterior probability vector of a single audio feature vector. Similarly,

$$CMd_VT \tag{3.9}$$

time units are needed for the visual posterior probability vector calculation. Therefore, in total it takes

$$CMT(d_A+d_V) \tag{3.10}$$

time units to complete the GMM stage of the tandem fusion approach.

In the LDC stage, equation (3.5) is implemented for each speech frame and the dominant operation in this stage is also the calculation of the term 3.7. The dimension of the input vector $\boldsymbol{x}$ is $2C$ where C is the number of classes in the dataset. So, for a single speech frame, computation time is $2CT$ time units. Considering that the discriminant function in equation (3.5) is evaluated for each C number of classes, the LDC stage is terminated in

$$2C^2T \tag{3.11}$$

time units.

As stated in section 2.3, the decision of the word spoken is given according to the likelihood computed by equation (2.45). In equation (2.45), a_{iN} is a scalar value and multiplication with a scalar value does not constitute much to the processing time, hence it is neglected. The main processing load consists of recursive calculation of $\phi_i(T)$ by equation (2.42). The scalar multiplicatives $\phi_i(t-1)$ and a_{ij} in equation (2.42) can be neglected since the major processing load is the calculation of $b_j(t)$ which is the probability distribution of a Gaussian mixture density for a state j in equation (2.23). The computational load analysis for a Gaussian Mixture was done for the GMM stage. Following a similar approach, for a single mixture of a state, the operation time is CT time units. Considering that there are M number of mixtures and N number states in a model and there are C number of models,

$$NMC^2T \tag{3.12}$$

time units are required for the HMM stage of the tandem fusion methodology.

Addition of equation (3.10), (3.11) and (3.12) gives the total computation time of the whole tandem fusion procedure as

$$CMT(d_A + d_V) + 2C^2T + NMC^2T. \tag{3.13}$$

3.3.2 Computation Time of MSHMM

The computation time of the MSHMM system can be calculated very similar to the regular HMM case but this time there are two Gaussian mixture distributions with different observation vector dimensions, one for each stream as seen in equation (3.2). Mathematically, equation (3.12) becomes

$$CNMT(d_A + d_V) \tag{3.14}$$

for the MSHMM framework.

There are two reasons that make the value of equation (3.13) smaller than equation (3.14). First, the dimensions d_A and d_V are much bigger in value than C. Second, GMM stage does not constitute much to the processing load because there is a GMM for each class whereas there is a GMM for each state of a class in the HMM.

The computational advantage of the tandem fusion approach against the MSHMM would be more apparent with a numeric comparison for the scenario in this work. There are $C = 11$ classes in the dataset, the audio feature vector is $d_A = 39$ dimensional, the visual feature vector is $d_V = 63$ dimensional, the number of mixtures, M of GMMs for all stages are 12 and HMMs with 10 emitting states are used. Evaluating equation (3.13) and (3.14) with these values, we arrive at $0.28 \times 10^5 T$ time units for the tandem fusion approach and $1.35 \times 10^5 T$ time units for the MSHMM respectively.

Chapter 4

Experiments and Results

This chapter is dedicated to experimental analysis of the proposed method. The experimental procedure is as follows: First, audio-only speech recognition systems are trained and tested to observe the performances of different audio feature types described in section 2.1. Second, visual-only speech recognition systems are trained to observe the performances of different visual feature types described in section 2.2. Both word-HMMs and phoneme HMMs are investigated for the single stream recognizers. Best performing audio and visual feature types are selected to be used in audio visual speech recognition scenarios. Lastly, MSHMM based and tandem fusion based audio visual speech recognition systems utilizing the previously selected feature types and HMM topology are established. The MSHMM and the tandem fusion approach are compared in terms of recognition accuracy and processing time. The chapter is organized as follows: Section 4.1 describes the database used for the experiments, section 4.2 lists the computational tools utilized and section 4.3 explains the noise addition procedure. The evaluation metric is given in section 4.4 and the HMM topology is presented in section 4.5. The results of Audio Speech Recognition experiments, Visual Speech Recognition experiments and Audio Visual Speech Recognition experiments are analysed in sections 4.6, 4.7 and 4.8 respectively.

4.1 Database

The experiments are conducted on M2VTS database [39] which consists of 5 different video recordings for each 37 subjects at 5 different times. The recordings are head-and-shoulder videos in an office environment with plain gray background. The subjects count digits from 0 to 9 in order in French. The database is organized in

5 tapes, each tape containing a single recording of each subject. The audio track of the recordings are sampled at 48kHz with 16-bits PCM coding. The video track has a frame rate of 25fps and 286x360 frame size.

4.2 Computational Tools

Various computational tools are used to conduct the experiments. Hidden Markov Models are built with the HMM Toolkit (HTK) [38]. Audio features are also extracted using the built-in functions in HTK. Visual features are extracted in Matlab whereas the lip regions are extracted from the face images with Visual C++ utilizing OpenCV. The Matlab Toolbox for Pattern Recognition (PRTools) and Voicebox toolbox are used for GMM and LDC training in the tandem fusion approach.

4.3 Noise Addition

Car noise at SNRs ranging from 20dB to -5dB are artificially added to analyse the performance of the recognition systems in noisy conditions. Noise addition is applied according to ITU-T P.56 standard [40] with the software provided by ITU.

4.4 Evaluation Metric

The evaluation metric used for the performance of a speech recognition system is the *Recognition Accuracy* as a percentage which is given by

$$A = \frac{N - D - S - I}{N} \cdot 100\% \tag{4.1}$$

where N is the total number of labels in the test dataset, D is the number of deletions, S is the number of substitutions and I is the number of insertions [38].

Since a label corresponds to a word (digit) in this case and the digits are counted in order in M2VTS database, the correct label sequence for each recording is {*zero-one-two-three-four-five-six-seven-eight-nine*}. Considering that one tape is used for testing and there are 37 speakers, N has the value 370. A deletion example would be {*...-three-four-six-seven...*}, a substitution example would be {*...-three-nine-five-six-...*} and an insertion example would be {*...-three-nine-four-five-...*}.

4.5 Hidden Markov Model Topology

Digit recognition is a limited vocabulary scenario where there are 11 words to discriminate, 10 digits together with the *silence/short pause*. For limited vocabulary tasks, either word HMMs or phoneme HMMs can be preferred. In this work, both are tested to conclude that word HMMs suit better.

As stated in section 2.3, number of states is a regularization parameter for HMMs. For this task, word HMMs are established with 10 emitting states and phoneme HMMs are established with 3 emitting states. Considering that the words in the database consist of 2-4 phonemes as shown in Table 2.1, a state in both word-HMM and phoneme-HMM approximately corresponds to the same segment of speech. The number of Gaussian mixtures for each state is also experimentally determined to be 12 for both audio and visual streams.

4.6 Audio Speech Recognition Experiments

Three types of audio features described in section 2.1 are investigated in a single modality speech recognition system. For all three, 12 static features are extracted from a speech frame. Together with the energy of the frame, there are a total of 13 static features. Dynamic information is attained by calculating the delta and acceleration coefficients over two neighbouring frames as described in section 2.1.3. Eventually, 39 dimensional acoustic observation vectors are obtained. Frame lengths are chosen to be 25ms and overlapping frames are extracted every 10ms. Four tapes of the database are used for training and the last tape for testing. The results are demonstrated for both word level and phoneme level modelling in Table 4.1. Figures 4.1, 4.2 and 4.3 show that phoneme level accuracy is superior to word level accuracy for MFCC and PLP and it is the opposite for AFE. Nevertheless, word level recognition will be preferred for audio visual scenarios because visual features perform significantly better in word level. Figure 4.4 exhibits that the performances of MFCC and PLP are close to each other whereas AFE outperforms the former two on account of its noise reduction scheme. Even so, AFE will not be used as the audio feature in the following audio visual scenarios because of its heavy computational load compared to MFCC and PLP. It is implemented to make a comparative analysis

of the performances of the audio visual frameworks with noise reduction algorithms since both methodologies aim to overcome the noise problem. Figure 4.4 denotes that MFCC is to be selected as the best performing audio feature for this task. Hence, MFCCs will be utilized in the audio visual speech recognition scenarios.

Noise Level	MFCC		PLP		AFE	
	Word	Phone	Word	Phone	Word	Phone
Clean	100.00	99.73	100.00	100.00	100.00	100.00
20dB	96.76	98.38	99.46	98.65	99.46	99.19
15dB	86.49	94.05	85.95	97.30	99.73	97.84
10dB	52.43	74.32	51.14	79.19	98.11	96.49
5dB	39.72	35.14	34.71	48.11	95.88	85.28
0dB	29.52	12.97	19.39	18.06	85.24	66.67
-5dB	0.00	5.68	0.00	9.41	70.00	47.71

Table 4.1: Acoustic ASR Accuracy (%)

4.7 Visual Speech Recognition Experiments

Similar to audio-only speech recognition experiments, visual-only speech recognition systems are implemented to decide on which visual feature to use in audio visual speech recognition. The preprocess of visual feature extraction is the lip region extraction. Accordingly, face detector is applied to the videos and lip region is extracted as the bottom 40% of the face region vertically and central 50% of it horizontally. The resulting lip images are fixed to the size 48x64.

Two dimensional DCT, PCA and LDA are applied on gray scale images. Their successes are analysed with different number of coefficients. As in the case of audio features, both word level and phoneme level systems are tested. Training and testing are carried on 25Hz data for the visual only recogniton. For all three cases, four tapes of the database are used for training and the last tape for testing.

The outcomes of the visual speech recognition system with DCT features are given in Table 4.2 and Figure 4.5. The letter T and S stand for triangle and square respectively and they represent the triangular or square region in the upper left

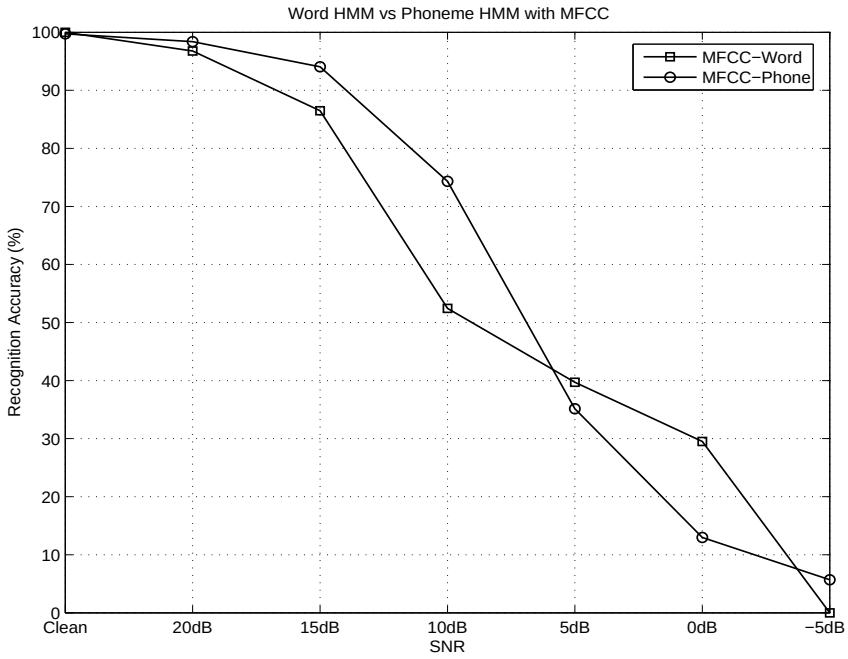


Figure 4.1: Acoustic ASR with MFCC

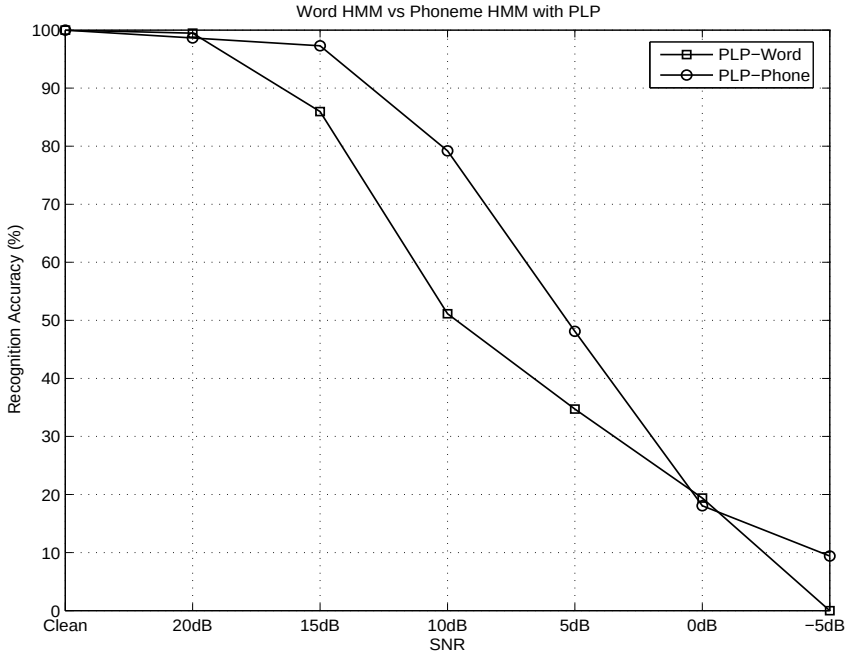


Figure 4.2: Acoustic ASR with PLP

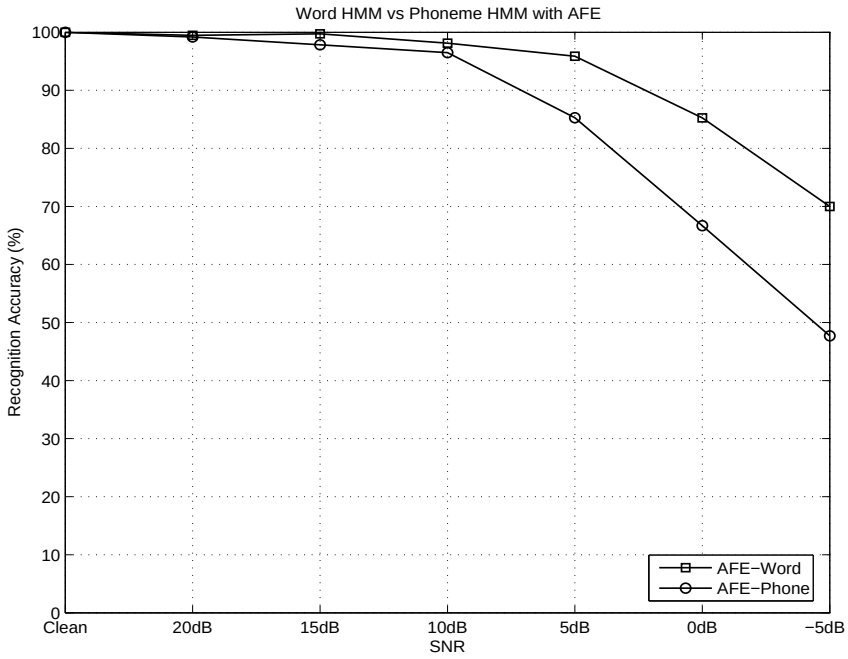


Figure 4.3: Acoustic ASR with AFE

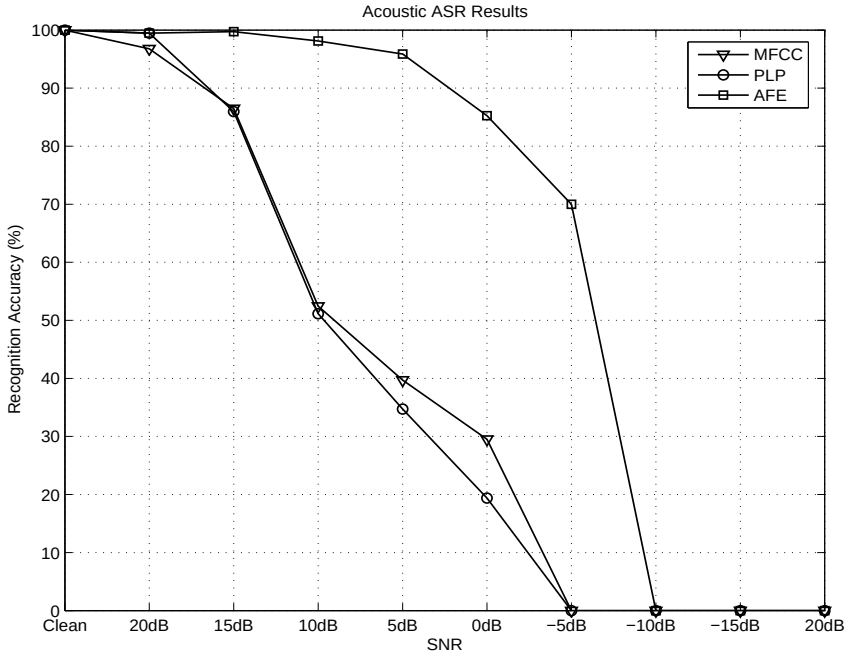


Figure 4.4: Word-level Acoustic ASR with Different Features

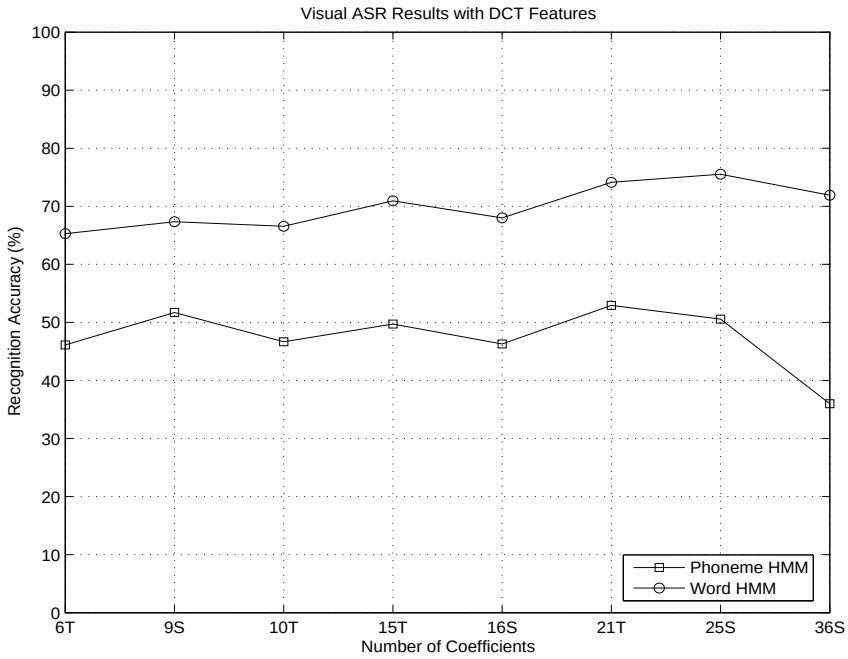


Figure 4.5: Visual ASR with DCT Features

corner of the DCT matrix. The coefficients within the specified region are used as the static features. The delta and acceleration coefficients are derived as in the case of audio feature extraction.

	Number of DCT coefficients							
	6T	9S	10T	15T	16S	21T	25S	36S
Phone HMM	46.11	51.71	46.67	49.71	46.29	52.94	50.57	36.00
Word HMM	65.29	67.35	66.56	70.94	68.00	74.14	75.52	71.90

Table 4.2: Visual ASR with DCT Coefficients (%)

It can be observed from Figure 4.5 that 21 DCT coefficients extracted from the upper left triangle of the DCT matrix would be a proper choice. A minor improvement with 25 DCT coefficients can be disregarded with the gain of 12 features (4 static features, 4 delta features and 4 acceleration features).

Success rate of PCA as a visual feature is investigated by trying different number of eigenvectors to create the PCA transformation matrix. The results are displayed in Table 4.3 and Figure 4.6. According to the table, highest recognition accuracy is reached with 20 eigenvectors.

LDA could not perform any better than random guess for this task, hence the

HMM Type	Number of Eigenvectors					
	5	10	15	20	25	30
Phone HMM	40.56	45.28	45.83	55.56	48.06	44.57
Word HMM	28.48	34.39	28.89	28.82	25.83	21.47

Table 4.3: Visual ASR with PCA Features(%)

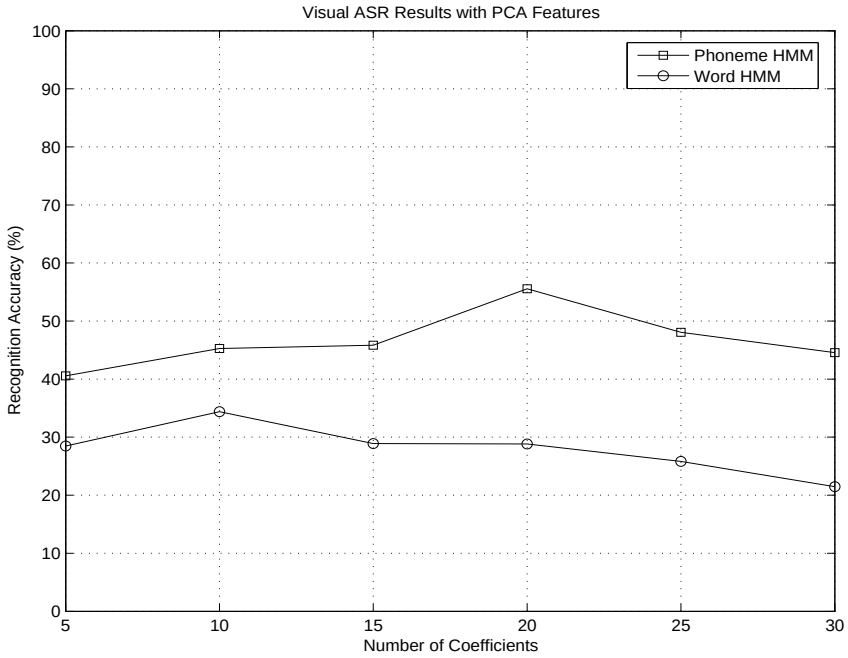


Figure 4.6: Visual ASR Accuracy with PCA Features

results are not provided. In theory, LDA is expected to perform better than PCA since it seeks for a tranformation in the most discriminative sense. However, it is stated that for the cases where the training dataset is small, PCA can outperform LDA [41] which is probably the case here.

Assembling the outcomes of DCT and PCA, DCT is selected for the word level audio visual speech recognition system.

4.8 Audio Visual Speech Recognition Experiments

Audio visual speech recognition experiments are carried out with MFCCs as the audio features and DCT coefficients as the visual features. MFCCs are extracted as 39 dimensional vectors at a frequency of 100Hz. DCT coefficients are extracted as 63 dimensional vectors at a frequency of 25Hz. Since, both MSHMM and the tan-

dem approach necessitates observation synchrony between the two streams, visual observations are upsampled to 100Hz by linear interpolation. Due to interpolation, recognition accuracy decreases at 100Hz compared to the recognition accuracy at 25Hz.

For the MSHMM case, three tapes are used for training the model. The trained model is evaluated on the fourth tape with varying weights and the best performing weights are determined for each SNR level according to the recognition accuracies. Finally, the system is tested on tape-5 with the pre-determined weights. The weights for each SNR level and the recognition accuracy of MSHMM system can be seen in Table 4.4.

	MSHMM			Tandem
Noise Level	Audio Weight	Video Weight	Accuracy	Accuracy
Clean	1	0	100.00	99.44
20dB	1	0	99.72	95.28
15dB	0.9	0.1	95.28	93.06
10dB	0.6	0.4	83.61	84.17
5dB	0.3	0.7	70.00	68.61
0dB	0.2	0.8	54.29	48.53
-5dB	0	1	53.14	31.00

Table 4.4: Audio Visual ASR (%)

In the tandem system, three tapes are used to train Gaussian Mixture Models with 12 mixtures for each data stream. Linear Discriminant Classifier is trained with the fourth tape. A class corresponds to a word both for GMM and LDC. The case where a class corresponds to a phoneme is also tried for the GMM and LDC stages but the phoneme level classification did not perform as well as the word level classification. Number of classes is 11 since there are 10 digits and an additional class of silence/short pause. On this account, the dimension of the output vectors of GMM are 11. Consequently, the inputs to the LDC stage are 22-dimensional vectors and the outputs are again 11-dimensional posterior probability vectors. Once the GMM-LDC stage is trained, posterior probability vectors are extracted as features for the whole dataset. The 11-dimensional posterior probability vectors are treated

as observations to single stream HMMs. HMMs are trained with the first four tapes and tested on the last tape. The results of the tandem approach are recorded in Table 4.4 indicating resembling performance with the MSHMM.

In addition to comparison of recognition accuracies, the two frameworks are compared according to the processing times of their testing stages. The processing times with HVite program in HTK running on Intel Xeon 2.0GHz processor are given in Table 4.5.

MSHMM	Tandem Fusion
129.978	31.112

Table 4.5: Processing Times (in seconds)

Figure 4.7 summarizes the main concept of this study. First of all, the dramatic performance degradation of MFCC based acoustic speech recognition system is represented. Visual information, although not much competent as a single modality, supports the acoustic stream in audio visual speech recognition frameworks. The proposed tandem approach shows compariable performance with the MSHMM which is a promising indication for future studies. Since the weights are manually assigned for each SNR level in MSHMM, there is a minimum limit for the recognition accuracy which is obtained by giving zero weight to the acoustic stream. On the other hand, no manual weight assignment is done in the tandem fusion approach, hence the recognition accuracy of the proposed approach is lower than MSHMM in very high noise levels. No matter which audio visual architecture is used, acoustic speech recogition systems employing AFE exhibits better performance with the help of noise reduction. This situation approves that there is still much work to be conducted in audio visual speech recognition research.

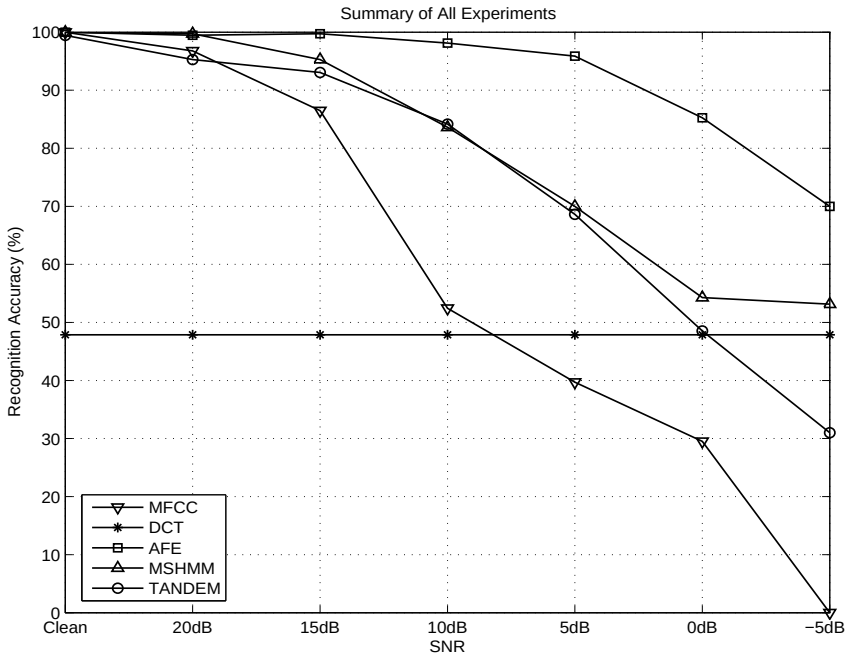


Figure 4.7: Summary of Experiments

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this study, a novel framework for audio visual information fusion which employs a preliminary decision stage prior to Hidden Markov Modelling is proposed. In the preliminary decision stage, the two data streams are separately modelled with GMMs and then combined with LDC as the combining classifier. Classifier combination is claimed to be superior to feature fusion and decision fusion methods. The reason is that the bimodal information is fused in a way to maximally discriminate among different classes. Hidden Markov Modelling stage following the preliminary decision fusion stage served for modelling the temporal evolutions in the data which could not be realized with the frame level decision of the first stage.

The proposed approach is compared with the MSHMM which is argued to be the principal audio visual information fusion framework [8]. The recognition accuracy results for the two systems were comparable. Also, the tandem fusion approach had a superiority to the MSHMM in run time. The tandem approach can be considered as a promising candidate for information fusion in audio visual speech recognition due to the fact that improved versions of it can be constructed using different singular classifiers and combining classifiers.

5.2 Future Work

As future work, different singular classifiers and combining classifier can be analysed. Especially, Neural Network (NN) as a singular classifier has the potential to be superior to GMM if computational needs are fulfilled.

Another future work is to test the novel method in large vocabulary tasks but first an extensive database is needed which is one of the main drawbacks of audio visual speech recognition research.

Bibliography

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[3] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, University of Illinois, Urbana, 1984.

[4] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 821–824, May 1996.

[5] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, pp. 165–168 vol.1, 2001.

[6] A. Adjoudani and C. Benoit, *Speechreading by Humans and Machines*. Berlin, Germany: Springer, 1996.

[7] P. Teissier, J. Robert-Ribes, J. Schwartz, and A. Guerin-Dugue, "Comparing models for audiovisual fusion in a noisy vowel recognition task," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 629–642, Nov 1999.

[8] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *Multimedia, IEEE Transactions on*, vol. 2, no. 3, pp. 141–151, Sep 2000.

[9] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, pp. 1–5, Nov. 2002.

[10] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech, Signal Processing*, 2001, pp. 169–172.

[11] S. Chu and T. Huang, "Bimodal speech recognition using coupled hidden markov models," in *Proc. of IEEE International Conference on Spoken Language Processing*, 2000, pp. 747–750.

[12] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.

[13] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Press, 1994.

[14] A. J. Robinson, L. Almeida, J. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J. P. Neto, S. Renals, M. Saerens, C. Wooters, H. Speechproducts, and H. Speechproducts, "A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The wernicke project," in *Proc. EUROSPEECH'93*, 1993, pp. 1941–1944.

[15] J. H. Aravind Ganapathiraju and J. Picone, "Support vector machines for speech recognition," in *Proceedings of the International Conference on Spoken Language Processing*, 1998, pp. 2348–2355.

[16] A. Ganapathiraju, J. Hamaker, and J. Picone, "Hybrid SVM/HMM architectures for speech recognition," in *in Speech Transcription Workshop*, 2000.

[17] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, Aug. 2004.

[18] A. Garcia-Moral, R. Solera-Urena, C. Pelaez-Moreno, and F. D. de Maria, "Hybrid models for automatic speech recognition: A comparison of classical ann and kernel based methods," in *Proc. of NOLISP*, 2007, pp. 152–160.

[19] M. Gordan, C. Kotropoulos, and I. Pitas, "Visual speech recognition using support vector machines," in *Proc. of Digital Signal Processing Conference*, 2002, pp. 1093–1096.

[20] S. E. Krger, M. Schaffner, M. Katz, E. Andelic, and A. Wendemuth, "Speech recognition with support vector machines in a hybrid system," in *in Proc. EuroSpeech*, 2005, pp. 993–996.

[21] M. Gurban and J. P. Thiran, "Audio-visual speech recognition with a hybrid SVM-HMM system," in *Proc. of EURASIP*, 2005, pp. 993–996.

[22] A. Hagen and A. Morris, "Recent advances in the multi-stream hmm/ann hybrid approach to noise robust asr," *Computer Speech and Language*, vol. 19, no. 1, pp. 3–30, 2005.

[23] J. K. Baker, "The dragon system - an overview," *IEEE transactions on Aocustic Speech Signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

[24] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, April 1976.

[25] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," in *ETSI ES 201 108 Ver.1.1.3*, Apr. 2000.

[26] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," in *ETSI ES 202 050 Ver.1.1.3*, Nov. 2002.

[27] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed)*. New York:Wiley, 1963, ch. 15, pp. 209–243.

[28] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *Journal of Acoustical Society of America*, vol. 87, pp. 1738–1753, 1990.

[29] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *The Journal of the Acoustical Society of America*, vol. 33, Feb. 1961.

[30] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[31] J. Li, B. Liu, R. Wang, and L. Dai, "A complexity reduction of ETSI advanced front-end for DSR," in *in Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 1, 2004, pp. 61–64.

[32] A. Agarwal and Y. M. Cheng, "Two-stage Mel-warped Wiener filter for robust speech recognition, asru keystones," in *in Proc. ASRU*, 1999, pp. 12–15.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[34] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. of the Int. Conf. on Image Proc.*, vol. 3, 1998, pp. 173–178.

[35] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Proc. Of Inter. Conf. on Spoken Language Processing*, 2002, pp. 1925–1928.

[36] P. Scanlon and R. Reilly, "Feature analysis for automatic speechreading," in *Proc. of Workshop on Multimedia Signal Processing*, 2001, pp. 625–630.

[37] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[38] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, UK: Cambridge University Engineering Dept., 2001.

[39] "M2VTS database," `http://www.tele.ucl.ac.be/M2VTS/m2fdb.html`.

[40] "ITU-T recommendation P.56 : Objective measurement of active speech level," 1993.

[41] A. Martinez and A. Kak, "PCA versus LDA," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228–233, 2001.