HMM Based Text-to-Speech Synthesis for Telugu

Gugulothu Narendhar

A Thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Electrical Engineering

June 2016

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

G. Narendhar

(Signature)

G. Narendhar

(Gugulothu Narendhar)

EE14MTECH11004

(Roll No.)

Approval Sheet

This Thesis entitled HMM Based Text-to-Speech Synthesis for Telugu by Gugulothu Narendhar is approved for the degree of Master of Technology from IIT Hyderabad

V-Siva Rama Krishna

(Dr. Siva Rama Krishna Vanjari) Examiner
Dept. of Electrical Engineering
IITH

(Dr. Siva Kumar K) Examiner
Dept. of Electrical Engineering
IITH

KSRmurty

(Dr. Sri Rama Murty Kodukula) Adviser
Dept. of Electrical Engineering
IITH

C. Krishna Mohan

(Dr. C. Krishna Mohan) Chairman
Dept. of Computer Science
IITH

Acknowledgements

First, I would like to express my sincere gratitude to Professor Dr. Sri Rama Murthy Kodukula, IIT, Hyderabad, my thesis adviser, for his support, encouragement, comments and guidance during this thesis work. I am also grateful for the motivation and confidence he transmitted to me throughout the thesis. He has always given me valuable comments during this work and has also guided me at every stage of my M.Tech studies. I am also thankful for allowing me to be a part of the SIP Lab group, where this thesis has been mainly developed. The group provided me with the perfect environment to achieve my goals.

I would like to thank Associate Professor Dr. Prassanna Mahadeva, IIT Guwahati and Research Scholar Nagaraj Adiga, IIT Guwahati for welcoming me so warmly during my time there.

Finally, I would like to give my special thanks to my family for all their love and emotional support over the years and I would like to thank the Almighty God for the grace to embark on and finish this project.

Dedication

This thesis is dedicated to my parents. For their endless love, support and encouragement.

Abstract

This thesis describes a novel approach to build a general purpose working Telugu text-to- speech synthesis system (TTS) based on hidden Markov model (HMM) which is reasonably intelligible, natural sounding and flexible. There have been several attempts proposed to use HMM for constructing TTS systems. Most of such systems are based on waveform concatenation techniques.

To fully convey information present in speech signals, text-to-speech synthesis systems are required to have an ability to generate natural sounding speech with arbitrary speakers individualities and emotions (e.g., anger, sadness, joy). To represent all these factors the Mel- cepstral coefficients are extracted as spectral parameters. Excitation parameters are extracted using fundamental frequency (F0).

In the proposed approach, on the contrary, speech parameter (Mel-generalized cepstral coefficients, F0) sequences are generated from HMM directly based on maximum likelihood criterion. By considering the relationship between static and dynamic parameters, smooth spectral sequences are generated according to the statistics of static and dynamic parameters modeled by HMMs. As a result, natural sounding speech can be synthesized.

To synthesize speech, fundamental frequency (F0) patterns are modeled and generated. The conventional discrete or continuous HMMs, however, cannot be applied for modeling F0 patterns, since observation sequences of F0 patterns are composed of one dimensional continuous values and discrete symbol which represents voiced and unvoiced respectively. To overcome this problem, the HMM is extended so as to be able to model a sequence of observation vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols. It is shown that by using this extended HMM, referred to as the multi-space probability distribution HMM (MSD-HMM), spectral parameter sequences and F0 patterns can be modeled and generated in a unified framework of HMM.

Contents

Chapter 1

Introduction

1.1 General Background

Since speech is obviously one of the most important ways for human to communication, a great effort is going on to incorporate speech into human-machine communication environments. Now a days machines becoming more functional and prevalent, it demands for technologies in speech processing area, such as speech recognition, dialogue processing, speech understanding, speech synthesis, is increasing to establish high-quality man-machine communication with voice. Text-to-speech synthesis (TTS), one of the important technologies in speech processing, is a technique for creating speech signal from arbitrarily given text in order to transmit information from a machine to a human by voice. To fully convey information contained in speech signals, text-to-speech synthesis systems are required to have an ability to generate natural sounding speech with arbitrary speaker's individualities and emotions such as anger, sadness and joy.

A text-to-speech system makes it possible for people who cannot read, to be able to listen to near natural sounding utterances of written text in a language that they understand. People who wants to learn a new language can use a Text-to-Speech synthesis system to learn a language by listening to how a given text in the language of their interest is pronounced. The Text-to-Speech synthesis systems can also integrated to work with systems that recognise text from scanned documents and those that recognise a person's handwriting in the form of digitals.

For constructing such a Text-to-Speech synthesis system, the use of hidden Markov models (HMMs) become large. Hidden Markov models were successfully applied to model the sequence of speech spectra in speech recognition systems, and the performance of HMM-based speech recognition systems have been improved by techniques which utilize the flexibility of HMMs: context-dependent modeling, dynamic feature parameters, mixtures of Gaussian densities, tying mechanism. Hidden Markov models based approaches for speech synthesis is categorized as follows:

1. Transcription of text and segmentation of the speech database [1]
2. Construct inventory of speech segments [2-5].
3. Run-time selection of multiple instances of speech segments [4,6].

4. Speech synthesis from HMMs themselves [7-10].

Since most of this methods $1-3$ are, by using a waveform concatenation algorithm, e.g., PSOLA algorithm, a high quality synthetic speech produced. However, to obtain various voice characteristics, we require large amounts of speech database are, and it is difficult to collect, segment and store the large speech database. On the other hand, in method 4, voice characteristics of synthetic speech can be changed by transforming HMM parameters. To get this parameter generation algorithms [11], [12] for HMM-based speech synthesis have been proposed, and a speech synthesis system [9], [10] has been built using these algorithms. Actually, voice characteristics of synthetic speech can be changed by applying a speaker adaptation technique [13], [14] or a speaker interpolation technique [15]. The main characteristics of the system is use to dynamic feature: by including the dynamic coefficients in the feature vector, the dynamic coefficients of the speech parameter sequence generated in synthesis are constrained to be realistic and smooth as defined by the parameters of the HMMs.

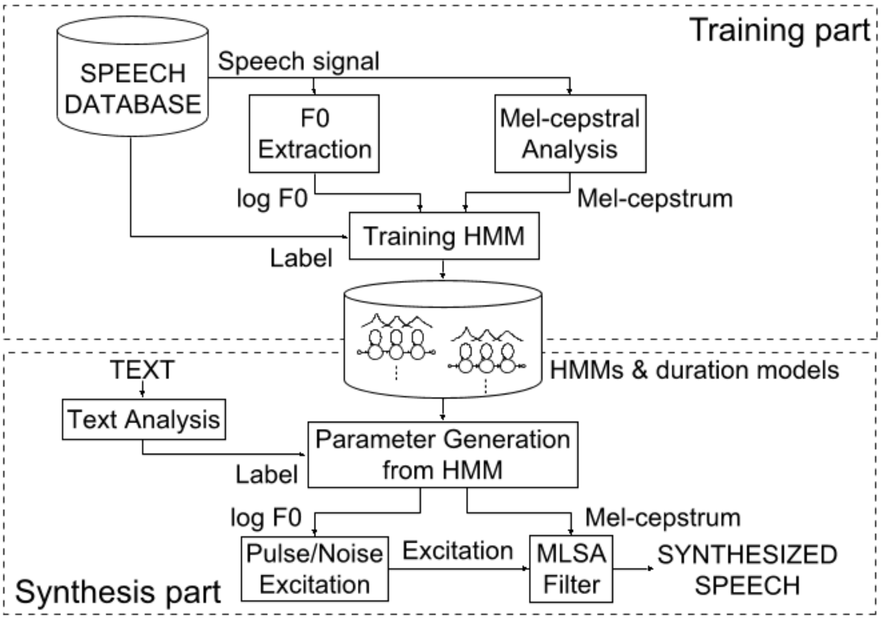

Figure 1.1: Flow chart of HMM-based Text-to-Speech Synthesis system.

1.2 Scope of thesis

The HMM-Based speech synthesis system is shown in Fig. 1.1. From this figure we have two parts. They are training and synthesis parts of the HMM-based synthesis system. In training phase, spectral parameters (e.g., mel-generalized cepstral coefficients) and excitation parameters (e.g., fundamental frequency F0) are extracted from the given or used speech database. These extracted speech parameters are modeled by context-dependent HMMs. In systhesis phase, a context-dependent label sequence is generated from the input text.From this context dependent label sequence a sentence HMM is built by concatenating context dependent HMMs for the given input text. By using the parameter generation algorithm, speech parameters (i.e. mel-generalized cepstral coefficients and fundamental frequencies) are generated from the sentence HMM. Lastly,a speech wave file is synthesized from the generated spectral(mel-generalized cepstral coefficients) and excitation parameters(fundamental frequencies) by using a synthesis filter.

In this report, it is assumed that spectral(mel-generalized cepstral coefficients) and excitation parameters(fundamental frequencies) include phonetic and prosodic information, respectively.

Chapter 2

Mel-cepstral Analysis and Synthesis

The speech analysis and synthesis technique is one of the key issues in vocoder dependent speech synthesis system, because characteristics of the spectral model, like stability of synthesis filter and performance of model parameters, effects the quality of synthesized speech, and speech synthesis system structure. Because of the above issues, the mel-generalized cepstral analysis adn synthesis technique[16] is employed to estimate spectral parameters and to synthesize the speech in the HMM-based speech synthesis system. This chapter deals with the mel-generalized cepstral analysis and synthesis technique, that is how the spectral feature parameters,i.e., mel-generalized cepstral coefficients, are extracted from speech signal from the speech database and speech is synthesized from by using these mel-generalized cepstral coefficients.

2.1 Source-filter Model

To deal with a speech waveform mathematically, a discrete-time model is used to represent sampled speech signals, as shown in Fig. 2.1. The transfer function $H(z)$ models the structure of vocal tract system part the speech. The excitation part contains voiced and unvoiced speech. For voiced speech the vocal folds oscillations are quasi periodic hence the voiced part is modellednu a quasi-periodic train of pulses. For unvoiced speech the vocal folds movements are random, i.e. it don't have periodicity associated with the unvoiced speech, hence random noise sequence is used to model the unvoiced sounds. To generate speech signals $x(n)$, the parameters of the model must change with time because speech is an outcome of time varying vocal-tract system driven by time-varying excitation.But for many speech sounds, it is reasonable to assume that the properties of the vocal tract and excitation remain fixed for some periods of 510 msec. Under such assumption, the excitation $e(n)$ is filtered by a slowly time-varying linear system $H(z)$ to generate speech signals $x(n)$.

By using this speech parameters i.e. the excitation $e(n)$ and the impulse response $h(n)$ of the vocal tract speech can be computed using the convolution sum expression.

$$x(n) = h(n) * e(n) \tag{2.1}$$

where $*$ represents discrete convolution. The details of digital signal processing and speech processing techniques are given in Ref. [17]

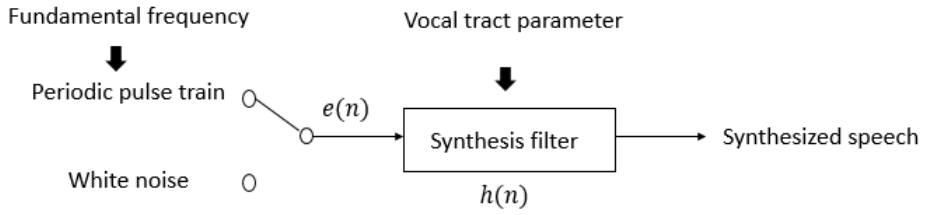

Figure 2.1: Source-filter modeling of speech.

2.2 Mel-Cepstral analysis

2.2.1 Spectral Model

Spectrum in the mel-cepstral analysis[16] $H(e^{jw})$ is represented by the M-th order mel-cepstral coefficients $\tilde{c}(m)$ as below

$$H(z) = exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m}, \tag{2.2}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |\alpha| < 1. \tag{2.3}$$

The phase of the all-pass transfer function $\tilde{z}^{-1} = e^{-j\tilde{w}}$ is given by

$$\tilde{w} = \arctan \frac{(1 - \alpha^2) \sin w}{(+ - \alpha^2) \cos w - 2\alpha} \tag{2.4}$$

The phase response w gives a good approximation to auditory frequency scale with appropriate value of α.

2.2.2 Spectral Criterion

In the unbiased estimation of log spectrum [18] it has been shown that the power spectral estimate $|H(e^{jw})|^2$, which is unbiased in a sense of relative power, is obtained by the following criterion E is minimized with respect to $\tilde{c}(m)_{m=0}^{M}$

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} (expR(w) - R(w) - 1)dw, \tag{2.5}$$

where

$$R(w) = logI_N(w) - log|H(e^{jw})|^2 \tag{2.6}$$

and $I_N(\omega)$ is the modified periodogram of a weakly stationary process $x(n)$ is given by

$$I_N(\omega) = \frac{|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\omega n}|^2}{\sum_{n=0}^{N-1} w^2(n)}$$

where $w(n)$ is the window whose length is N. To take the gain factor K outside from $H(z)$, we rewrite Eq.(2.2) as:

$$H(z) = exp \sum_{m=0}^{M} b(m)\phi_m(z) = K \cdot D(z), \tag{2.7}$$

where

$$K = exp\ b(0), \tag{2.8}$$

$$D(z) = exp \sum_{m=0}^{M} b(m)\phi_m(z), \tag{2.9}$$

and

$$b(m) = \begin{cases} c(m) & \text{if } m = M; \\ c(m) - \alpha b(m+1) & \text{if } 0 \leq m < M. \end{cases} \tag{2.10}$$

$$\phi_m(z) = \begin{cases} 1 & \text{if } m = 0; \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)} & \text{if } m \geq 1. \end{cases} \tag{2.11}$$

Since $H(z)$ is a minimum phase system, we can show that the minimization of E with respect to $\tilde{c}(m)_{m=0}^{M}$ is equivalent to that of

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{jw}|} dw, \tag{2.12}$$

with respect to

$$b = [b(1), b(2), ..., b(M)]^T \tag{2.13}$$

The gain factor K that minimizes E is obtained by setting $\frac{\partial E}{\partial K}$=0

$$K = \sqrt{\varepsilon_{min}} \tag{2.14}$$

where ε_{min} is the minimized value of ε. There exists only one minimum point because the criterion E is convex with respect to $\tilde{c}$. Consequently, the minimization problem of E can be solved using efficient iterative algorithm based on FFT and recursive formulas. In addition, the stability of model solution H(z) is always guaranteed.

2.3 Synthesis Filter

To synthesize speech from the mel-cepstral coefficients, itneeds to realize the exponential transfer function $D(z)$ of Eq. 2.9. Even the transfer function $D(z)$ is not a rational function, the MLSA (Mel Log Spectral Approximation) filter [19] approximates $D(z)$ with sufficient accuracy and becomes minimum phase IIR system. The complex exponential function $exp\omega$ is approximated by a rational function

$$exp\ \omega \simeq R_L(F(z))$$

$$= \frac{1 + \sum_{l=1}^{L} A_{L,l}\omega^l}{1 + \sum_{l=1}^{L} A_{L,l}(-\omega)^l} \tag{2.15}$$

Thus $D(z)$ is approximated as below

$$R_L(F(z) \simeq exp\,(F(z)) = D(z) \tag{2.16}$$

where $F(z)$ is defined by

$$F(z) = \sum_{m=1}^{M} b(m)\phi_m(z) \tag{2.17}$$

When $F(z)$ is expressed as

$$F(z) = F_1(z) + F_2(z) \tag{2.18}$$

the exponential transfer function is approximated in a cascade form

$$D(z) = exp\,F(z)$$

$$= exp\,F_1(z) \cdot exp\,F_2(z)$$

$$\simeq R_L(F_1(z)) \cdot R_L(F_2(z)) \tag{2.19}$$

as shown in Fig. 2.2. If

$$max_w|F_1(e^{j\omega})|, max_w|F_2(e^{j\omega})| < max_w|F(e^{j\omega})|, \tag{2.20}$$

we know that $R_L(F_1(e^{j\omega})) \times R_L(F_2(e^{j\omega}))$ approximates $D(e^{j\omega}))$ precisely than $R_L(F(e^{j\omega})$

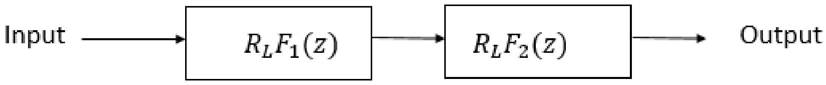

Figure 2.2: Two stage cascade structure of synthesis filter D(z).

Chapter 3

Modelling Speech Parameters Based on HMM

The performance of speech recognition system based on hidden Markov models was improved by including the dynamic features of speech parameters. Thus, if there exists is a technique to generate speech parameters from HMMs which includes the dynamic features, it will be useful for speech synthesis system to get the smooth synthesized speech. In this chapter a speech parameter generation technique from HMMs which include the dynamic features is discussed.

3.1 Modeling Spectal Parameters

3.1.1 Introduction to Continuous density HMM

In this report, a continuous density HMM is used to model the the vocal tract as in the speech recognition systems. The continuous density Markov model is a finite state machine which makes one state transition at each time unit. Firstly, a decision has to be taken in order to decide which state to occupy i.e. to succeed to next state which may be that state also. Then have to generate output vector based on the probability density function(pdf) for the present state. An HMM is a doubly stochastic random process, i.e. it is models the state transition probabilities between states and the output probabilities for each state as in [33].

One way to understand HMMs is to consider each state as a model of a segment of speech. In figure 3.1 how the speech utterance using a N-state left-to-right HMM, in which each state is modeled by a multi-mixture Gaussian model is discussed. Assume that this utterance(parameterized by speech analysis as the D-dimensional observation vector o_t) is divided into N segments d_i which are represented by the states S_i . The transition probability a_{ij} defines the probability of making transition from state i to state j and satisfies the stochastic constraint $a_{ii} + a_{ij} = 1$. Then, each state is modeled by a M -mixtures Gaussian density function:

$$b_j(o_t) = \sum_{k=1}^{M} c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})$$

$$= \sum_{k=1}^{M} c_{jk} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} exp\{-\frac{1}{2}(o_t - \mu_{jk})^T \Sigma_{jk}^{-1} (o_t - \mu_{jk})\}, \quad (3.1)$$

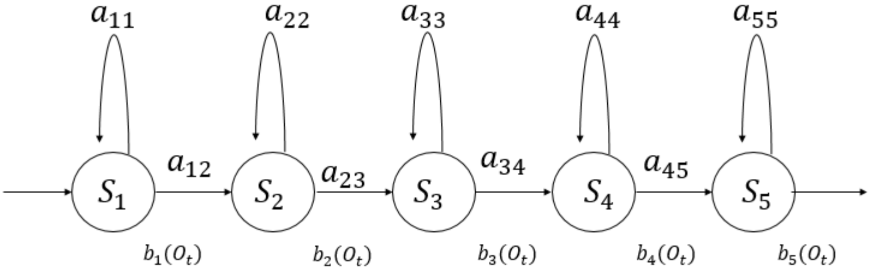

Figure 3.1: A five state HMM.

where c_{jk}, μ_{jk} and Σ_{jk} represnts the mixture coefficients, D-dimensional mean vector and $D \times D$ covariance matrix(full covariance matrix) for the k-th mixture component in the j-th state, respectively. Above covariance matrix can be restricted to the diagonal elements if the elements of the feature vector are independent. $|\Sigma_{jk}|$ is the determinant of Σ_{jk}, and Σ_{jk}^{-1} is the inverse of Σ_{jk}. The mixture gains c_{jk} satisfy the stochastic constraint

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 1 \leq j \leq N \quad (3.2)$$

$$c_{jk} \geq 0, \quad 1 \leq j \leq N, 1 \leq k \leq M \quad (3.3)$$

so pdf is properly normalized, i.e.,

$$\int_{-\infty}^{\infty} b_j(o)do, \quad 1 \leq j \leq N \quad (3.4)$$

Since the pdf of Eq. (3.1) can be used to approximate, any finite, continuous density function, it can be used in a wide range of problems and is widely used for acoustic modeling.

For simpicity and convinience, the complete parameter set of the HMM model is represented by a notation

$$\lambda = (A, B, \pi) \quad (3.5)$$

where $A = \{a_{ij}\}, B = \{b_j(o)\}$ and $\pi = \{\pi_i\}$. π_i is the initial state distribution of state i, and it have the property

$$\pi = \begin{cases} 0 & \text{if } i \neq 1; \\ 1 & \text{if } 1 = 1. \end{cases} \quad (3.6)$$

in the left-to-right HMM model.

3.1.2 Probability

To calculate the probanility of the obeservation sequence $O = (o_1, o_2, , o_T)$ given the model λ, i.e.$P(O|\lambda)$, forward-backward algorithm is used. Direct calculation $P(O|\lambda)$ without using farward-backward algorithm the computational complexity is high. It requires order of $2TN^2$ order calculations. On the other hand, If the forward-backward algorithm is epmloyed, then computational complexity is less and it requires on the order of N^2T calculations, and it is computationally feasible. Forward-backward algorithm is discussed in the following section as in [33].

The forward algorithm

Forward variable $\alpha_t(i)$ is defined as

$$\alpha_t(i) = P(o_1, o_2, , o_T, q_t = i|\lambda) \tag{3.7}$$

that is, the probability of the partial observation sequence from time 1 to time t and state i at time t, given the model parameters λ . We solve $\alpha_t(i)$ inductively, as discussed below

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N. \tag{3.8}$$

2. Induction

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)a_{ij}]b_j(o_{t+1}), \quad 1 \leq t \leq T-1, 1 \leq i \leq N. \tag{3.9}$$

3. Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.10}$$

The backward algorithm

As in the forward algorithm, let us assume the backward variable$\beta_t(i)$ is defined as

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, , o_T|q_t = i, \lambda) \tag{3.11}$$

that is, the probability of the partial observation sequence from time t to time T , given state i at time t and the model parameters λ. We solve $\beta_t(i)$ using induction as discussed below

1. Initialization

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \tag{3.12}$$

2. Induction

$$\beta_t(i) = \sum_{i=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, 1 \leq i \leq N \tag{3.13}$$

3. Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \beta_1(i) \tag{3.14}$$

By using trellis structure we calculate forward-backward probability as shown in Fig. 3.2. In this figure, x-axis represents obeservation sequence and y-axis represents states of Markov model.All the possible state sequence will remerge into these N nodes no matter how long the observation sequence in Trellis structure. For forward algorithm, at times $t = 1$, we have to calculate the values of $\alpha_1(i), 1 \leq i \leq N$, and at time $t = 2, 3, ..., T$ we o only calculate values of $\alpha_t(j), 1 \leq j \leq N$, in which each calculation involves only the N previous values of $\alpha_{t-1}(i)$ because each of the N states can be reached from only the N states at the previous time slot. Because of this advantage the order of probabilty is reduced in the forward-backward algrorithm.

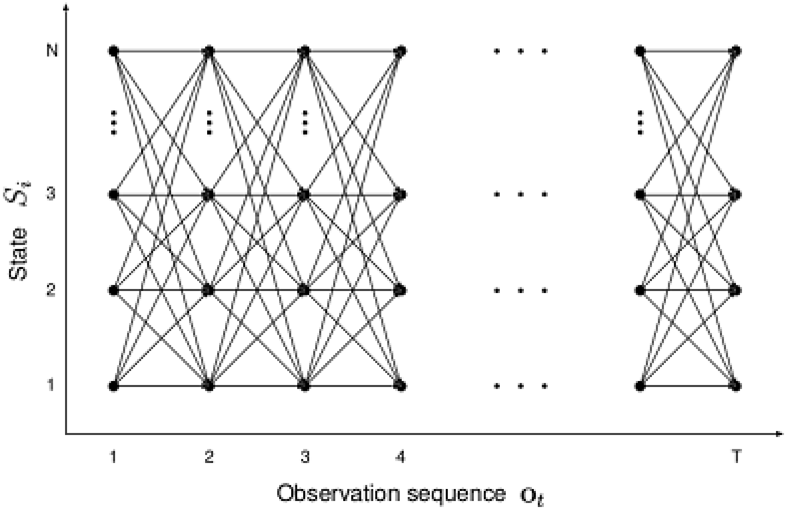

Figure 3.2: Computation of forward-backward algorithm by using a trellis structure of observation t and state i.

3.1.3 Parameter estimation of continuous density HMM

Adjusting the model parameters (A, B, π) by a method is very difficult. It has to satisfy a certain optimization criterion.There is no exiting technique to analytically solve for the model parameter set which maximizes the probability of the observation sequence. But we can choose a method in which $\lambda = (A, B, \pi)$ such that its likelihood, $P(O|\lambda)$, is locally maximized using an iterative procedure which is known as the EM(expectation-maximization) methosd [21], [22].

Now we define a parameter that is the probability of being in state i at time t, and state j at time $t+1$, given the model and the observation sequence to reestimate HMM parameters, is defined below

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O, \lambda), \tag{3.15}$$

By using forward and backward variables, $\xi_t(i,j)$ can be written as

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j|O,\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}\beta_{t+1}(j))}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i) a_{ij} b_j(o_{t+1}\beta_{t+1}(j))} \tag{3.16}$$

By using $\xi_t(i,j)$, we now define a new parameter, i.e. the probability of being in state i at time t, given the entire observation sequence and the model parameters, and it is represented as

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{3.17}$$

Q-function

We use Baum's auxiliary function by maximizing it to reesitmate the formulas over λ.

$$Q(\lambda, \lambda^{'}) = \sum_{q} P(O, q|\lambda^{'}) \log P(O, q|\lambda) \tag{3.18}$$

Because

$$Q(\lambda, \lambda^{'}) \geq Q(\lambda, \lambda^{'}) \Rightarrow P(O, q|\lambda^{'}) \geq P(O, q|\lambda) \tag{3.19}$$

We can maximize the function $Q(\lambda, \lambda^{'})$ over λ to improve $\lambda^{'}$ in the sense of increasing the likelihood $P(O, q|\lambda)$

Maximization of Q-function

We derive parameters of λ for the given observation sequence O and model parameters λ, which maximize $Q(\lambda, \lambda^{'})$. $P(O, q|\lambda)$ is written as below

$$P(O, q|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{3.20}$$

$$\log P(O, q|\lambda) = \log \pi_{q_0} + \sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log b_{q_t}(o_t) \tag{3.21}$$

Now the Q-function in Eq 3.18 can be written as

$$Q(\lambda, \lambda^{'}) = Q_{\pi}(\lambda^{'}, \pi) + \sum_{t=1}^{T} Q_{a_i}(\lambda^{'}, a_i) + \sum_{t=1}^{T} Q_{b_i}(\lambda^{'}, b_i)$$

$$= \sum_{i=1}^{N} P(O, q_0 = i|\lambda^{'}) \log \pi_i + \sum_{j=1}^{N} \sum_{t=1}^{T} P(O, q_{t-1} = i, q_t = j|\lambda^{'}) \log a_{ij} + \sum_{t=1}^{T} P(O, q_t = i|\lambda^{'}) \log b_i(o_t) \tag{3.22}$$

where

$$\pi = [\pi_1, \pi_2, ..., \pi_N] \tag{3.23}$$

$$a_i = [a_{i1}, a_{i2}, ..., a_{iN}], \tag{3.24}$$

and b_i is the parameter vector that defines $b_i()$. The following the stochastic constraints have to be satisfied to maximize the Eq. 3.22 and the parameter set is λ

$$\sum_{j=1}^{N} \pi_j = 1, \tag{3.25}$$

$$\sum_{j=1}^{N} a_{ij} = 1, \tag{3.26}$$

$$\sum_{k=1}^{M} c_{jk} = 1, \tag{3.27}$$

$$\int_{-\infty}^{\infty} b_j(o) do = 1, \tag{3.28}$$

can be derived as

$$\pi_i = \frac{\alpha_0(i)\beta_0(i)}{\sum_{j=1}^{N} \alpha_T(j)} = \gamma_0(i) \tag{3.29}$$

$$a_{ij} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} b_j(o_t) \beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} = \frac{\sum_{t=1}^{T} \xi_{t-1}(i,j)}{\sum_{t=1}^{T} \gamma_{t-1}(i)} \tag{3.30}$$

The reestimation formulas for the coefficients of the mixture density, i.e, c_{jk} , μ_{jk} and Σ_{jk} are of the form

$$c_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{3.31}$$

$$\mu_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.32}$$

$$\Sigma_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (o_t - \mu_{jk})(o_t - \mu_{jk})^{'}}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.33}$$

where $\gamma_t(j,k)$ is the probability of being in state j at time t with the kth mixture component accounting for o_t , i.e.,

$$\gamma_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \frac{c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{m=1}^{M} N(o_t, \mu_{jk}, \Sigma_{jk})} \tag{3.34}$$

3.2 Modeling F0 Parameters

In the excitation part the F0 pattern contains continuous values in the voiced region because of the periodic movement of vocal folds during speech production and a discrete symbol in the unvoiced region because of the random movement of vocal folds at the time speech production as shown in Fig.3.3. Hence, we can not employ the discrete or continuous HMMs to model the F0 patterns. There are several methods[23] to handle the unvoiced region of the speech: (i) We use a random vector generated from a probability density function (pdf) with a large variance to replace each unvoiced symbol and then we use the continuous HMMs to model the random vectors explicitly [24], (ii) We assume that F0 values exists always but they are unobservable in the unvoiced region and applythe expectation maximization (EM) algorithm [25]. In this section, we discuss a new form of HMM to model the F0 patterns, in this new HMM the state output probabilities are modelled by multi-space probability distributions (MSDs) is described. That is nothing but modelling the F0 patterns as linear combination of continuous and discrete HMMs

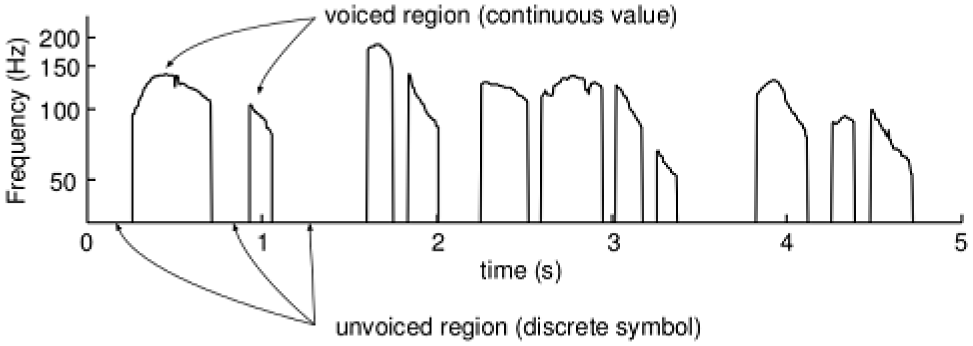

Figure 3.3: F0 patterns

3.2.1 Introduction to Multi-Space Probability Distribution

Now let us consider a sample space Ω which consists of G spaces as in [33] as shown in Fig. 3.4

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{3.35}$$

where Ω_g is an n_g -dimensional real space R^{n_g} , and a space index g is use to specify it. Each space Ω_g has its probability w_g , i.e., $P(\Omega_g) = w_g$, subject to the constraint $\sum_{g=1}^{G} w_g = 1$. Each space has a probability density function $N_g(x), x\epsilon R^{n_g}$ If n_g ¿ 0, where $\int_R^{n_g} N_g(x)dx = 1$. Let us assume t Ω_g contains only one sample point if $n_g = 0$. Accordingly, let $P(E)$ be the probability distribution, then

$$P(\Omega) = \sum_{g=1}^{G} P(\Omega_g) = \sum_{g=1}^{G} \Omega_g \int_R^{n_g} N_g(x)dx = 1 \tag{3.36}$$

it is observed that, although $N_g(x)$ does not exist for $n_g = 0$. since Ω_g contains only one sample point, for notational simplicity we define as $N_g(x) \equiv 1$ for $n_g = 0$.

Each event E, which we take in this report is represented by a random variable o, it consists of a continuous random variable $x \epsilon R^n$ and a set of space indices X, that is,

$$o = (x, X) \tag{3.37}$$

where all spaces specified by X are n-dimensional. The observation probability of o is defined by

$$b(o) = \sum_{g \epsilon S(o)} \omega_g N_g(V(o)) \tag{3.38}$$

where

$$V(o) = x, \;\; S(o) = X. \tag{3.39}$$

Observations patterns are shown in Fig. 3.4. An observation o_1 consists of three-dimensional vector $x_1 \epsilon R^3$ and a set of space indices X_1 = 1, 2, G. Thus the random variable x is drawn from one of three spaces $\Omega_1, \Omega_2, \Omega_G \, \epsilon \, R^3$, and its probability density function is given by $w_1 N_1(x) + w_2 N_2(x) + w_G N_G(x)$.

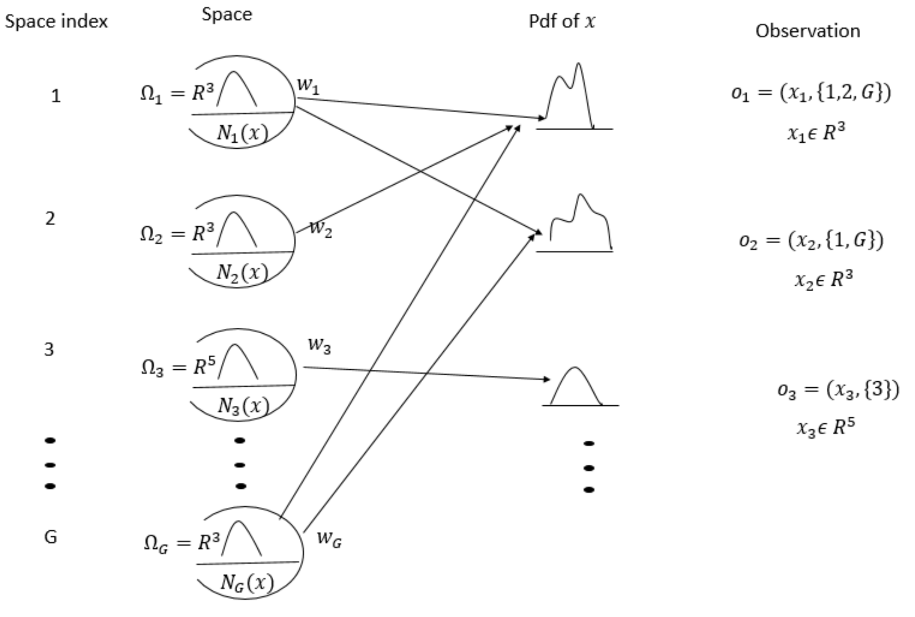

Figure 3.4: Example of multi-space probability distribution and observations.

Above defined probabilty distribution is known as multi-space probability distribution (MSD) which is same as the discrete distribution and the continuous distribution when $n_g \equiv 0$ and $n_g \equiv m > 0$, respectively. Further, if $S(o)1, 2, ..., G$, the continuous distribution is represented by a G-mixture probability density function.

3.2.2 Multi-space distribution HMM

The MSD-HMMs output probability in each state is given by the multi-space probability distribution as in [33]. An N -state MSD-HMM λ is specified by initial state probability distribution $\pi = \{\pi_j\}_{j=1}^{N}$, the state transition probability distribution $A = \{a_{ij}\}_{i,j=1}^{N}$, and state output probability distribution $B = b_i(\cdot)_{i=1}^{N}$, where

$$b_i(o) = \sum_{g \epsilon S(o)} \omega_{ig} N_{ig}(V(o)), \quad i = 1, 2, \cdots, N. \tag{3.40}$$

As shown in Fig. 3.5, each state i has G probability density functions $N_{i1}(\cdot), N_{i2}(\cdot), \cdots, N_{iG}(\cdot)$, and their weights $w_{i1}, w_{i2}, \cdots, w_{iG}$.

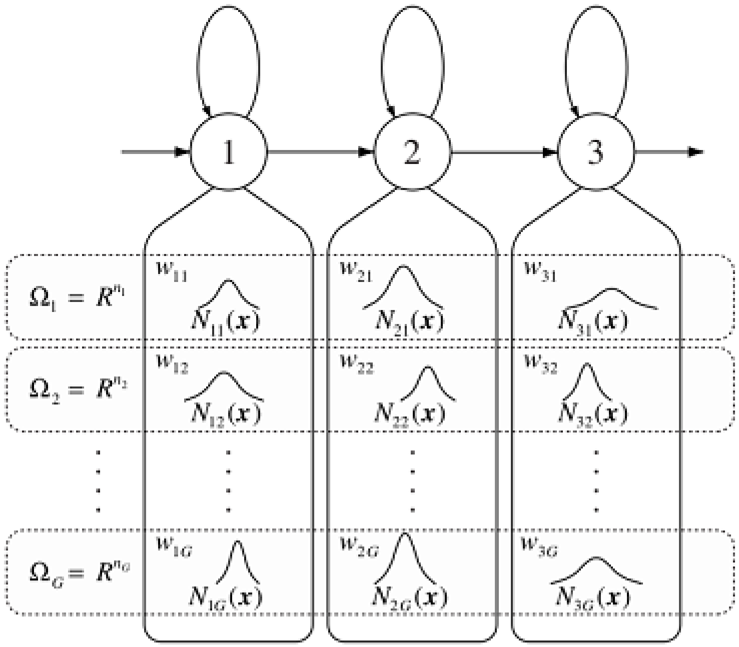

Figure 3.5: An HMM based on multi-space probability distribution.

Observation probability of $O = o_1, o_2, \cdots, o_T$ is written as

$$P(O|\lambda) = \sum_{all\ q} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t)$$

$$P(O|\lambda) = \sum_{all\ q} \prod_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t l_t} N_{q_t l_t}(V(o_t)) \tag{3.42}$$

where $q = q_1, q_2, \cdots, q_T$ is a possible state sequence, $l = l_1, l_2, \cdots, l_T \epsilon S(o_1) \times S(o_2) \times \cdots \times S(o_T)$ is a sequence of space indices which is possible for the observation sequence O, and $a_{q_0 j}$ denotes π_j. Now let us define the forward and backward variables

$$\alpha_t(i) = P(o_1, o_2, , o_T, q_t = i|\lambda) \tag{3.43}$$

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, , o_T | q_t = i, \lambda) \tag{3.44}$$

Now Eq 3.42 can be calculated as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \beta_1(i). \tag{3.45}$$

We use forward and backward variables to calculate the reestimation formulas which will be derived in the following section

3.2.3 MSD-HMM training using Reestimation algorithm

For a particular choice of MSD-HMM and a observation sequence O the objective in maximum likelihood estimation is to maximize the observation likelihood $P(O|\lambda)$ given by (3.42), over all parameters in λ. In a manner similar to [21], [22], we derive reestimation formulas for the maximum likelihood estimation of MSD-HMM.

Q-function

An auxiliary function $Q(\lambda^{'}, \lambda)$ of current parameters $\lambda^{'}$ and new parameter λ is defined as follows:

$$Q(\lambda^{'}, \lambda) = \sum_{all\ \ q,l} P(O, q, l|\lambda^{'}) \log P(O, q, l|\lambda) \tag{3.46}$$

Let us assume $N_{ig}(\cdot)$ to be the Gaussian density with mean vector μ_{ig} and covariance matrix Σ_{ig} .

$$Q(\lambda^{'}, \lambda) \geq Q(\lambda^{'}, \lambda^{'}) \rightarrow P(O, \lambda) \geq P(O, \lambda^{'})$$

Maximization of Q-function

For the given model parameter λ and observation sequence O, let us derive the parameters of λ which maximize $Q(\lambda^{'}, \lambda)$. From Eq 3.42, $\log P(O, q, l|\lambda)$ can be written as

$$\log P(O, q, l|\lambda) = \sum_{t=1}^{T} (\log a_{q_{t-1}q_t} + \log \omega_{q_t l_t} + \log N_{q_t l_t}(V(o_t))) \tag{3.47}$$

Hence Q-function (3.46) can be written as

$$\begin{aligned} Q(\lambda^{'}, \lambda) = \sum_{i=1}^{N} P(O, q_1 = i|\lambda^{'}) \log \pi_i + \sum_{i,j=1}^{N} \sum_{t=1}^{T-1} P(O, q_t = i, q_{t+1} = j|\lambda^{'}) \log a_{ij} \\ + \sum_{i=1}^{N} \sum_{g=1}^{G} \sum_{t \epsilon T(o,g)} P(O, q_t = i, l_t = g|\lambda^{'}) \log N_{ij}(V(o_t)) \end{aligned} \tag{3.48}$$

where

$$T(O,g) = \{t | g \,\epsilon\, S(o_t)\} \tag{3.49}$$

The parameter set $\lambda = (\pi, A, B)$ which used to maximize Eq.3.48, subjected to the following stochastic constraints $\sum_{i=1}^{N} \pi_i = 1, \sum_{j=1}^{N} a_{ij} = 1$ and $\sum_{g=1}^{G} \omega_g = 1$, can be derived as

$$\pi_i = \sum_{g \epsilon S(o_1)} \gamma^{'}(i,g) \tag{3.50}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi^{'}(i,j)}{\sum_{t=1}^{T-1} \sum_{g \,\epsilon\, S(o_t)} \gamma^{'}(i,g)} \tag{3.51}$$

$$\omega_{ig} = \frac{\sum_{t \epsilon T(O,g)} \gamma^{'}(i,g)}{\sum_{h=1}^{G} \sum_{t \,\epsilon\, T(O,h)} \gamma^{'}(i,h)} \tag{3.52}$$

$$\mu_{ig} = \frac{\sum_{t \,\epsilon\, T(O,g)} \gamma^{'}(i,g) V(o_t)}{\sum_{t \,\epsilon\, T(O,g)} \gamma^{'}(i,g)}, n_g > 0 \tag{3.53}$$

$$\Sigma_{ig} = \frac{\sum_{t \,\epsilon\, T(O,g)} \gamma^{'}(i,g)(V(o_t) - \mu_{ig})(V(o_t) - \mu_{ig})^T}{\sum_{t \,\epsilon\, T(O,g)} \gamma^{'}(i,g)} \tag{3.54}$$

Now the forward variable $\alpha_t(i)$ and backward variable $\beta_t(i)$ are used to calculate $\gamma_t(i,h)$ and $\xi_t(i,j)$ as below

$$\gamma_t(i,h) = P(q_t = i, l_t = h | O, \lambda)$$

$$\gamma_t(i,h) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \cdot \frac{\omega_{ih} N_{ih}(V(o_t))}{\sum_{g \,\epsilon\, S(o_t)} \omega_{ig} N_{ig}(V(o_t))} \tag{3.55}$$

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j | O, \lambda)$$

$$\xi_t(i,h) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{h=1}^{N} \sum_{k=1}^{N} \alpha_t(h) a_{hk} b_k(o_{t+1}) \beta_{t+1}(k)} \tag{3.56}$$

3.2.4 Application to F0 pattern modeling

The MSD-HMM contains both the continuous mixture HMM and the discrete HMM as special cases because the multi-space probability distribution includes both discrete distribution and the continuous distribution as in [33].

From the observation of F0, it has a continuous value in the voiced region, and there exist no value for the unvoiced region. Hence we model this observation sequence assuming that the observed F0 value occurs from one-dimensional spaces for the voiced and the the zero-dimensional space defined for 'unvoiced' symbol as in section 3.2.1, that is by setting $n_g = 1$ $(g = 1, 2, \cdots, G-1)$,$n_g = 0$ and

$$S(o_t) = \begin{cases} \{1, 2, \cdots, G-1\}, & \text{(voiced)}; \\ \{G\} & (unvoiced) \end{cases} \tag{3.57}$$

the MSD-HMM can handle F0 patterns including the unvoiced region without anyassumption. In this case, the observed F0 value is assumed to be drawn from a continuous $(G-1)$-mixture probability density function.

Chapter 4

Generation of Speech parameters from HMM

Speech recognition systems performance imoroved by including the dynamic features of speech. If there exists a technique for speech parameter generation from HMMs which integrates the dynamic features also, will be very useful in speech synthesis. Here we derive a technique for speech parameter generation from HMMs which integrates the dynamic features as in [33].

4.1 Speech parameter generation based on maximum likelihood criterion

We derive an algorithm to determine speech parameter vector sequence for a given continuous mixture HMM λ,

$$O = [o_1^T, o_2^T, \cdots, o_T^T]^T \tag{4.1}$$

in such a way that

$$P(O|\lambda) = \sum_{all\ Q} P(O, Q|\lambda) \tag{4.2}$$

is maximized with respect to O, where

$$Q = \{(q_1, i_1), (q_2, i_2), \cdots, (q_T, i_T)\} \tag{4.3}$$

is the state and mixture sequence, i.e.,(q, i) indicates the i-th mixture of state q. Let us consider speech parameter vector o_t contains the static feature $c_t = [c_t(1), c_t(2), \cdots, c_t(M)]^T$ e.g., mel-generalized cepstral coefficients) and dynamic feature $\Delta c_t, \Delta^2 c_t$ (e.g., delta and delta-delta cepstral coefficients, respectively), i.e., $o_t[c_t^T, \Delta c_t^T, \Delta^2 c_t^T]^T$ Now we use following linear combination to calculate the dynamic feature vectors

$$\Delta c_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} \omega^{(1)}(\tau) c_{t+\tau} \tag{4.4}$$

$$\Delta^2 c_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} \omega^{(2)}(\tau) c_{t+\tau} \tag{4.5}$$

Let us consider the algorithms in [11], [12] to solve the following conditions

Case 1.
For given λ and Q, maximize $P(O|Q,\lambda)$ with respect to O under the conditions (4.4),(4.5)

Case 2.
For given λ, maximize $P(O,Q|\lambda)$ with respect to Q and O under the conditions (4.4),(4.5).

Case 3.
For given λ, maximize $P(O|\lambda)$ with respect to O under the conditions (4.4),(4.5).

4.1.1 Maximizing $P(O|Q,\lambda)$ with respect to O

Now let us consider maximizing $P(O|Q,\lambda)$ with respect to O for a fixed state and mixture sequence Q. The logarithm of $P(O|Q,\lambda)$ can be written as

$$\log P(O|Q,\lambda) = -\frac{1}{2} O^T U^{-1} O + O^T U^{-1} M + K \tag{4.6}$$

where

$$U^{-1} = diag[U_{q_1,i_1}^{-1}, U_{q_2,i_2}^{-1}, \cdots, U_{q_T,i_T}^{-1}] \tag{4.7}$$

$$M = [U_{q_1,i_1}^{T}, U_{q_2,i_2}^{T}, \cdots, U_{q_T,i_T}^{T}] \tag{4.8}$$

μ_{q_t,i_t} and U_{q_t,i_t} are the $3M \times 1$ mean vector and the $3M \times 3M$ covariance matrix, respectively, associated with i_t -th mixture of state q_t, and the constant K is independent of O.

It is evident that $P(O|Q,\lambda)$ is maximized when $O = M$ without the conditions (4.4), (4.5), i.e., the speech parameter vector sequence becomes a sequence of the mean vectors. Now arranege the Conditions (4.4), (4.5) in a matrix form:

$$O = WC \tag{4.9}$$

where

$$C = [c_1, c_2, \cdots, c_T]^T \tag{4.10}$$

$$W = [\omega_1, \omega_2, \cdots, \omega_T]^T \tag{4.11}$$

$$\omega_t = [\omega_t^{(0)}, \omega_t^{(1)}, \omega_t^{(2)}] \tag{4.12}$$

$$\omega_t^{(n)} = \begin{array}{l} [O_{M\times M_{1st}}, \cdots, O_{M\times M}, \omega^{(n)}(-L_-^{(n)})_{t-L_-^{(n)}-th} I_{M\times M}, \\ \cdots, \omega^{(n)}(0) I_{M\times M}, \cdots, \omega^{(n)}(L_+^{(n)})_{t+L_+^{(n)}-th} I_{M\times M}, \\ O_{M\times M}, \cdots, O_{M\times M_{T-th}}]^T, \quad n = 0,1,2 \end{array} \tag{4.13}$$

where $L_-^0 = L_+^0$ =0, and $\omega^{(0)}(0)$=1.Under the condition (4.9), maximizing $P(O|Q,\lambda)$ with respect to O is equivalent to that with respect to C. By setting

$$\frac{\partial \log P(WC|Q,\lambda)}{\partial C} = 0 \tag{4.14}$$

we obtain a set of equations

$$W^T U^{-1} WC = W^T U^{-1} M^T \tag{4.15}$$

For the solution of Eq.4.15 direclty, we require $O(T^3M^3)$ operations because $W^TU^{-1}W$ is a $TM \times TM$ matrix. By using the special structure of $WU^{-1}W$, Eq.4.15 solved by using the Cholesky decomposition with less operations [27].

4.1.2 Maximizing $P(O,Q|\lambda)$ with respect to O and Q

This problem is solved by calculating $\max_c P(O,Q|\lambda) = \max_c P(O|Q,\lambda)P(Q|\lambda)$for all Q. But it is not possible in practical because there are too many combinations of Q. Hence We use the algorithms developed in [11], [12].

state duration densities are incorporated in HMMs to control temporal structure of speech parameter sequence . The probability $P(O,Q|\lambda)$ can be written as $P(O,i|q,\lambda)P(q|\lambda)$, where $q = \{q_1, q_2, \cdots, q_T\}, i = \{i_1, i_2, \cdots, i_T\}$, and the state duration probability $P(q|\lambda)$ is given by

$$\log P(q|\lambda) = \sum_{n=1}^{N} \log p_{q_n}(d_{q_n}) \tag{4.16}$$

where the total number of states which have been visited during T frames is N , and $p_{q_n}(d_{q_n})$ is the probability of d_{q_n} consecutive observations in state q_n . If we determine the state sequence q only by $P(q|\lambda)$ independently of O, maximizing $P(O,Q|\lambda) = P(O,i|q,\lambda)P(q|\lambda)$ with respect to O and Q is equivalent to maximizing $P(O,i|q,\lambda)$ with respect to O and i. Furthermore, if we assume that state output probabilities are single-Gaussian, i is unique. Therefore, the solution is obtained by solving (4.15) in the same way as the Case 1.

4.1.3 Maximizing $P(O|\lambda)$ with respect to O

In this case we derive an algorithm based on an EM algorithm, which find a critical point of the likelihood function $P(O|\lambda)$. An auxiliary function of new parameter vector sequence O' and current parameter vector sequence O is defined as

$$Q(O,O') = \sum_{all\ \ Q} P(O,Q|\lambda) \log P(O',Q|\lambda) \tag{4.17}$$

It can be shown that by substituting O' which maximizes $Q(O,O')$ for O, the likelihood increases unless O is a critical point of the likelihood. Equation (4.17) can be written as

$$Q(O,O') = P(O|\lambda)\{-\frac{1}{2}O^{'T}\overline{U^{-1}}O' + O^{'T}\overline{U^{-1}M} + \overline{K} \tag{4.18}$$

where

$$\overline{U^{-1}} = diag[\overline{U_1^{-1}}, \overline{U_2^{-1}}, \cdots, \overline{U_T^{-1}}] \tag{4.19}$$

$$\overline{U_t^{-1}} = \sum_{q,i} \gamma_t(q,i) U_{q,i}^{-1} \tag{4.20}$$

$$\overline{U^{-1}M} = [\overline{U_1^{-1}\mu_1}^T, \overline{U_2^{-1}\mu_2}^T, \cdots, \overline{U_T^{-1}\mu_T}^T]^T \tag{4.21}$$

$$\overline{U_t^{-1}\mu_t} = \sum_{q,i} \gamma_t(q,i) U_{q,i}^{-1} \mu_{q,i} \tag{4.22}$$

and the constant $\overline{K}$ is independent of $O^{'}$. The occupancy probability $\gamma_t(q,i)$ defined by

$$\gamma_t(q,i) = P(q_t = (q,i)|O, \lambda) \tag{4.23}$$

can be calculated with the forward-backward inductive procedure. Under the condition $O^{'} = WC^{'}$, $C^{'}$ which maximizes $Q(O, O^{'})$ is given by the following set of equations:

$$W^T\overline{U^{-1}}WC^{'} = W^T\overline{U^{-1}M} \tag{4.24}$$

The above set of equations has the same form as (4.15). Accordingly, it can be solved by the algorithm for solving (4.15).
If we determine the state sequence q only by $P(q|\lambda)$ independently of O in a manner similar to the previous section, only the mixture sequence i is assumed to be unobservable. Further, we can also assume that Q is unobservable but phoneme or syllable durations are given.

4.2 Dynamic features effect and Telugu Example

A simple experiment of synthesizing speech was carried out by using the parameter generation algorithm.I used phonetically balanced 1000 sentences from Telugu speech database for training. The HMMs used are continuous Gaussian model. All the HMM models were 3-state left-to-right models with no skips. After the training the duration densities were calculated. Output feature vector consists of 25 mel-generalized cepstral coefficients including the zeroth coefficient, and their delta and delta-delta coefficients. Mel-generalized cepstral coefficients were calculated from the mel-cepstral analysis which we discussed earlier. A 25-ms Blackman window is applied to the signal with a window shift of 5-ms.

I observed the parameter generation in the case 1, in which parameter sequence O maximizes $P(O|Q, \lambda)$. From the results of Viterbi alignment of natural speech we estimate the state sequence Q. The spectral parameters(mel-generalized cepstral coefficients) are calculated from the mel-spectral analysis. If we don't consider the dynamic features, the parameter sequence which maximizes $P(O|Q, \lambda)$ becomes a sequence of the mean vectors which results in glitches in the synthesized speech. Hence we can see that integration of dynamic features improves smoothness of generated speech spectra.

Chapter 5

HMM-based Text-to-Speech System for Construction

In this phonetic informations and prosodic informations are modeled simultaneously by HMM. In this system, mel-generalized cepstral coefficients, fundamental frequency (F0) and state duration are modeled by continuous density HMMs, multi-space probability distribution HMMs and multi-dimensional Gaussian distributions, respectively. The distributions for mel-spectrum, fundamental frequency (F0), and the state duration are clustered independently by using a decision-tree based context clustering technique. In this chapter we discuss about the HMM modelling of th efeature vector, HMM structure and how to train context-dependent HMM as discussed in [33].

5.1 Dynamic feature calculation

Here spectral parameters are Mel-cepstral coefficients. These Mel-cepstral coefficient vectors c are obtained by a mel-cepstral analysis technique [16]. Their dynamic features Δc and $\Delta^2 c$ are calculated as linear combination of present and previous cofficients

$$\Delta c_t = -\frac{1}{2}c_{t-1} + \frac{1}{2}c_{t+1} \tag{5.1}$$

$$\Delta^2 c_t = \frac{1}{4}c_{t-1} - \frac{1}{2}c_t + \frac{1}{4}c_{t+1} \tag{5.2}$$

We follow the same to calculate dynamic features F0

$$\delta p_t = -\frac{1}{2}p_{t-1} + \frac{1}{2}p_{t+1} \tag{5.3}$$

$$\delta^2 p_t = \frac{1}{4}p_{t-1} - \frac{1}{2}p_t + \frac{1}{4}p_{t+1} \tag{5.4}$$

5.2 Modeling of spectrum and F0

The sequence of mel-cepstral coefficient vector and F0 pattern were modeled by a continuous density HMM and multi-space probability distribution HMM, respectively.
By using embedded traing we construct spectrum and F0 models because the embedded training does not need label boundaries when appropriate initial models are available. Speech segmentations may be discrepant if spectrum models and F0 models are embedded and trained separately.
Hence context dependent HMMs are trained with feature vector which consists of spectrum, F0 and their dynamic features as in Fig.5.2. Which results in HMM which has four streams as shown in Fig.5.3.

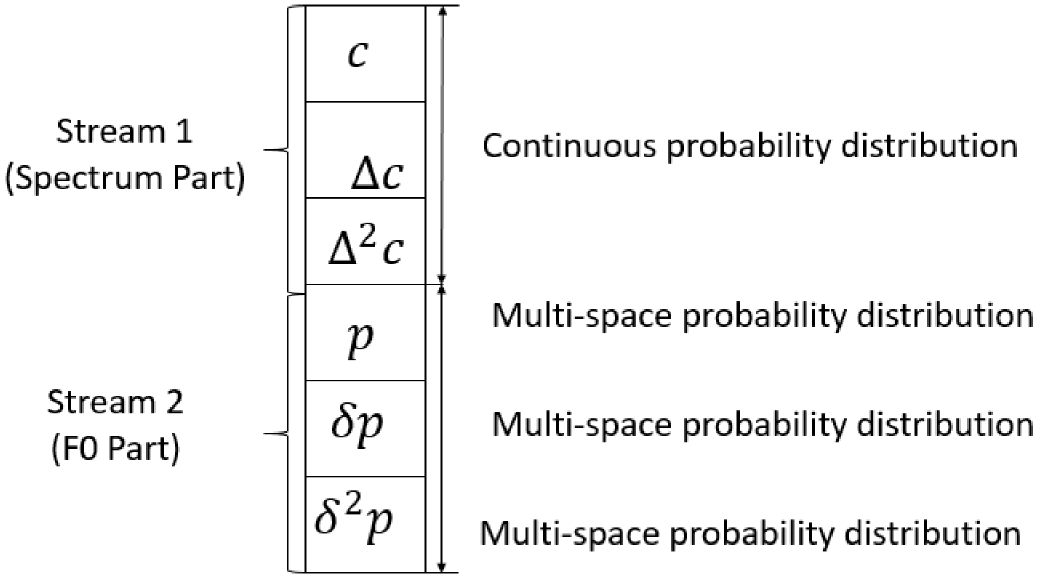
Stream 1
(Spectrum Part)
Stream 2
(F0 Part)
c
Δc
$\Delta^2 c$
p
δp
$\delta^2 p$
Continuous probability distribution
Multi-space probability distribution
Multi-space probability distribution
Multi-space probability distribution

Figure 5.1: Output feature vector.

5.3 Modeling of Durations

5.3.1 Overview

There exists many proposed techniques for training HMMs and their state duration densities simultaneously (e.g.,[28]). But these technique require large computations. so, state duration densities are estimated by using state occupancy probabilities which are obtained in the last iteration of embedded re-estimation as in [29].

In the HMM-based speech synthesis system described above, state duration densities were modeled by single Gaussian distributions estimated from histograms of state durations which were obtained by the Viterbi segmentation of training data[33]. In this procedure, however, it is impossible to

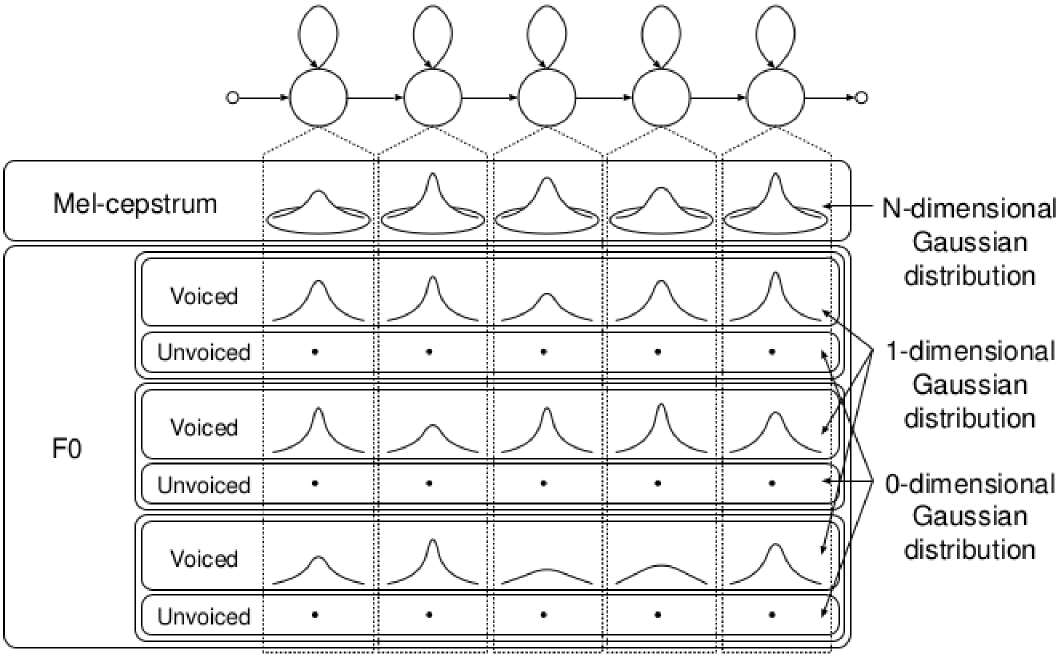
Mel-cepstrum
F0
Voiced
Unvoiced
Voiced
Unvoiced
Voiced
Unvoiced
N-dimensional Gaussian distribution
1-dimensional Gaussian distribution
0-dimensional Gaussian distribution

Figure 5.2: Structure of HMM.

obtain variances of distributions for phonemes which appear only once in the training data.

To overcome it Gaussian distributions of state durations are calculated on the trellis(Section 3.1.2) which is made in the embedded training stage. State durations of each phoneme HMM are regarded as a multi-dimensional observation, and the set of state durations of each phoneme HMM is modeled by a multi-dimensional Gaussian distribution. Dimension of state duration densities is equal to number of state of HMMs, and nth dimension of state duration densities is corresponding to nth state of HMMs.

5.3.2 State duration models training

From the trellis structure which are obtained in embedded training state duration densities are estimated. The mean $\xi(i)$ and the variance $\sigma^2(i)$ of duration density of state i are determined by

$$\xi(i) = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)(t_1-t_0+1)}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)} \tag{5.5}$$

$$\sigma^2(i) = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)(t_1-t_0+1)^2}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)} - \xi^2(i) \tag{5.6}$$

respectively, where $\chi_{t_0,t_1}(i)$ is the probability of occupying state i from time t_0 to t_1 and can be written as

$$\chi_{t_0,t_1} = (1-\gamma_{t_0-1}(i)) \cdot \prod_{t=t_0}^{t_1}\gamma_t(i) \cdot (1-\gamma_{t_1+1}(i)) \tag{5.7}$$

where $\gamma_t(i)$ is the occupation probability of state i at time t, and we define $\gamma_{-1}(i) = \gamma_{T+1}(i) = 0$

5.4 Context dependent model

5.4.1 Contextual factors

In this we consider the relation between phonemes. Context is nothing but the factor of speech variations. There are many contextual factors which affect spectrum, F0 and duration. In this report following contextual factors are taken into account:

Phoneme
- {preceding, succeeding} two phonemes
- current phoneme

Syllable
- no. of phonemes at {preceding, current, succeeding} syllable
- {accent, stress} of preceding, current, succeeding syllable
- Position of current syllable in current word
- no. of {preceding, succeeding} {accented, stressed} syllable in current phrase
- no. of syllables {from previous, to next} {accented, stressed} syllable
- Vowel within current syllable

Word
- Part of speech of {preceding, current, succeeding} word
- no. of syllables in {preceding, current, succeeding} word
- Position of current word in current phrase
- no. of {preceding, succeeding} content words in current phrase
- no. of words {from previous, to next} content word

Phrase
- no. of of syllables in {preceding, current, succeeding} phrase

Note that a context dependent HMM corresponds to a phoneme.

5.4.2 Context clustering

We build context dependent models considering many possible combinations of the above mentioned contextual factors, and we expect that we able to obtain appropriate models. But, as contextual factors increase,then their possible combinations also increase exponentially. Therefore, model parameters with sufficient accuracy cannot estimated with the given limited training data. And it is very difficult to prepare speech database with all these possible combinations of contextual factors

Introduction to context clustering

So as to overcome the above problem, we apply a decision-tree based context clustering technique as in [32] for distributions for spectrum, F0 and state duration.

The decision-tree based context clustering algorithm have been extended for MSD-HMMs in [31].Because each of spectrum, F0 and duration have its own influencing contextual factors, the distributions for spectral parameter and F0 parameter and the state duration are clustered independently

5.4.3 Databese Desctiption

We used phonetically balanced 1000 sentences from Telugu speech database for training. Given speech database were sampled at 16 kHz sampling frequency and a Blackman window of 25-ms id applied with a windowshift of 5 ms. Then mel-generalized cepstral coefficients were calculated by the mel-cepstral analysis which is presented in chapter 2. Output feature vector consists of spectral(mel-generalized spectral coeffecients) and excitation(F0 parameter) vectors. Spectral parameter vector consists of 25 mel-generalized cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. F0 parameter vector consists of log F0, its delta and delta-delta. We used 3-state left-to-right HMM models with no skip and with single diagonal Gaussian output distributions. Decision trees for spectrum, F0 and duration models were constructed.

Chapter 6

HMM-based Text-to-Speech Synthesis

Speech waveform was generated by using an speech parameter generation algorithm from HMM and a source-filter based vocoding technique. By listening tests, I confirmed that the proposed system successfully synthesizes natural-sounding speech which resembles the speaker in the training databas

6.1 Overview

HMM-based text-to-speech synthesis systems synthesis part is shown in Fig. 6.1. In the synthesis, we convert the given text as context based label sequence. Then,from these context dependent label sequence, a sentence HMM is constructed by concatenating context dependent HMMs. and the state durations of the sentence HMM are determined as in [31]. By considering the obtained state durations, a sequence of mel-cepstral coefficients and F0 values are generated by using speech parameter generation algorithm. Which also includes voiced and unvoiced decisions . Finally, speech is synthesized directly from the generated mel-cepstral coefficients and F0 values by using the MLSA filter [16], [19].

6.2 Text analysis

Phonetic transcription is done for the given text input which is in Telugu. Now for the trancscripted text we form a context dependent label sequence.

6.3 Duration determination

For a given speech length T , the goal is to obtain a state sequence $q = q_1, q_2, \cdots, q_T$ which maximize

$$\log P(q|\lambda, T) = \sum_{k=1}^{K} \log p_k(d_k) \tag{6.1}$$

under the constraint

$$T = \sum_{k=1}^{K} d_k \tag{6.2}$$

where $p_k(d_k)$ is the probability of duration d_k in state k, and K is the number of states in HMM λ.

Since each duration density $p_k(d_k)$ is modeled by a single Gaussian distribution, state durations $\{d_k\}_{k=1}^{K}$ which maximize (6.1) are given by

$$d_k = \xi(k) + \rho \cdot \sigma^2(k) \tag{6.3}$$

$$\rho = \frac{T - \sum_{k=1}^{K} \xi(k)}{\sum_{k=1}^{K} \sigma^2(k)} \tag{6.4}$$

where $\xi(k)$ and $\sigma^2(k)$ are the mean and variance of the duration density of state k, respectively.

6.4 Results

According to the estimated state duration, spectral and excitation parameters are generated from a sentence HMM constructed by concatenating context dependent HMMs. Fig. 6.2 shows spectra of natural speech and synthesized speech for a Telugu phrase ”aayanabhaarya”. Fig. 6.4 shows F0 pattern of natural speech and synthesized speech for a Telugu sentence ”aayanabhaarya”.

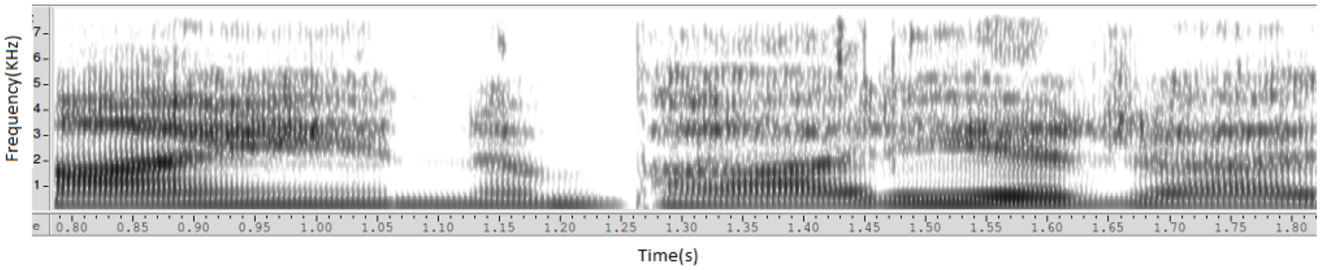

Figure 6.1: Spectrogram of the original phrase ”aayanabhaarya”

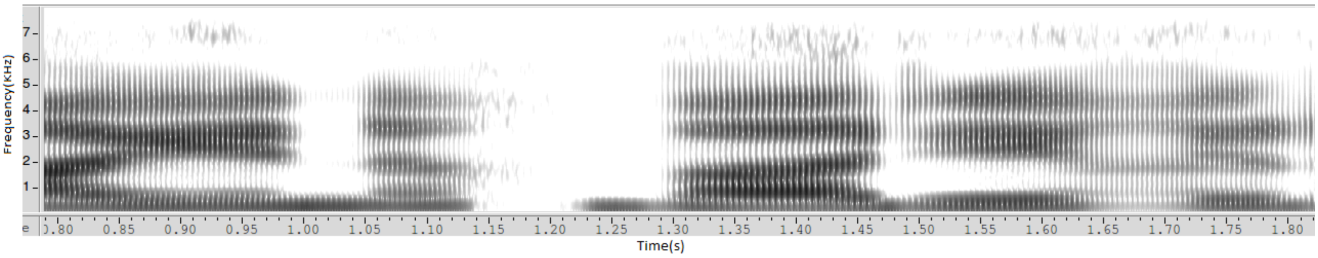

Figure 6.2: Spectrogram of the synthesized phrase ”aayanabhaarya”

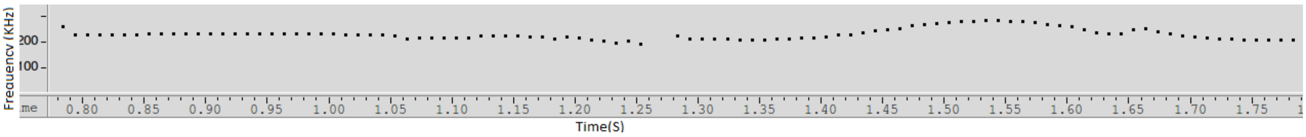

Figure 6.3: Pitch contour for the original phrase ”aayanabhaarya”

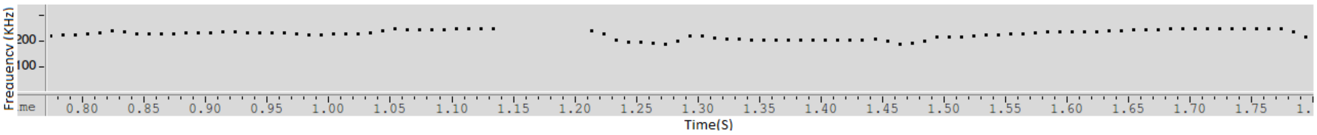

Figure 6.4: Pitch contour for the synthesized phrase ”aayanabhaarya”

Chapter 7

Conclusions

This thesis deals with HMM-Based Text to speech synthesis system for Telugu language.A working Telugu HTS is built. Source-filter model is discussed. How to model the speech parameters based on HMM is presented in this thesis. Spectral Parameter modelling and excitation (F0) parameter modelling is dealt in this thesis. Generation of speech parameter vector from the HMM and construction of HMM Based Text-to-Speech Synthesis system for Telugu is presented. Effects of dynamic features are also discussed.

References

[1] A. Ljolje, J. Hirschberg and J. P. H. van Santen, Automatic speech segmenta-tion for concatenative inventory selection, *Progress in Speech Synthesis,ed. J.P.H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, Springer-Verlag, New York,* 1997

[2] R. E. Donovan and P. C. Woodland, Automatic speech synthesiser parameter estimation using HMMs, *Proc. of ICASSP,* pp 640–643 , 1995

[3] R. E. Donovan and P. C. Woodland, Improvements in an HMM-Based Synthesizer, *Proc. of EUROSPEECH,* pp 573–576 , 1995

[4] H. Hon, A.Acero, X. Huang, J. Liu and M. Plumpe, Automatic generation of synthesis units for trainable text-to-speech synthesis, *Proc. of ICASSP,* pp 293–306, 1998

[5] H. Hon, A.Acero, X. Huang, J. Liu, S. Meredith and M. Plumpe, Recent improvements on Microsofts trainable text-to-speech system-Whistler, *Proc. of ICASSP,* pp 959–962, 1997

[6] R. E. Donovan and E. M. Eide, The IBM Trainable Speech Synthesis System, *Proc. of ICSLP,*, vol. 5, pp 1703–1706 , Nov. 1997

[7] A. Falaschi, M. Giustiniani and M. Verola, A hidden Markov model approach to speech synthesis, *Proc. of EUROSPEECH,* pp 187–190, 1997

[8] M. Giustiniani and P. Pierucci, Phonetic ergodic HMM for speech synthesis, *Proc. of EUROSPEECH,*, pp 349–352 , 1991

[9] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, Speech synthesis from HMMs using dynamic features, *Proc. of ICASSP,* pp 389–392, 1996

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis, *Proc. of EUROSPEECH, S11.PO2.16,*, vol. 5, pp 2347–2350, 1999

[11] K.Tokuda, T. Kobayashi and S. Imai, peech parameter generation from HMM using dynamic features, *Proc. of ICASSP,* pp 660–663, 1995

[12] K.Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features, *Proc. of EUROSPEECH,* pp 757–760, 1995

[13] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, Voice characteristics conversion for HMM-based speech synthesis system, *Proc. of ICASSP,* vol. 3 , pp 1611–1614, 1997

[14] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR, *Proc. of The Third ESCA/COCOSDA workshop on Speech Synthesis,* pp 273–276, Dec. 1998

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, Speaker Interpolation in HMM-Based Speech Synthesis System, *Proc. of EUROSPEECH,, Th4C.5,*, vol. 5, pp 2523–2526, Sep. 1997

[16] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, An adaptive algorithm for mel-cepstral analysis of speech, *Proc. of ICASSP,* vol. 1 , pp 1137–1140, 1992

[17] A.V. Oppenheim and R.W. Schafer, Discrete-time signal processing, 2nd edition. Prentice-Hall, Englewood Cliffs, N.J., 1989

[18] S. Imai and C. Furuichi, Unbiased estimator of log spectrum and its application to speech signal processing, *Proc. of EURASIP,* pp 203–206, Sep. 1988

[19] S. Imai, Cepstral analysis synthesis on the mel frequency scale, *Proc. of ICASSP,* pp 93–96, Feb. 1983

[20] T. Kobayashi and S. Imai, Complex Chebyshev approximation for IIR digital filters using an iterative WLS technique, *Proc. of ICASSP,* pp 377–380, Apr. 1990

[21] L.A.Liporace, Maximum Likelihood Estimation for Multivariate Observations of Markov Sources, *IEEE Trans. Information Theory, IT-28 729734*, 1982

[22] B.H. Juang, Maximum-likelihood estimation for mixture multivariate stachastic observations of Markov chains, *ATT Technical Journal, 64, no. 6, 12351249*, 247–260, Aug. 1994

[23] U. Jensen, R. K. Moore, P. Dalsgaard and B Lindberg, Modeling intonation contours at the phrase level using continuous density hidden Markov models, *Computer Speech and Language,* vol. 8 , pp 247–260, Aug. 1994

[24] G. J. Freij and F. Fallside, Lexical stress recognition using hidden Markov models, *Proc. of ICASSP,* pp 135–138, 1988

[25] K. Ross and M. Ostendorf, A dynamical system model for generating F0 for synthesis, *Proc. of ESCA/IEEE Workshop on Speech Synthesis,* pp 131–134, 1994

[26] M. Nishimura and K. Toshioka, HMM-based speech recognition using multi-dimensional multi-labeling, *Proc. of ICASSP,* pp 1163–1166, 1987

[27] K. Koishida, K. Tokuda, T. Masuko and T. Kobayashi, Vector quantization of speech spectral parameters using statistics of dynamic features, *Proc. of ICSP,* pp 247–252, 1997

[28] S. E. Levinson, Continuously Variable Duration Hidden Markov Models for Speech Analysis, *Proc. of ICASSP*, 1241–1244, 1986

[29] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, Duration Modeling in HMM-based Speech Synthesis System, *Proc. of ICSLP,* 2 pp 29–32, Nov. 1998

[30] J. J. Odell, The Use of Context in Large Vocabulary Speech Recognition, *Ph.D. thesis,Cambridge University*, 1995

[31] N. Miyazaki, K. Tokuda, T. Masuko and T. Kobayashi, A Study on Pitch Pattern Generation using HMMs Based on Multi-space Probability Distributions, *Technical Report of IEICE,* SP98-12, Apr. 1998

[32] K. Shinoda and T. Watanabe, Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle, *Proc. of ICASSP,* pp 717–720, May. 1996

[33] T. Yoshimura, Simultaneous modelling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-to-speech systems, Ph.D. thesis, Nagoya Institute of Technology, Japan, 2002.