Chapter 5

Duration Modeling - Part II

Duration models predict the duration of individual segments to be used for synthesizing speech, by considering the various factors affecting the duration. While estimating the effects of the various factors, speech is considered to be the resultant of a series of separate articulatory actions or gestures\(^1\) occurring one after another. Even though the perception and representation of speech segments like phonemes, syllables or words are discrete, speech produced is manifested as a continuous variation of acoustic parameters. The neuro-muscular commands issued from the brain for speech production are discrete, while the resulting articulations are continuous and hence influenced by preceding and succeeding articulations. The articulatory gestures corresponding to the phonetic segments overlap, leading to the effects like assimilations, neutralizations and nasalization [144]. The overlap of gestures occur, since in continuous speech the articulation of a phoneme begins before the completion of articulation of the preceding one. The anticipation of

\(^{1}\)The gestures include tightening and slackening of vocal chords, generating sufficient subglottal pressure, producing sufficient degree of mouth opening, raising or lowering of the velum, advancing the tongue root, lowering the jaw, lowering the larynx etc.
succeeding phonemes and the influence of preceding phonemes changes the acoustic realization of phonemes\(^2\).

The merging of articulatory actions and the resulting mixing up of acoustic features is one of the major challenges in artificial speech production and recognition. To highlight the difficulty of the problem, Hockett's analogy of conveyor belt of coloured eggs (phonemes), which are smashed and smeared by rollers into mess of yolk, albumen and shell fragments (phonetic realizations) is often cited. Once the eggs are smashed it is no longer a series of discrete entities, but a continuous flow of a mixture, having the features of all nearby eggs. The challenge in recognition is to identify the colors of the eggs in correct sequence from the mess. For synthesized speech to appear natural, this mess has to be simulated. Otherwise the generated speech will be robotic. The articulatory gestures follow one after the other in a linear fashion, only when we pronounce a word phoneme by phoneme, (coloured eggs passing one after the other, without being smashed) which rarely occurs in continuous speech. Hence it is now considered that this beads-on-a-string representation is not adequate for speech modeling, production or recognition [152].

This is especially true when it comes to adding rhythm or prosody to speech. Each sentence, phrase and word we utter has a rhythm as a whole. This cannot be recreated by considering the constituent speech segments in isolation and manipulating the prosodic features (duration, \(f_0\), intonation) of these individual units. The sentences, phrases and words have to be considered as entities and the duration patterns for these entities have to be

\(^2\)For example the acoustic manifestation of /t/ in the words ‘two’ and ‘tea’ are influenced by the phonemes following /t/
created considering them in a wholistic manner. Natural sounding speech can be obtained by mimicking the way in which human speech production system introduces prosody. The duration model presented in this chapter is an attempt in this direction and is inspired by the basic aspects of three theories i) exemplar theory, ii) memory prediction framework and iii) theory of analogy which explain the human cognition, brain structure and language acquisition respectively. Duration modeling has not been previously done in this approach. The practical feasibility of the model named as memory based model is underlined by Zipf’s law, which deals with the nonuniform distribution of words in any natural language.

5.1 Exemplar Theory

Exemplar theory is a psychological model of similarity and classification [146]. It was first introduced as a model of perception and categorization. According to exemplar theory, all linguistic materials are stored in memory in the form of exemplars or patterns, after it has been either produced or perceived [145]. During language acquisition, memory is incrementally filled with exemplars. The exemplars are categorized into clouds of memory traces, with similar traces lying close to each other and dissimilar traces lying more distant in a cognitive map in our memory. Each category is represented by a large cloud of remembered tokens [146]. Category is a mental construct, which relates two levels of representation, a discrete level and a parametric level [147]. The discrete level refers to the labels of the category and parametric level to the parameters stored in memory, corresponding to the linguistic construct. The exemplars associates with each category of the system, a
cloud of detailed perceptual memories.

When a new input stimuli arrives, it is compared with the existing exemplar clouds and is classified according to its similarity with the stored exemplars. The input stimuli is perceptually encoded into parameters. The similarity of the new token to any stored exemplar can be computed as its distance from the exemplar in the parametric space [146]. If it is not similar to any of the existing exemplar clouds, a new category (exemplar cloud) will be generated. The detailed phonetic knowledge that native speakers have about their language can be understood, in terms of the acquisition of a large number of memory traces of experiences [146]. Speakers access the stored representations during speech production and perception.

Speech production corresponding to a linguistic token starts with the computation of a plan by the brain, regarding the sequence of neuromuscular commands to be issued to the vocal organs. The plan comprising of a sequence of articulatory gestures is computed so as to achieve a perceptual target. When we try to reproduce a speech segment we have already perceived, a auditory feedback is enacted, so that the articulations are adjusted to achieve a perceptual target that is stored in the brain. During speech production, we always try to reduce the difference between what we have perceived before and stored in memory (perceptual target) and what we perceive at the time of production. The exemplars stored in memory serve as perceptual targets. The phonetic details are stored in a parameter space including acoustic correlates of prosodic features like $f_0$, amplitude and duration. The speakers access the stored representation of the prosodic events (tonal and temporal structure) and this serve as a reference in speech production [155].
The learners of a language acquire internal prosodic model by storing in memory representations of large numbers of perceived acoustic realizations. Since the prosody for each language and dialect vary with different types of speaking style including reading style, socially driven styles, situation specific style and emotional style, the speakers will be internalizing prosodic models corresponding to each of these [155]. The prosody also varies with communicative and situational settings. When the situation demands a fast or loud speech, the speaker may access a prefabricated model that implements the appropriate strategies like reducing the duration and increasing intensity etc. The evidence of exemplar model in speech perception and production has been investigated in the following works.

By comparing the syllable duration, Antje Schweitzer and Bernd Mobius [148] has suggested that infrequent syllables are assembled from smaller units during speech production, where as frequent syllables are accessed as units, from stored exemplars. The infrequent syllables have to be assembled from smaller units, because they are not represented by enough exemplars, to establish a syllable level target region. If exemplar cloud exists, then the speakers are likely to access the durations of the stored exemplars and use them as reference in speech production. This hypothesis was supported by the simulation results using an exemplar theoretic computational model of syllable frequency effects by Michael Walsh et al [149].

Uta Bene et al [150] studied the difference in coarticulation in high frequency and low frequency syllables, and have found that there is a tendency towards stronger co-articulation, with greater co-articulatory variability in high frequency syllables than in low frequency syllables. For most frequent syllables, the stored motor program corresponding to one of the many exem-
The concept of storing prosodic features of most frequently occurring speech constructs as exemplars and retrieving them for speech production can be utilized for developing a novel duration model. Such a model stores the duration patterns of commonly used sentences, phrases, words, morphemes, syllables and phonemes as exemplar clouds and predicts duration of speech segments of the given text by retrieving the most similar exemplar. The duration patterns of a word, that is not there in the data base, is taken in comparison with that of a similar word. This is done by a process of analogical reasoning, by which an exemplar based model, when presented with a novel situation, finds the most resembling exemplar and imparts its features on the new input.

5.2 Analogical modeling

The term analogy is derived from the Greek word analogy which means regularity. It has been studied and widely used in mathematics, philosophy and linguistics. Analogy is process of generalization from one set of expressions to another [153]. The generalization by analogy can be defined as the inferential process, by which an unfamiliar object is seen as an analogue of known objects of the same type, so that what ever properties are known about the latter are assumed to be transferable to the former. The concepts of comparison, correspondence and resemblance are closely related to analogy [153].

Linguists observe analogy as a creative mechanism in language [154].
It plays a major role in the development of languages. The theory has been widely used in various fields of linguistics like syntax, morphology and phonology. Its impact can be seen in the process of language acquisition. A child who acquires the knowledge that the plural of ‘pen’ is ‘pens’, will extend that knowledge analogically to derive the plural of ‘pencil’.

The theory of analogy can also be extended to prosody generation. The rhythm of a new utterance will be generated analogous to a similar word, whose prosodic pattern is already there in memory. For example the prosody of ‘deflection’ will be generated analogous to that of ‘reflection’. This concept can be used in memory based duration model, to derive the duration patterns of exemplars, which are not stored in memory.

5.3 Memory Prediction Framework

Memory prediction framework (MPF) is a theory of brain function proposed by Jeff Hawkins in his book On Intelligence [107]. It is based on the hierarchical function of the human neocortex. The neocortex stores sequence of patterns and is capable of memorizing frequently observed sequences of patterns. The brain finds solution to any problem, by retrieving patterns stored in memory. According to this theory, the entire cortex is a memory system. The inputs to our brain are patterns. The brain process these patterns by classifying them and storing them in the different layers of neocortex.

The neocortex has a hierarchical network structure, with all regions of the hierarchy performing the same basic operations. The lowest level in the hierarchy receives inputs from our senses and represents it by spatial and temporal patterns. Each hierarchy level remembers the frequently ob-
served temporal sequences and generates labels or invariant names for these sequences, which is propagated up the hierarchy. The invariant name is a form of abstraction, which means that only the essence is stored and not the details. For example, suppose that we see the photographs of a rose flower taken in different angles and lighting conditions. Even though the retinal signals that reach the lowest level of hierarchy is different, the higher levels forms the invariant name, that is ‘flower’. A conventional pattern recognition program will not be able to perform this. It will not be able to recognize if the picture is rotated or shifted laterally, mainly because of the absence of abstraction. Similarly the different instantiations of a spoken word whose acoustic parameters may be quiet different in terms of amplitude, pitch and/or duration will be perceived by the higher levels of neocortex, by forming invariant representation of it.

The higher levels take these patterns of labels and memorize higher order sequences. Each level up the hierarchy does the same function, of receiving labels from the lower levels and forming patterns of the labels. Thus each layer stores higher and higher structures. In case of vision, bottom up information starts at the lower level from retinal signals. At the higher levels it may be lines, regions, specific objects like faces, mountains, landscapes and their behaviour.

The neocortex can recall the stored patterns autoassociatively, which means that the complete pattern can be retrieved by partial or distorted input. When an input pattern arrives, the corresponding stored pattern is retrieved by looking into the similarity between the two. If there is a match at any level, the sequence of corresponding labels or names are propagated up.
When there is mismatch between the input and the memorized sequence at any level, a top down signal is propagated to the lower levels and a more complete representation is propagated upwards, causing alternate sequences to be activated at higher levels. When we read any text in a familiar language, we do not actually read all the individual letters, but instead recognize entire words and often entire phrases at a glance. We recognize words and phrases as entities. We start perceiving the details or individual letters only when we stumble upon an unfamiliar word, which is not already stored in our vocabulary. The mismatch of the unfamiliar word with stored patterns result in a more detailed representation being propagated up the hierarchy.

Memory prediction framework can be used to model speech perception, production and recognition. Memory prediction framework or model is expected to provide a foundation for a comprehensive new theory and model for automatic and human processing of speech [152]. Since the theory is novel, a few theoretical or practical studies of the memory-prediction framework exist today. The application of memory prediction framework for visual pattern recognition has been reported in the works of Dileep George et al [116] and Saulius J. Garalevicius [109].

The memory prediction framework can be used for duration modelling, where the duration patterns of sentences, phrases, words, morphemes and syllables can be stored in a hierarchical manner and retrieved at the time of synthesis.
5.4 Zipf’s law

The distribution of words in a corpus of any natural language is non uniform. A few occur very often while many others occur rarely [159]. The rank of a word (in terms of its frequency) in a given corpus of natural language utterances is approximately inversely proportional to its actual frequency, and so produces a hyperbolic distribution according to Zipf’s law. It is an empirical law named after the linguist George Kinsley Zipf. If the words of a sample text are ordered by decreasing frequency, the frequency of the kth word \( P(k) \), is given by \( P(k) \propto k^{-\alpha} \), with \( \alpha \approx 1 \) [156]. The word with rank 2 (second most common word) will occur half as often as the word with rank one and so on. One of the significant implications of this hyperbolic relation between rank and frequency is the fact that, most of the vocabulary in a given text can be represented using a finite number of words. The half of the Brown corpus with over 1 million words can be covered with 135 words [157]. Zipf has also observed that, not only words but phrases and whole sentences also follow Zipf’s law [158].

The significance of Zipf’s law is that, a major portion of any speech corpus can be represented by the most frequently occurring words. Hence for duration modeling, the duration patterns of most frequently occurring words and phrases need be stored and made use of.

5.5 Memory based duration modelling

The language acquisition by infants starts by perceiving the speech utterances of people around him or her. The global rhythmic and prosodic aspects of the
utterances are identified first before the child focuses on small speech units [15]. The prosody model of the first language is internalized by the child by storing the rhythm of a large number of utterances. Later other speaking styles are learned by storing the exemplars of those styles. The prosody model of a TTS system can be designed similar to the internal prosodic model we have in our brain.

It can be observed that, when we produce speech corresponding to frequently used phrases like ‘how do you do’, ‘all the best’, ‘many many happy returns of the day’ etc., the overall rhythm we impose on it, is recollected from the stored exemplars of the corresponding representation in the memory. These exemplars have been filled in to memory, when we perceived and produced these phrases many times during the course of language acquisition. Similarly the prosodic features of frequently occurring morphemes, words and syllables are stored and will be used for speech production. Infrequent segments will be assigned the durational patterns of the most similar exemplar stored. For example if the word ‘flight’ is encountered for the first time, it will be pronounced with the rhythm of previously produced words like ‘slight’, ‘might’ or ‘right’. If a similar exemplar is not found, then the speech segment will be assembled from its constituents and the duration of the corresponding smaller units will be assigned to it. These concepts are in fact explained by the theories of i) memory prediction framework ii) exemplar theory and iii) theory of analogy.

The memory based duration model developed based on these concepts, attempts to mimic the way in which human brain acquires prosodic features and recreates it during speech production. The central concepts behind the proposed model are the presumptions that i) the duration patterns of speech
segments are stored in memory as exemplars and retrieved for speech production ii) for new inputs whose exemplars are not stored, the duration patterns of the most similar exemplar is retrieved and iii) the memory has a hierarchical structure with each layer storing exemplars of different constituents of speech. This possibility is empirically evident from Zipf’s law.

Before proceeding to the basic concepts of the memory based model, i) the duration patterns of speech constructs in natural speech is investigated and ii) Zipf’s law is verified for Malayalam corpus. The duration patterns in natural speech are analyzed by plotting the patterns of phrases, words, morphemes and rhyming words. Since we are concerned with durational variation, it is better to use z-score, which gives the deviation of duration from its mean value (normalized using its standard deviation). The normalized duration or z-score [104] is obtained from the phonetic duration ‘d’ as

$$z - \text{score} = \frac{d - \mu}{\sigma}$$  \hspace{1cm} (5.1)

The duration patterns of phrases, words, morphemes and rhyming words are analyzed, by plotting the z-scores of phonemes to observe the similarity of patterns.

### 5.5.1 Duration patterns of words and morphemes

To analyze the duration patterns of words, different instantiations of the words occurring in different contexts, were chosen. It was observed that, the durations of individual phonemes may differ in values, but the their z-scores when the word as a whole is looked into, show similarity. The figure 5.1 is the plot of the z-scores of phonemes in the word /paRanju/. The z-scores of ten
instantiations of the word was taken. The plot shows that, the occurrences of the word have produced similar patterns. The difference is mainly in the boundary, which may be due to the influence of the neighboring speech segments (context in which they occur).

The z-scores of the duration values of the speech segment corresponding to the linguistic unit /sambandhich/ in words /sambandhicha/, /sambandhichatu/, and /sambandhichu/ is shown in Figure 5.2. Figure 5.3 show the duration patterns of the sub word /keralathi/ in the words /keralathile/, /keralathinte/ and /keralathil/. The duration patterns of the morpheme /mUnnaR/ in the words /mUnnaRil/, /mUnnaRile/, /mUnnaRileKu/ and /mUnnaR/ are given in Figure 5.4. The duration patterns depicted in the figures show that, even though the duration values of each phoneme vary, each linguistic unit have a particular pattern. As in the case of words, the values differ generally near the boundaries which is due to the effect of the neighboring speech segments. All the instantiations show much similarity in the middle phonemes.

5.5.2 Duration patterns of rhyming words

Different sets of rhyming words were taken to study the similarity of duration patterns of rhyming words. The examples are shown in Figure 5.5, Figure 5.6, Figure 5.7, Figure 5.8 and Figure 5.9. In these plots, the duration values are taken, instead of z-scores. Figure 5.5 shows the patterns of the words /peTi/, /paTi/, /vaTi/, /kATi/ and /kuTi/. In all words, the last syllable is the same (/Ti/). All the words have two vowels, the first vowel is different (/e/, /a/, /A/, /u/) and second the same (/i/). It can be seen
Figure 5.1: Duration pattern of the word /paRanju/

Figure 5.2: Duration pattern of the word /sambandhichu/
Figure 5.3: Duration pattern of the segment /keralathi/

Figure 5.4: Duration pattern of the segment /munnar/
that, the patterns are similar. The difference is mainly in the first syllable, especially in the first vowel, which is different in different words. It means that the duration patterns of a word can be taken from that of a rhyming word, by changing the duration of the phonemes in which they differ.

The duration patterns of /pachati/, /kichati/ and /achati/ are shown in figure 5.6. The words vary only in the first syllable. The duration patterns are similar except at the boundaries. The Figure 5.7 depicts the duration patterns of six rhyming words /mekham/, /medam/, /melam/, /vedam/, /vesham/ and /venam/. The words have same vowels, while consonants differ. For the first three vowels, the first consonant is /m/ and for the rest it is /v/. The duration pattern of any of these can be obtained from that of others, by replacing the duration pattern of the phoneme in which they differ. (For example duration pattern of /medam/ can be obtained from that of /melam/ by replacing the duration of /l/ with that of /d/.

The Figure 5.8 shows the duration patterns of seven rhyming words. The words differ in terms of the constituent consonants, the vowels being the same. The patterns appear to be more similar, when the vowels are the same (when compared with Figure 5.5 where vowels are different). The plot in black colour, is the mean profile of the patterns. This means that the duration patterns of a cluster of similar words can be represented by the mean profile. The Figure 5.9 shows the duration patterns of five compound words (/apamanam/, /kurimanam/, /abhimanam/, /varumanam/ and /anumanam/) having two morphemes each. The first morpheme is different, while the second is the same (/manam/). It can be seen that, the duration patterns of the second morpheme (last five phonemes) in all words show similarity. The mean profile of the second morpheme /manam/ can be stored and for
Figure 5.5: Duration patterns of 5 rhyming words with last syllable as /Ti/
synthesizing any word in that set the second morpheme can be taken from
the mean profile.

5.5.3 Duration patterns of phrases

The duration patterns of similar phrases also showed similarity in patterns.
The examples are shown in Figure 5.10 and Figure 5.11. The Figure 5.10
shows two instantiations of the phrase /charcha natathi/ and Figure 5.11
gives three instantiations of the phrase /manthri paranju/. As in the case of
words, the phrases also show similarity in duration patterns. In the case of
words, it is observed that the duration patterns of similar words differed at
the boundaries due to influence of neighbouring words. Hence the synthesis
of most frequent phrases can be made natural sounding by storing their
duration pattern and using it. The duration patterns of the phrases will
Figure 5.6: Duration patterns of 3 rhyming words differing in the first syllable only

Figure 5.7: Duration patterns of 5 rhyming words with last syllable as /am/
Figure 5.8: Duration patterns of 7 rhyming words and the mean profile

Figure 5.9: Duration patterns of 5 rhyming words with second morpheme as /manam/
5.5.4 Clustering of duration patterns

The analysis of duration patterns of words in natural speech has shown that, sets of similar words have similar duration patterns. In other words the duration patterns of words can be grouped into different clusters. The evidence of the clusters in duration patterns is investigated using k-means clustering. Cluster analysis classifies a set of observations into two or more mutually exclusive groups in such a way that, the degree of association between two objects is maximal, if they belong to the same group and minimal otherwise [160].

In k-means clustering, the data is partitioned into k mutually exclusive clusters. The k-means clustering function, returns a vector of indices indicat-
to which of the k clusters, it has assigned each observation. A silhouette plot can be drawn using the cluster indices returned by the k-means function, to get an idea of how well separated are the clusters. The silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters. The measure ranges from +1 (indicating data that are distant from neighboring clusters), through 0 (data that are not distinctly in one cluster or another) to -1 (data that are probably assigned to the wrong cluster). The higher positive values indicate more similarity and negative values indicate dissimilarity or wrong classification.

Figure 5.12 depicts the clusters of duration patterns of words with first syllable as ‘na’. It is seen that all the words can be clustered into three groups. In Figure 5.13, the clusters of words with first and second syllable as /h/ is shown (all other syllables are different eg. ‘valathu’, ‘kadalil’, ‘paranju’,

Figure 5.11: Duration pattern of phrase /mantri paranju/
Figure 5.12: Silhouette plot of clusters with /na/ as first syllable etc.). It is seen that there are only two clusters, which means that when two vowels are coming in a sequence the duration patterns show similarity. The influence of vowels on duration patterns is further investigated by clustering the data values of words with same vowel in three consecutive syllables. It can be seen from Figure 5.14 that, almost all the values comes under one cluster. Only three values are there in the second cluster, of which one is wrongly clustered. This means that it is the sequence of vowels that primarily determines the duration patterns.

5.5.5 Verification of Zipf’s law for Malayalam corpus

To verify Zipf’s law a corpus of 26865 words were compiled from Malayalam newspaper reports. The frequency distribution of Malayalam words has been observed for this corpus. The words in the corpus were ranked according to
Figure 5.13: Silhouette plot of clusters with vowel /a/ in first and second syllable

Figure 5.14: Silhouette plot of clusters with vowel /a/ in first, second and third syllable
It can be seen that (Figure 5.15) the rank and frequency show a hyperbolic relation. It is observed that, the first 100 words (in the order of rank) cover 25.6% of the whole corpus. First 3006 words can represent 75% of the corpus (table 5.1). This means that natural sounding speech can be synthesized by using duration patterns of most frequently occurring speech constructs.

### 5.6 Memory based duration model

The analysis of duration patterns of natural speech has shown that the similar linguistic units have similar duration patterns. The linguistic units as a whole has a particular duration pattern, irrespective of the individual values. This shows that the concepts of exemplar theory, memory prediction model and theory of analogy can be utilized for generating natural sounding speech. Hence a new duration model is developed using the basic aspects of these theories. This is feasible since we can represent most of the corpus with frequently used phrases, words and syllables as suggested by Zipf’s law.

The inputs to the duration model are phonetic representations coming from the text analysis and phonology module. The phonetic representations
correspond to the actual form in which the phonemes are manifested in each particular context and are obtained after applying the phonological rules of each language. The phonetic representations include all the phonetic manifestations of phonemes\(^3\).

The duration model consists of a hierarchical framework with six layers (Figure 5.16), the highest layer corresponding to frequently used sentences and lowest layer the phonetic representations.

Similar to human speech acquisition, the model learns incrementally and is trained using recorded speech data of the required style (Dordarsan news). Initially all the layers in the model are empty. When trained with each sentence, the exemplar inventories of different classes of each layer are populated,

\(^3\)For example the phonetic representation of the phoneme /a/ includes all the allophones /a/
Figure 5.16: Memory Prediction Model for duration modeling

with the corresponding exemplars.

5.6.1 Storage of exemplars

The storage of exemplars start from the bottom layer and moves up, forming and attaching feature vectors to each exemplar.

- The phonetic representations of phonemes in each data along with the z-scores of the duration are fed to layer 1. The different classes in layer 1, stores the details of different phonetic representations (For the sentence 'How are you' ([haua:ju]), layer 1 stores the z-scores of [h], [au], [a:], [j] and [u]). Corresponding to each phonetic representation two vectors are there, a duration vector storing the z-scores and a
feature vector, storing the different factors (positional and contextual).
The factors are taken based on the statistical analysis of duration.
The factors include the ‘name’ or label of the phonetic representation, identity of the phonetic representation, identity of the preceding and succeeding phonemes, position of the phoneme in the word and word position (For example feature vector of [au] in the sentence [hauaju] is \( (p_5, au, h, a :, 2, 1) \)).

- The ‘names’ or labels of these representations \( p_i \) are passed on to layer 2. Layer 2 stores the feature vector corresponding to the syllables \( s_j \), which are constituted of phonemes in layer 1. The feature vector is comprised of factors including syllable identity or ‘name’ (say \( s_2, s_3, s_5 \) etc), identity of the constituent phonemes, number of phonemes in the syllable, identity of constituent phonemes, identity of the preceding and succeeding syllables, position of the syllable in the word and word position (For example feature vector of [haul is] \( s_2, p_2, p_5, 2, wb, h, s_3, 1, 1 \)).

- The labels of the syllables \( s_j \) are passed on to layer 3. The morphemes constituted by the syllables in layer 2, forms the exemplars in layer 3. The feature vector attached to each class of exemplar cloud is comprised of factors including morpheme identity, number of syllables in the morpheme, identity of constituent syllables, identity of the syllables preceding and succeeding the morpheme and position of the word in

\[
\begin{align*}
\text{Say } p_2, p_5, p_6, p_7 \text{ and } p_8, \text{ where } p_2 & \rightarrow [h], p_5 \rightarrow [au], p_6 \rightarrow [a :], p_7 \rightarrow [j] \text{ and } p_8 \rightarrow [u] \\
\text{Say } s_2 & \rightarrow [p_2, p_5], s_3 \rightarrow [p_6] \text{ and } s_5 \rightarrow [p_7, p_8] \\
\text{For syllables at the boundaries, the identity of the preceding syllable is denoted as 'W' if it is in the word beginning and succeeding syllable is denoted as 'we' if it is in the word end respectively}
\end{align*}
\]
the phrase.

- Layer 4, receives labels of the morphemes $m_i$ passed on from layer 3. The words are constituted of one or more morphemes. The feature vector attached to each word is comprised of factors including word identity, number of morphemes in the word, identity of constituent morphemes\(^7\), identity of the words preceding and succeeding the word and position of the word in the phrase.

- The labels of the syllables $w_k$ are passed on to layer 5. The phrases constituted by the words in layer 4, forms the exemplars in layer 5. The feature vector attached to each phrase is comprised of factors including phrase identity, number of words in the phrase, identity of constituent words\(^8\), identity of the syllables preceding and succeeding the phrase and position of the phrase in the sentence.

- The labels of the syllables $ph_{m_i}$ are passed on to layer 6. The sentences $s_n$, constituted by the phrases in layer 5, forms the exemplars in layer 6. The feature vector attached to each class of exemplar cloud is comprised of factors including sentence identity, number of phrases in the sentence, identity of constituent phrases, identity of the syllables preceding and succeeding the sentence and position of the sentence in the utterance.

The bottoms up signals propagate up in this fashion, until sentence level is reached. Top down signals are now initiated which tags to each of the lower level symbol, the labels of the corresponding higher level symbols (say $h \rightarrow [m_2, w_2, ph_1]$). This is required since the same word (say ‘are’) can occur

\(^7\)Say $w_2 \rightarrow [m_2], w_3 \rightarrow [m_3]$ and $w_4 \rightarrow [m_4]$ \\
\(^8\)Say $ph_1 \rightarrow [w_2, w_3, w_4]$
in different phrases and sentences. The different instantiations of syllables, morphemes, words and phrases can thus be accounted for. The feature vectors of the exemplars at each layer is hence augmented to include the labels of the higher layers. Each exemplar in a layer is thus attached with a feature vector, which contains factors denoting the exemplars to which it is affiliated in all the other layers, along with general factors that affect duration. For each sentence in the training database, at each level, the given feature vector is compared with the stored exemplars. If the symbols match, the new duration pattern will be stored as another exemplar. Otherwise a new class is formed.

5.6.2 Prediction of durational variation

When the phonetic representations corresponding to an input sentence are fed for prediction to layer 1, the matched labels at each layer are passed on to the higher layer. The labels of matched phonemes \((p_i)\) to layer 2, syllables \((s_j)\) to layer 3 and so on. If matched labels are found at all layers, the bottom up signals propagate, up to the sentence level. If a sentence level match is found, the stored duration pattern of z-score values, is retrieved for prediction. If match is not found at a level, the exemplars in other classes will be checked to find the nearest match at that level. Those symbols will be passed on to higher layers. The nearest match is found by using the distance measure given as follows,

The distance between the exemplars \(i\) and \(j\) is taken as

\[
d_{ij} = \sum_{m=1}^{M} w_m |x_{im} - x_{jm}|
\]

(5.2)
where $x_{im}$ and $x_{jm}$ denote the values of exemplars i and j on dimension m, respectively, $w_m$ is the weight given to dimension m and M denotes the number of dimensions along which the stimuli vary. The weight parameters are constrained to vary between 0 and 1 ($0 \leq w_m \leq 1$) and $\sum_{m=1}^{M} w_m = 1$.

To find the distance measure, factors in the feature vector are used, which constitute the different dimensions. The weights for the factors were determined empirically. The dimensions and their weights taken for the different speech constructs (sentence, phrase, word, morpheme and syllable) are enlisted below.

1. Position of the speech construct : 0.06
2. Number of constituent speech constructs (N) : 0.06
3. Identity of constituent speech constructs : 0.8 (weight factor of each constituent is 0.8/N)
4. Identity of preceding syllable : 0.04
5. Identity of succeeding syllable : 0.04

The nearest matching or minimum distance exemplar was found. If there were more than one exemplar having zero distance, the average of corresponding values were taken. In case of morphemes, if more than one stored morpheme show nearest match or minimum distance, the sequence of the vowels is looked into. The morpheme having the most similar vowel sequence will be chosen. The maximum threshold of the distance is set as 0.6, which indicates atleast 40% similarity. If similarity cannot be found within this threshold, then a top down signal is propagated to the lower layer and the
duration patterns of the constituent segments will be taken from the lower layer and a new label will be passed to top. Thus a series of bottom up and top down signal excursions will occur, finally resulting in the prediction of duration patterns, of the given sentence. A frequency count is kept for the new labels formed and if it exceeds a threshold value, it is stored as a new class. Thus the model learns incrementally even after training. The duration of the phonetic segments can be found from the z-scores.

The model was tested with a data base of 1048 words, consisting of 9208 phonemes. All the syllables in the test data base were there in the training data base. For those constructs (phrases, morphemes and words) whose match was found, the stored z-score values were retrieved. The z-score values of syllable durations were obtained from the syllable layer, for those constructs whose match was not found. Of the 9208 phonemes, match was found at phrase, word or morpheme level for 7206 phonemes (82.56%). For the rest (1606 phonemes), the z-scores of the nearest matching syllables were taken. The root mean square error obtained was 4.86 ms, correlation was 0.96 and mean opinion score was 3.82.

5.7 Summary of results

The conventional approach of considering speech as a linear succession of discrete speech segments, occurring one after another, cannot capture the complex durational patterns of natural speech, beyond a certain extent, since the articulatory gestures that result in the speech segments merge together, producing continuous variation of acoustic parameters. The durational patterns of speech being dependent on a large number of mutually interacting
factors, including socio-cultural, physiological and linguistic factors, the duration model should be trained to capture these patterns by following the way in which, human cognition system acquires internal prosodic models, corresponding to different speaking styles.

The analysis of duration patterns of natural speech has shown that

1. Even though the duration values of individual phonemes vary, the linguistic constructs like phrases, words, subword and morphemes have a particular duration pattern, for a particular speaking style. The variation in individual values is due to factors including positional and contextual factors.

2. The rhyming words have similar duration patterns. The similarity is more if the sequence of vowels is the same, which indicates that vowels are the deciding factor in case of duration values.

3. The clustering of words using k-means method provided evidence of clusters of duration patterns. It confirmed the notion that, it is the sequence of vowels that dictates the duration patterns of speech constructs.

When the memory based duration model was tested with a database of 9208 phonemes, match was obtained at phrase, word or morpheme level for 82.56% phonemes (7206). The root mean square error obtained was 4.86 ms, correlation was 0.96 and mean opinion score was 3.62.