J. Software Engineering & Applications, 2010, 3, 1107-1117
doi:10.4236/jsea.2010.312129 Published Online December 2010 (http://www.scirp.org/journal/jsea)

Selection of Suitable Features for Modeling the Durations of Syllables

Krothapalli S. Rao, Shashidhar G. Koolagudi

School of Information Technology, Indian Institute of Technology, Kharagpur, India.
Email: ksrao@iitkgp.ac.in, koolagudi@yahoo.com

Received November 7th, 2010, revised November 21th, 2010, accepted December 7th, 2010.

ABSTRACT

Acoustic analysis and synthesis experiments have shown that duration and intonation patterns are the two most important prosodic features responsible for the quality of synthesized speech. In this paper a set of features are proposed which will influence the duration patterns of the sequence of the sound units. These features are derived from the results of the duration analysis. Duration analysis provides a rough estimate of features, which affect the duration patterns of the sequence of the sound units. But, the prediction of durations from these features using either linear models or with a fixed rulebase is not accurate. From the analysis it is observed that there exists a gross trend in durations of syllables with respect to syllable position in the phrase, syllable position in the word, word position in the phrase, syllable identity and the context of the syllable (preceding and the following syllables). These features can be further used to predict the durations of the syllables more accurately by exploring various nonlinear models. For analying the durations of sound units, broadcast news data in Telugu is used as the speech corpus. The prediction accuracy of the duration models developed using rulebases and neural networks is evaluated using the objective measures such as percentage of syllables predicted within the specified deviation, average prediction error (μ), standard deviation (σ) and correlation coefficient (γ).

Keywords: *Prosody, Syllable Duration, Syllable Position, Syllable Context, Syllable Identity, Feed Forward Neural Network*

1. Introduction

Human beings use durational and intonation patterns on the sequence of sound units, while producing speech. It is these prosody constraints (duration and intonation), that lend naturalness to human speech. Lack of this knowledge can easily be perceived, for instance, in the speech synthesized by a machine. Even though human beings are endowed with this knowledge, they are not able to express it explicitly. But it is necessary to acquire, represent and incorporate this prosody knowledge for synthesizing speech from a text. Speech signal carries information about the message to be conveyed, speaker and language in the prosody constraints, and these prosody cues aid human beings to get the message, and identify speaker and language. The prosody knowledge also helps to overcome perceptual ambiguities. Thus, acquisition and incorporation of prosody knowledge is essential for developing speech systems [1-3].

1.1. Manifestation of Prosody Knowledge

Prosody can be viewed as speech features associated with larger units (than phonemes) such as syllables, words, phrases and sentences. Consequently, prosody is often considered as suprasegmental information. The prosody appears to structure the flow of speech, and is perceived as melody and rhythm. The prosody is represented acoustically by a pattern of duration and intonation (F0 contour). The prosody can be distinguished at four principal levels of manifestation [4]. They are at 1) Linguistic intention level, 2) articulator level, 3) acoustic realization level and 4) perceptual level.

At the linguistic level, prosody refers to relating different linguistic elements to each other, by accentuating certain elements of a text. For example, the linguistic distinctions that can be communicated through prosodic means are the distinction between question and statement, or the semantic emphasis of an element with respect to previously enunciated material.

At the articulator level, the prosody is physically manifested as a series of articulator movements. Thus, prosody manifestations typically include variations in the amplitudes of articulator movements, variations in air pressure, and specific patterns of electric impulses in nerves leading to the articulator musculature.

Muscle activity in the respiration system, and along the vocal tract leads to emission of sound waves. The acoustic realization of prosody can be observed and quantified using acoustic signal analysis. The main acoustic parameters bearing on prosody are fundamental frequency ($F0$), intensity and duration. For example, stressed syllables have higher fundamental frequency, greater amplitude and longer duration than unstressed syllables.

Finally, speech sound waves enter the ear of the listener who derives the linguistic and paralinguistic information from prosody via perceptual processing. At the level of perception, prosody can be expressed in terms of subjective experience of the listener, such as pauses, length, melody and loudness.

It is difficult to process or analyze the prosody through speech production or perception mechanisms. Hence the acoustic properties of speech are exploited for analyzing the prosody. In the next section we will discuss some of the sources of knowledge that are present in the speech signal.

1.2. Implicit Knowledge in Speech Signal

For illustration, we demonstrate some of the knowledge sources present in speech signal. **Figure 1** shows a speech signal and its transcription, energy con- tour, pitch contour and spectrogram.

The waveform shown in **Figure 1(a)** represents the time domain representation of a speech signal, the abscissa (X-axis) indicates the timing information and the ordinate (Y-axis) indicates the amplitude of speech samples.

The transcription **(Figure 1(b)**) represents the sequence of sound units and their boundaries. This gives information about the identities of sound units present in the speech signal and their durations. Energy contour (**Figure 1(c)**) indicates the distribution of energy in different regions of the speech signal, and also gives a rough indication of the voiced and nonvoiced regions. Pitch contour (**Figure 1(d)**) indicates the global and local patterns of intonation. Global intonation pattern refers to the characteristics of the whole sentence or phrase. A rising intonation pattern at the global level indicates that the sentence (phrase) is interrogative, and a declining intonation pattern indicates a declarative sentence. Local fallrise patterns indicate the nature of words and basic sound units. The spectrogram (**Figure 1(e)**) is used to represent the speech intensity in different frequency bands as a function of time. In **Figure 1(e)**, the ordinate is the frequency axis, and the grey value indicates the energy (intensity) of speech signal. The dark bands in the spectrogram represent the resonances of the vocal tract system. These resonances are also called formant frequencies, which represent the high energy regions in the frequency spectrum of a speech signal. These formant frequencies are distinct for each sound unit. The shape of the sequence of dark bands indicates the changes in the shape of the vocal tract from one sound unit to the other. Speech signal also contains information about semantics, language, speaker identity and emotional state of the speaker, which are difficult to represent quantitatively.

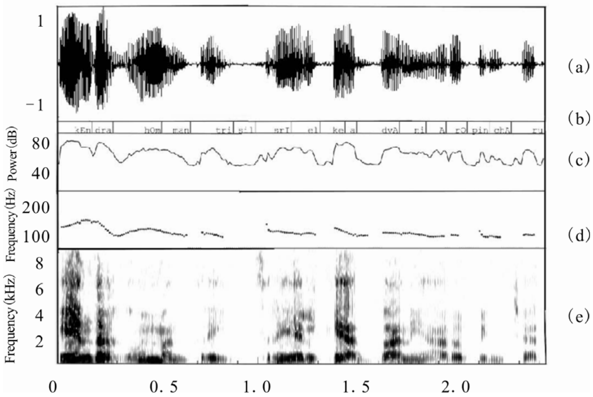

Figure 1. (a) Speech signal; (b) Transcription for the utterance "*kEndra hOm mantri srI el ke advAni ArOpinchAru*"; (c) Energy contour; (d) Pitch contour and (e) wideband spectrogram.

The basic goal of the paper is to identify the basic factors and important features, which influence the durations of the sequence of sound units present in speech. It is known that these durations depend on the linguistic and production constraints, and expressing these constraints using categorical features is a difficult task [5-7]. In this paper we present the analysis of durations of sound units with respect to phonological, positional and contextual factors. The analysis is performed on the broadcast news data for the Indian language Telugu. It is the Dravidian language with the largest number of speakers (approximately 75 million), the second most spoken language in India after Hindi. A similar analysis can be carried out for deriving the features required for modeling the intonation patterns.

The paper is organized as follows: The database used for duration analysis is described in Section 2. Computation of average durations and their deviations from the base durations for initial and final syllables is discussed in Section 3. Section 4 describes the analysis of durations of syllables using positional and contextual factors. Detailed analysis is performed in Section 5 by categorizing the syllables based on the size of the word and position of the word in the utterance. In Section 6, broad observations from the duration analysis carried out in Sections 3-5, are presented. Prediction of durations of syllables using rulebases and neural networks is discussed in Section 7. Brief summary and future extensions to the paper are presented in Section 8.

2. Speech Database

The database for this study consists of 20 broadcast news bulletins in Telugu. These news bulletins are read by male and female speakers. The total duration of speech in the database is 4.5 hours. The speech signal was sampled at 16k Hz and represented as 16 bit numbers. The speech is segmented into short utterances of duration of around 2 to 3 seconds. The speech utterances are manually transcribed into text using common transliteration code (ITRANS) for Indian languages [8]. The speech utterances are segmented and labeled manually into syllable like units. Each bulletin is organized in the form of syllables, words and orthographic text representations of the utterances. Each syllable and word file contains the text transcriptions and timing information in number of samples. The database consists of 6,484 utterances with 25,463 words and 84,349 syllables [9,10].

In this work, we have chosen syllable as the basic unit for the analysis. The syllable is a natural and convenient unit for speech in Indian languages. In Indian scripts characters generally correspond to syllables. A character in an Indian language script is typically in one of the following forms: V, CV, CCV, CCVC and CVCC, where C is a consonant and V is a vowel. Syllable boundaries are more precisely identified than phonemic segment in both the speech waveform and in the spectrographic display [11].

3. Computation of Durations

In order to analyze the effects of positional and contextual factors, syllables need to b categorized into groups based on position and context. Syllables at word initial position, middle position and final position are grouped as initial syllables, middle syllables and final syllables, respectively. Syllables next to initial syllables are grouped as following syllables, while syllables before the final syllables are grouped as preceding syllables. Words with only one syllable are treated as monosyllabic words, and the syllables are known as monosyllables. In Telugu language the occurrence of monosyllables is very less.

To analyze the variations in durations of syllables due to positional and contextual factors, reference duration of the syllable is needed [12]. The reference duration, also known as base duration, the effect of any of the factors should be minimum. “For this analysis only the middle syllables are considered as neutral syllables, where the effects of positional and contextual factors are minimum. The base duration of a syllable is obtained by averaging the durations of all the middle syllables of that category. In this analysis, some of the initial/final syllables have no reference duration, because they are not available in the word middle position. This subset of syllables is not considered for analysis.

For analyzing the effect of positional factors, initial and final position syllables are considered. This analysis consists of computation of mean duration of initial/final syllables an their deviation from their base durations. The deviations are expressed in percentage. For analyzing the gross behavior of positional factors, a set of syllables is considered, whose frequency of occurrence is greater than a particular threshold (threshold = 20) across all bulletins. This set consists of about 60 to 70 syllables, and is denoted as the set of common syllables. Most of these syllables are terminated with vowels and a few are terminated with consonants. **Table 1** shows the percentage deviations of durations of the initial syllables terminating with vowels. **Table 2** shows the percentage deviations of durations of the final syllables terminating with vowels. In these tables, the leftmost column indicates the consonant part of the syllable (CV or CCV) and the top row indicates the vowel part of the syllable. The other entries in the tables represent the percentage deviations of durations of the syllables. The blank entries in the tables correspond to syllables whose frequency of occurrence is less than a threshold (threshold = 20) across all bulletins.

Contextual factors deal with the effect of the preceding and the following unit on the current unit. In this analysis, middle syllables are assumed as neutral syllables. Therefore, initial and final syllables are analyzed for contextual effects. For initial syllable, only the effect of the following unit is analyzed. For final syllable, only the effect of the preceding syllable is analyzed. To perform the analysis, the following and preceding syllables need to be identified. The percentage deviation of duration is computed for all initial and final syllables. For each following syllable, the mean of the percentage deviation of durations of all corresponding initial syllables is computed. Likewise for each preceding syllable, the mean of the percentage deviation of durations of all corresponding final syllables is computed. These average deviations represent the variations in durations of initial and final syllables due to their following and preceding units, respectively. **Table 3** shows the percentage deviations of durations of the initial syllables due to their following syllables terminating with vowels. **Table 4** shows the percentage deviations of durations of final syllables due to their preceding syllables terminating with vowels.

4. Analysis of Duration

Durations of the syllables are analyzed using positional and contextual factors. The effect of positional factors is analyzed by observing the durations of initial and final syllables. The effect of contextual factors is analyzed by observing the durations of initial and final syllables with respect to their following and preceding syllables. The following subsections summarize the effects of positional and contextual factors on syllable duration.

4.1. Positional Factors

From **Table 1**, it is observed that most of the syllables at word initial position have durations more than their base durations. The percentage deviations of durations of all the initial syllables are not uniform. They vary based on manner of articulation, place of articulation and voicing nature associated with the production of the syllable. At a primary level, it is noticed that syllables with voicing nature (consonant within a syllable is of voicing nature) have more deviations in durations compared to their unvoiced counterparts. Again, within the voiced and unvoiced categories, a variation in duration is observed based on the manner and place of articulation and on the nature of vowel present in the syllable. From the analysis of word final position syllables (**Table 2**), it is observed that most of the syllables terminating with vowels (i.e., CV type) have larger duration compared to their base duration. Bilabial stops, bilabial nasals and fricative group of syllables do not belong to the set of common syllables. From the final syllables of the set of common syllables, a broad grouping can be performed based on the vowel inside the syllable. **Table 2** shows that syllables with the vowel */a/* have deviations in duration between 20% and 30%, while syllables with vowel */i/* and */u/* have about 40% to 60% deviations.

4.2. Contextual Factors

The effect of contextual factors on the initial and final syllables is given (in the form of percentage deviations of durations of syllables) in **Tables 3** and **4**, respectively. The initial syllable duration is close to its reference duration in the case of syllables with semivowels or fricatives in following position. The duration of initial syllable increases by 10% to 20% of its base duration, when nasalized syllable is the following unit. Trills and liquids increase the duration of their preceding syllables by 25% to 35%. Syllables with unvoiced stops at the following position affect the durations of their preceding syllables (initial syllables) more compared to syllables with voiced stops at the following position.

The final syllable duration increases by 20% to 30%, if the preceding syllable contains unvoiced stop consonant or semivowel. Syllables with voiced stop consonants and trills affect the duration of the following final syllables by 30% to 40%. The final syllable duration is increased by 45%, if liquid category syllables are in the preceding position. Nasals in the preceding position increase the duration of final syllables by approximately 30%. Fricative based syllables increase the duration of

Table 1. Percentage deviations of durations of the initial syllables terminating with vowels. The entries in the leftmost column and top row indicate the consonant and vowel parts of the syllable.

	a	**A**	**i**	**I**	**u**	**U**	**e**	**E**	**o**	**O**
k	23	14	31		20	3		14	23	27
ch			32				1	3	10	
t	-2	19	2	11			29	0		
p	5	2	18	9	2		14	24	22	31
g	103	45			70					67
j	37	44								
d	51	36	42					24		
b	71	31	81	60						
bh	40	48								
m	51	44	56		34	48	61			
n	51	15	48	37				40		
r	47	61	58		60			40		8
v	40	39	33	62			40	34		
s	17	15								
sh		19								

Table 2. Percentage deviations of durations of the final syllables terminating with vowels. The entries in the leftmost column and top row indicate the consonant and vowel parts of the syllable.

	a	**A**	**i**	**I**	**u**	**U**	**e**	**E**	**o**	**O**
k	28		54		55					
ch			90						28	
T	17		54		45					
t	34	25	46		53	33				12
g		18			62					
j		18	54		51					
D	26		84		53					
d			65		49					
n	22		52		44					
l	17	0			58					31
y	37	-7	53							
r	20		43		62					
v	22									
Sh	13		5							

Table 3. Percentage deviations of durations of initial syllables due to their following syllables terminating with vowel. The syllables indicated are following syllables. The entries in the leftmost column and top row indicate the consonant and vowel parts of the following syllable.

	a	**A**	**i**	**I**	**u**	**U**	**e**	**E**	**o**	**O**
k	32				-1					
ch	34	36	9							
T	30		19	18	21					
t	58	27	28	58	17					
p	21	8	49		34			22		
g	30	21	29		32					
j	29				10					
D	16	4	17		23					
d	17		9	8	15					
b			12		3					
bh	11									
m	11	21	20		21		23			22
n	23	15	15		18					
y	9	4	-1		-7					
r	35	35	21		28					
l	25	30	33	27	25			24		13
v	7	6	8	11				8		
s	8	4	10		5			-1		
Sh	3	4								

Table 4. Percentage deviations of durations of the final syllables due to their preceding syllables terminating with vowel. The syllables indicated are preceding syllables. The entries in the leftmost column and top row indicate the consonant and vowel parts of the preceding syllable.

	a	**A**	**i**	**I**	**u**	**U**	**e**	**E**	**o**	**O**
k	11	20			21					18
ch	13	17	21					22		
T	28	20	22		29					
t	14	27	26	26	20					
p	22	26	19		28			26		-3
g	32	22	39		41					
j	37				32					
D	31	23	34		39					
d	35	23	29		33			29		
m	31	26	28							
n	28	33	30							
l	41		34		52					47
y	24	27	25		28			34		
r	32	32	36		33					
v	18	32	28		28					
s	6	8	14							
Sh	8		1							

the following syllables by 10%. The important contextual effect observed is that syllables with unvoiced stop consonants affect the durations of their preceding syllables more compared to their following syllables, whereas syllables with voiced stop consonants affect the durations of their following syllables more.

5. Detailed Duration Analysis

In the analysis (Section 4), a wide range of durational variations are observed in both initial and final syllables. This is due to the dependency of syllable duration on size of the word and position of the word in the utterance. Hence, for a detailed duration analysis, the initial/final syllables need to be categorized further based on word size and position, and the analysis needs to be performed separately on different categories. Analysis of durations of the syllables with respect to position of the word and size of the word is performed in the following subsections.

5.1. Analysis of Durations Based on Position of the Word

To perform this analysis, words are classified into groups based on their position in the utterance. The word positions considered for the analysis are first, middle, and last, denoted by W_f, W_m and W_l, respectively. From each group of words the following analysis is performed: Initial and final syllables, their adjacent syllables and their associated durations are derived. The average deviations of the durations of the initial and final syllables are computed using positional and contextual factors. The set of syllables present in each category above some threshold of frequency of occurrence is used for analysis. **Table 5** shows the percentage deviations of durations of initial and final syllables present in the words at different positions in the utterance. **Table 6** shows the percentage deviations of durations of the initial and final syllables due to their associated context (following and preceding syllables) present in the words at different positions in the utterance.

The following are the inferences drawn from the **Table 5**:

- Initial syllables present in W_m have more duration compared to initial syllables in W_f and W_l.
- Initial syllables with unvoiced consonants present in W_f have durations lesser than their base durations.
- Initial syllables from nasal and trill categories present in W_f have greater durations compared to other word positions.
- The initial syllables terminating with long vowels have more duration in W_f, whereas the initial syl-

lables terminating with consonants have more duration in W_l.

- The final syllables of W_l have larger duration compared to final syllables of W_f and W_m, and among the syllables of W_f and W_m, syllables of W_f have larger durations.
- In comparison with the initial syllables of W_l, the final syllables of W_l have larger deviations in durations.

The following are the inferences drawn from the **Table 6**

- The initial syllables present in W_m and W_l are more lengthened due to their following syllables.
- The final syllables of W_f and W_l are more lengthened compared to the initial syllables of W_f and W_l due to their context.
- The durations of the final syllables of W_l are more lengthened due to their preceding syllables, compared to other word positions.

5.2. Analysis of Durations Based on Size of the Word

In this analysis, words are categorized into groups based on the number of syllables they contain. In this study words are classified into six groups. They are monosyllabic, bisyllabic, trisyllabic, tetrasyllabic, pentasyllabic and polysyllabic words, containing of one, two, three, four, five and more than five syllables, respectively. Monosyllabic words are not considered for analysis, since they are very few in number. Analysis is performed separately for the other five groups. **Table 7** shows the percentage deviations of durations of the initial syllables and final syllables present in different word sizes. **Table 8** shows the percentage deviations of durations of the initial and final syllables due to their adjacent syllables.

The results of the analysis indicates that, the durations of the initial and final syllables in different word sizes are inversely related to word size. That is, the durations of the initial/final syllables in bisyllabic words are more compared to the durations of the initial/final syllables in polysyllabic words. The mean durations of the final syllables from various word sizes are more, compared to the initial syllables of their corresponding categories.

The deviations of durations of the initial and final syl-

Table 5. Percentage deviations of durations of initial and final syllables present in the words that occur at different positions in the utterance.

Initial Syllables				Final Syllables			
	First	Mid	Last		First	Mid	Last
a	7	46	30	**Du**	38	41	95
ba	47	85	51	**TI**	33	19	47
da	47	55	43	**Tu**	44	21	98
ga	94	113	77	**chi**	80	79	104
ka	-8	43	34	**du**	43	29	71
na	62	54	52	**ga**	54	27	115
pa	-25	22	13	**ju**	37	37	77
reN	32	49	51	**ka**	28	19	82
ta	-29	24	17	**ku**	37	34	87
bA	72	59	38	**la**	14	9	68
dE	45	39	28	**lu**	49	46	88
ja	64	67	37	**na**	22	14	48
kAr	-3	11	20	**pu**	18	11	72
nA	47	42	13	**ra**	25	8	79
rA	68	62	53	**ram**	10	1	14
vi	20	45	40	**sham**	7	1	20
mukh	11	33	57	**ta**	26	22	76
rASh	17	25	26	**tri**	42	31	52
				vam	5	-3	16
				ya	35	22	97

Table 6. Percentage deviations of durations of initial and final syllables due to their adjacent syllables (following and preceding syllables for initial and final syllables, respectively) in the words that occur at different positions in the utterance. Syllables in the first column are following syllables to the initial syllables and syllables in the fifth column are preceding syllables to the final syllables.

Following Syllables				Preceding Syllables			
	First	Mid	Last		First	Mid	Last
Du	35	45	49	**Da**	30	27	34
TI	-7	14	10	**Sha**	9	4	24
bhut	0	31	32	**Ta**	20	13	46
da	6	16	28	**bhut**	42	43	64
di	-6	16	13	**da**	21	20	63
ga	40	32	22	**dhA**	4	2	44
ju	38	52	62	**du**	19	8	68
la	11	31	22	**ga**	14	1	43
lu	0	34	35	**la**	29	27	61
ma	-15	35	33	**ma**	24	13	67
na	13	25	31	**man**	41	28	50
ni	6	21	8	**na**	28	16	49
ra	22	45	34	**ni**	19	66	72
ru	6	39	36	**pa**	36	6	41
ta	52	64	59	**ra**	17	23	67
va	28	27	25	**ri**	31	28	58
ya	-4	12	14	**sa**	8	1	34
				ta	12	13	25
				tu	27	24	51
				va	17	6	43
				ya	29	12	63

Table 7. Percentage deviations of durations of initial and final syllables present in different word sizes. legend: Tr-Tri, Te-Tetra, Pe-Penta, P-Poly.

Initial Syllables						Final Syllables					
	Bi	Tr	Te	Pe	P		Bi	Tr	Te	Pe	P
a	41	37	29	27	34	**Du**	46	41	61	59	33
chE	13	4	3	3	3	**Ti**	68	54	29	31	24
ka	52	20	30	30	13	**chi**	93	83	94	88	82
ku	30	21	21	18	16	**da**	80	26	17	23	-5
ma	48	52	52	51	46	**gu**	64	20	-1	10	24
nA	31	8	16	6	31	**ka**	37	24	21	23	47
na	76	71	28	51	40	**la**	59	17	8	13	13
ni	61	45	50	38	49	**nu**	50	43	40	41	37
pa	28	10	0	6	2	**ra**	52	6	15	-6	18
ra	27	21	23	21	21	**ru**	45	24	19	15	24
sa	19	17	19	11	19	**si**	61	34	38	11	-5
tI	22	13	4	19	8	**ti**	38	29	43	21	25
vi	42	35	32	33	32	**ya**	69	25	27	30	27
kAr	27	26	-1	-7	-7						

lables due to contextual factors are less in magnitude, compared to deviations in durations due to positional factors. The percentage deviations of durations of the initial syllables due to their following units are inversely related to size of the word, whereas for the final syllables, the deviations in durations due to their preceding units are proportional to size of the word.

So far duration analysis is performed on the whole set of initial/final syllables, and on the categorized initial/final syllables based on position/size of the word in the utterance. In the analysis a large variation in duration within each category of syllables is observed. For more detailed analysis, there is a need to classify the syllables further by considering the size of the word and position of the word together. With this classification, words can be categorized into 15 groups (3 × 5 = 15, using position of the word 3 groups, and size of the word 5 groups).

6. Observation and Discussion

From the above analysis (Section 4 and Section 5) it is observed that duration patterns for the sequence of syllables depends on different factors at various levels. It is difficult to derive a finite number of rules which characterizes the behavior of the duration patterns of the syllables. Even to fit the linear models to characterize the durational behavior of the syllables is also difficult. Since the linguistic features associated to different factors have complex interactions at different levels, it is difficult to derive a rulebase or linear model to characterize the durational behavior of the syllables.

To overcome the difficulty in modeling the duration patterns of the syllables, nonlinear models can be explored. Nonlinear models are known for their ability to capture the complex relations between the input and output. The performance of the model depends on the quality and quantity of the training data, structure of the model and training and testing topologies [13]. From the analysis carried out in sections 4 and 5, we can identify the features that affect the duration of a syllable. The basic factors and its associated linguistic features that affect the durations of syllables are given in **Table 9**. Nonlinear models can be developed using these features as input and the corresponding duration as the output to predict the durations of syllables [14-16]. Similar analogy can be used for modeling the intonation patterns [17,18].

Even though the list of the factors affecting the duration may be identified, but it is difficult to determine their effect independently. This is because the duration patterns of the sequence of sound units depend on several factors and their complex interactions. Formulation of

Table 8. Percentage deviations of durations of initial and final syllables due to their adjacent syllables (following and preceding syllables for initial and final syllables, respectively) present in different word sizes. Syllables in the first column are following syllables to the initial syllables and syllables in the seventh column are preceding syllables to final syllables. Legend: T-Tri, Te-Tetra, Pe-Penta, P-Poly.

Following Syllables						Preceding Syllables					
	Bi	T	Te	Pe	P		Bi	T	Te	Pe	P
Du	52	57	45	44	30	**chA**	10	24	21	13	36
Ta	25	21	21	13	16	**da**	-25	30	30	32	49
da	45	15	16	18	3	**gA**	27	17	20	32	39
di	22	-2	14	-10	7	**ku**	-22	24	41	23	33
li	51	43	23	12	40	**la**	28	24	47	41	38
lu	30	26	23	21	22	**ma**	12	14	40	56	62
mi	38	20	21	16	-11	**man**	32	39	40	27	43
nu	22	7	19	25	11	**nA**	23	24	20	20	33
ni	37	35	16	5	15	**na**	11	15	30	30	54
pu	53	13	13	-8	40	**ni**	27	0	24	43	58
ra	46	45	25	43	24	**rA**	9	26	35	33	48
ri	30	18	25	18	9	**ra**	34	18	43	40	68
ru	51	22	31	39	27	**shA**	20	13	33	27	27
si	25	8	4	10	-2	**vA**	26	27	36	43	49
su	12	4	7	8	0	**yA**	19	36	35	28	41
ta	70	50	47	65	52						
ti	34	36	27	24	19						
va	51	24	23	31	22						
ya	16	5	8	10	10						

Table 9. List of the factors and its associated features affecting the syllable duration.

Factors	Features
Syllable position in the phrase	Position of syllable from beginning of the phrase Position of syllable from end of the phrase Number of syllables in the phrase
Syllable position in the word	Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables
Word position in the phrase	Position of word from beginning of the phrase Position of word from end of the phrase Number of words in a phrase
Syllable identity	Segments of the syllable (consonants and vowels)
Context of the syllable	Previous syllable Following syllable
Syllable nucleus	Position of the nucleus Number of segments before the nucleus Number of segments after the nucleus
Gender identity	Gender of the speaker

these interactions in terms of either linear or nonlinear relations is a complex task. For example, in the analysis of positional factors (**Tables 1** and **2**) the deviation in the durations include the effect of contextual factors, nature of the syllable, word level and phrase level factors. Similarly in the analysis of contextual factors, the contributions of other factors are also included.

From the duration analysis one can identify the list of features affecting the duration, but it is very difficult to derive the precise rules for estimating the duration. The analysis performed in this paper may be useful for some speech applications where the precise estimation of durations is not essential. For example, in the case of speech recognition, rule-based duration models can provide a supporting evidence to improve the recognition rate [19]. This is particularly evident in speech recognition in noisy environment [20]. In the case of speaker recognition, speaker-specific duration models will give an additional evidence, which can be further used to enhance the recognition performance [21]. Duration models are also useful in language identification task, since the duration patterns of the sequence of sound units are unique to a particular language [22-24].

In the duration analysis, the numbers shown in the tables indicate the average deviations. These models may not be appropriate for Text-to-Speech (TTS) synthesis application. In TTS synthesis, precise duration models produce speech with high naturalness and intelligibility [25]. The naturalness mainly depends on the accuracy of prosody models. The derived duration models in the paper may not be appropriate for high quality TTS applications, but they can be useful in developing other speech systems such as speech recognition, speaker recognition and language identification. Precise duration models can be derived by using nonlinear models such as neural networks, support vector machines and classification and regression trees. A nonlinear model will give the precise duration by providing 1) all the factors and features responsible for the variation in duration as input, 2) each category of the sound unit has enough examples and (3) the database should contain enough diversity [13].

7. Prediction of Durations

In this section, prediction performance of the duration-models is illustrated using (a) Rulebase derived from the manual analysis and (b) Feed forward neural network model. For carrying out this prediction analysis, 15 news bulletins (60,763 syllables) are used for training and 5 news bulletins (23, 586 syllables) are used for testing. The prediction analysis using rulebase is carried out with three different models: 1) Rulebase is derived from the duration analysis of positional and contextual factors of all the syllables together. 2) Rulebase derived from the duration analysis of positional and contextual factors of the syllables categorized based on the position of the words. 3) Rulebase derived from the duration analysis of positional and contextual factors of the syllables categorized based on the size of the words. In each case the rulebase is derived from the syllables of 15 news bulletins, and evaluated using the syllables of 5 news bulletins. The prediction accuracy of the duration models is analyzed using (a) % of syllables predicted within different deviations from the actual durations and (b) objective measures such as (i) average prediction error, (ii) standard deviation and (iii) correlation coefficient. For each syllable the deviation (D_i) is computed as follows:

$$Di = \frac{|x_i - y_i|}{x_i} \times 100$$

where *xi* and *yi* are the actual and predicted durations, respectively. The definitions of average prediction error (μ), standard deviation (σ), and linear correlation coefficient ($\gamma X, Y$) are given below:

$$\mu = \frac{\sum_i |x_i - y_i|}{N}$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \quad d_i = ei - \mu, \quad ei = x_i - y_i,$$

where x_i, y_i are the actual and predicted durations, respectively, and *ei* is the error between the actual and predicted durations. The deviation in error is d_i, and *N* is the number of observed syllable durations. The correlation coefficient is given by

$$\gamma X, Y = \frac{VX, Y}{\sigma X . \sigma Y}, \quad \text{where} \quad VX, Y = \frac{\sum_i | x_i - \bar{x} | . | y_i - \bar{y} |}{N}.$$

The quantities σ_x, σ_y are the standard deviations of the actual and predicted durations, respectively, and V_x, *Y* is the correlation between the actual and predicted durations. The prediction performance of the duration models using different rulebases is given in **Table 10** The first column indicates three different rulebases derived from the training data. The irst one indicate the rulebase derived from the gross duration analysis (i.e., analyzing the durations of all syllables). Second and third rulebases are derived from the refined duration analysis (i.e., analyzing the durations of syllables based on posi tion of the word and size of the word). Second column of the **Table 10** indicates the basic influencing factors of the durations of the sound units, with which the rule-bases are derived. Columns 3-7, indicate the % of syllables predicted within 2%, 5%, 10%, 25% and 50% deviations from their actual durations. Columns 8-10, indicate objective measures of the prediction accuracy. From the results, it is observed that the accuracy of prediction has improved by using the rulebases derived from the syllables categorized based on position/size of the words, compared to whole set of the syllables. Here, in the manual analysis, the durations are predicted separately using positional and contextual factors. Combining the positional and contextual factors in the analysis of durations is very hard in deriving the rulebase, because of the difficulty in capturing the complex nonlinear interactions between them. Therefore, we propose a neural network model, which suppose to capture the complex interactions among the factors as well as their associated duration patterns.

A four layer Feed forward Neural Network (FFNN) is used for modeling the durations of syllables. The general structure of the FFNN is shown in **Figure 2** Here the FFNN model is expected to capture the functional relationship between the input and output feature vectors of the given training data.

The mapping function is between the 25-dimensional input vector and the 1-dimensional output. It is known that a neural network with two hidden layers can realize any continuous vector-valued function [26]. The first layer is the input layer with linear units. The second and third layers are hidden layers. The second layer (first hidden layer) of the network has more units than the input layer, and it can be interpreted as capturing some local features in the input space. The third layer (second hidden layer) has fewer units than the first layer, and can be interpreted as capturing some global features [13,27]. The fourth layer is the output layer having one unit representing the duration of a syllable. The activation function for the units at the input layer is linear, and for the units at the hidden layers, it is nonlinear. Generalization by the network is influenced by three factors: The size of the training set, the architecture of the neural network, and the complexity of the problem. We have no control over the first and last factors. Several network structures were explored in this study. The (empirically arrived) final structure of the network is 25*L* 50*N* 12*N* 1*N*, where *L* denotes a linear unit, and *N* denotes a nonlinear unit. The integer value indicates the number of units used in that layer. The nonlinear units use *tanh*(*s*) as the activation function, where *s* is the activation value of that unit. For studying the effect of the positional and contextual factors on syllable duration, the network structures 14*L* 28*N* 7*N* 1*N* and 13*L* 26*N* 7*N* 1*N* are used, respectively. The proportions of the number of units in each layer are similar as in the earlier network. The inputs to these networks represent the positional and contextual factors. All the input and output features are normalized to the range [-1, +1] before presenting them to the neural network. The back propagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error for each syllable duration [27].

Table 10. Performance of the duration models using different rulebases. Legend: Po-Positional, Co-Contextual.

Type of rulebase	Factors	% of predicted syllables within dev.					Objective measures		
		2%	5%	10%	25%	50%	μ (ms)	σ (ms)	γ
General	Po	1	5	19	51	81	45	37	0.63
	Co	1	4	20	47	79	48	40	0.61
Position of word	Po	2	6	25	57	85	41	35	0.65
	Co	1	5	24	56	83	43	37	0.62
Size of word	Po	2	7	24	53	87	40	33	0.66
	Co	1	5	22	56	83	44	38	0.62

For studying the effect of positional and contextual factors on syllable duration, the features associated with the syllable position and syllable context were used separately. The features representing the positional factors are: 1) Syllable position in the phrase (3-dimensional feature), 2) syllable position in the word (3-dimensional feature), 3) word position in the phrase (3-dimensional feature), 4) syllable identity (4-dimensional feature) and 5) identity of gender. Features representing the contextual factors are the identities of the present syllable, its previous and following syllables and the identity of gender. Altogether, we have developed three neural network models 1) Model developed using only positional features, 2) Model developed using only contextual features and 3) Model developed using both positional and contextual features. The prediction performance of these three models is given in **Table 11**. It is observed that the accuracy of prediction is superior for the neural network models (see **Table 11**), compared to rule based models (see **Table 10**). This is due to fact that the neural net work captures the hidden interactions that are present in the features of different levels.

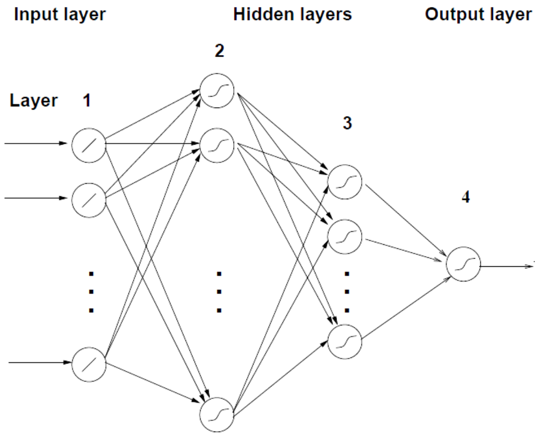

Figure 2. Four layer feed forward neural network.

Table 11. Performance of the duration models using neural networks. Legend: Pos-Positional, Con-Contextual.

Factors	% of predicted syllables within dev.					Objective measures		
	2 %	5%	10 %	25 %	50 %	μ (ms)	σ (ms)	γ
Pos	4	11	32	61	87	32	26	0.74
Con	2	9	31	59	85	35	28	0.71
All	5	12	34	66	91	29	24	0.78

8. Summary and Conclusion

Factors affecting the durations of syllables in continuous speech were identified. They are positional, contextual and phonological factors. Durations were analyzed using positional and contextual factors. In the analysis of positional factors, it was noted that the deviations in durations of the syllables depend on voicing, place and manner of articulation, and nature of vowel present in the syllable. From the analysis of contextual factors, it was mainly observed that syllables with unvoiced stop consonants affect the durations of their preceding syllables, and syllables with voiced consonants affect the durations of their following syllables. In the analysis, a wide range of durational variations was observed. For detailed analysis, categorization of syllables was suggested. In this work syllables were categorized based on size of the word and position of the word in the utterance, and the analysis was performed separately in each group. From the analysis, list of various factors and its associated features which affect the duration patterns of the sequence of sound units were identified. Nonlinear models were suggested for predicting the accurate durations of syllables from the complex interactions of various factors at different levels. Prediction performance of the duration models using rulebase and neural networks was evaluated. It was observed that the accuracy of prediction using neural network models was better compared to the models derived from rulebases. This is because of the capability of the neural networks to capture the complex nonlinear relations across the factors influencing the durations of sound units. The duration analysis can be further extended by analyzing the factors at higher level such as accent and prominence of syllables, part-of-speech (syntactic factors), semantics and the emotional state of the speaker.

REFERENCES

[1] K. S. Rao, “Acquisition and Incorporation Prosody Knowledge for Speech Systems in Indian Languages,” Ph.D. Thesis, Indian Institute of Technology Madras, Chennai, May 2005.

[2] L. Mary, K. S. Rao, S. V. Gangashetty and B. Yegnanarayana, “Neural Network Models for Capturing Duration and Intonation Knowledge for Language and Speaker Identification,” *International Conference on Cognitive and Neural Systems*, Boston, May 2004.

[3] A. S. M. Kumar, S. Rajendran and B. Yegnanarayana, “Intonation Component of Text-to-Speech System for Hindi,” *Computer Speech and Language*, Vol. 7, No. 3, 1993, pp. 283-301. doi:10.1006/csla.1993.1015

[4] S. Werner and E. Keller, “Prosodic Aspects of Speech,” *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, the Future Challenges*, E. Kelle Edition, John Wiley, Chichester, 1994. pp. 23-40.

[5] K. K. Kumar, "Duration and Intonation Knowledge for Text-to-Speech Conversion System for Telugu and Hindi," Master's Thesis, Indian Institute of Technology Madras, Chennai, May 2002.

[6] S. R. R. Kumar, "Significance of Durational Knowledge for a Text-to-Speech System in an Indian Language," Master's Thesis, Indian Institute of Technology Madras, Chennai, March 1990.

[7] O. Sayli, "Duration Analysis and Modeling for Turkish Text-to-Speech Synthesis," Master's Thesis, Bogaziei University, Istanbul, 2002.

[8] A. Chopde, "Itrans Indian Language Transliteration Package Version 5.2 Source." http://www.aczone.com/itrans/.

[9] A. N. Khan, S. V. Gangashetty and S. Rajendran, "Speech Database for Indian Languages—A Priliminary Study," *International Conference on Natural Language Processing*, Mumbai, December 2002, pp. 295-301.

[10] A. N. Khan, S. V. Gangashetty and B. Yegnanarayana, "Syllabic Properties of Three Indian Languages: Implications for Speech Recognition and Language Identification," *International Conference on Natural Language Processing*, Mysore, December 2003, pp. 125-134.

[11] O. Fujimura, "Syllable as a Unit of Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 23, No. 1, 1975, pp. 82-87. doi:10.1109/TASSP.1975.1162631

[12] D. H. Klatt, "Review of Text-to-Speech Conversion for English," *Journal of Acoustic Society of America*, Vol. 82, No, 3, 1987, pp. 737-793. doi:10.1121/1.395275

[13] S. Haykin, "Neural Networks: A Comprehensive Foundation", Pearson Education Aisa, Inc., New Delhi, 1999.

[14] M. Riedi, "A Neural Network Based Model of Segmental Duration for Speech Synthesis," *Proceedings of European Conference on Speech Communication and Technology*, Madrid, September 1995, pp. 599-602.

[15] K. S. Rao and B. Yegnanarayana, "Modeling Syllable Duration in Indian Languages Using Neural Networks," *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*, Montreal, May 2004, pp. 313-316.

[16] W. N. Campbell, "Predicting Segmental Durations for Accommodation within a Syllable-Level Timing Framework," *Proceedings of European Conference on Speech Communication and Technology*, Berlin, Vol. 2, September 1993, pp. 1081-1084.

[17] K. S. Rao and B. Yegnanarayana, "Intonation modeling for Indian languages," *Proceedings of International Conference on Spoken Language Processing*, Jeju Island, October 2004, pp. 733-736.

[18] M. Vainio and T. Altosaar, "Modeling the Microprosody of Pitch and Loudness for Speech Synthesis with Neural Networks," *Proceedings of International Conference on Spoken Language Processing*, Sidney, December 1998.

[19] S. Lee, K. Hirose and N. Minematsu, "Incoporation of Prosodic Modules for Large Vocabulary Continuous Speech Recognition," *Proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding*, New Jersey, 2001, pp. 97-101.

[20] K. Ivano, T. Seki and S. Furui, "Noise Robust Speech Recognition Using F0 Contour Extract by Hough Transform," *Proceedings of International Conference on Spoken Language Processing*, Denver, 2002, pp. 941-944.

[21] L. Mary and B. Yegnanarayana, "Prosodic Features for Speaker Verification," *Proceedings of International Conference on Spoken Language Processing*, Pittsburgh, September 2006, pp. 917-920.

[22] L. Mary, "Multi Level Implicit Features for Language and Speaker Recognition," Ph.D. Thesis, Indian Institute of Technology Madras, Chennai, June 2006.

[23] L. Mary and B. Yegnanarayana, "Consonant-Vowel Based Features for Language Identification," *International Conference on Natural Language Processing*, Kanpur, December 2005, pp. 103-106.

[24] L. Mary, K. S. Rao and B. Yegnanarayana, "Neural Network Classifiers for Language Identification Using Phonotactic and Prosodic Features," *Proceedings of International Conference on Intelligent Sensing and Information Processing* (*ICISIP*), Chennai, January 2005, pp. 404-408. doi:10.1109/ICISIP.2005.1529486

[25] S. R. R. Kumar and B. Yegnanarayana, "Significance of Durational Knowledge for Speech Synthesis in Indian Languages," *Proceedings of IEEE Region 10 Conference Convergent Technologies for the Asia-Pacific*, Bombay, November 1989, pp. 486-489.

[26] E. D. Sontag, "Feedback Stabilization Using Two Hidden Layer Nets," *IEEE Transactions on Neural Networks*, Vol. 3, No. 6, November 1992, pp. 981-990. doi:10.1109/72.165599

[27] B. Yegnanarayana, "Artificial Neural Network*s*," Printice-Hall, New Delhi, India, 1999.