SIGNIFICANCE OF DURATIONAL KNOWLEDGE FOR SPEECH SYNTHESIS SYSTEM IN AN INDIAN LANGUAGE

S. R. Rajesh Kumar and B. Yegnanarayana

Department of Computer Science and Engineering

Indian Institute of Technology

MADRAS 600 036 INDIA

ABSTRACT

Duration is one of the prosodic features of speech, the other two being stress and intonation. This paper demonstrates the significance of durational knowledge in speech synthesis for the Indian language, Hindi. We are developing a speech synthesis system based on parameter concatenation (Linear Prediction) model. The characters of Hindi have been taken as the basic units. A framework for collection of speech data for the basic units has been evolved, keeping in mind the flexibility needed to incorporate prosodic features during speech synthesis. The various durational effects in Hindi have been identified and categorised. Some of the durational effects have been studied and a set of durational rules formulated for use in a speech synthesis system for Hindi.

1. INTRODUCTION

The function of a text-to-speech (or speech synthesis) system is to convert the input text into a speech signal. This paper discusses some issues involved in the development of a text-to-speech system for the Indian language, Hindi. In particular, the importance of 'durational' knowledge for speech synthesis in Hindi is examined.

Quantity (duration) is one of the 'suprasegmental' (also called 'prosodic') features of speech. The other two prosodic features are stress (intensity) and intonation (pitch). The terms in the brackets refer to the acoustic manifestation of the corresponding suprasegmental feature [1]. Generally the basic unit in speech is either a phoneme or a character depending upon the language. The segmental features refer to the features which determine the phonetic quality (for example voicing or aspiration) of a basic unit. A suprasegmental feature is an overlaid function of the segmental features, as for example, pitch is an overlaid function of voicing. A further difference between segmental and suprasegmental features appears in the fact that suprasegmental features are established by a comparison of items in sequence, whereas segmental features can be defined without reference to the sequence of segments. The various prosodic features and their effect at various levels are summarized in Table 1.

In general, suprasegmental information is difficult to extract compared to segmental information. Among suprasegmental features, intensity, pitch and duration represent the increasing order of the degree of difficulty to extract. But in terms of usefulness, intensity is not very promising. On the other hand, both pitch and duration carry a lot of information. Each basic unit has a certain duration in the time domain. The perception of duration is associated with the suprasegmental feature of quantity, which involves manipulation of durations of basic units. At the sentence level, the effects of duration contribute to 'tempo' or 'rhythm'.

Table 1. Prosodic Features and their Correlates

Prosodic feature	Acoustic feature	Linguistic function	
		word level	sentence level
Quantity	Duration	Quantity	Tempo or Rhythm
Tonal	Pitch	Tone	Intonation
Stress	Loudness	Word stress	Sentence stress

In a concatenation model, the speech is produced by concatenation of prestored basic units corresponding to the input text. A basic unit could be one of the linguistic units like a phoneme, syllable, word, etc. The basic speech units may be stored in the form of raw signal data (waveform concatenation model) or in the form of parameters (parameter concatenation model). The 'context' is completely known in a text-to-speech system as opposed to speech recognition. The 'context' refers to the preceding character, the following character, the position of the character in the word, the type of the word, the location of phrase boundaries, etc. Hence text-to-speech systems can easily exploit the durational knowledge. By assigning a 'base' duration for each character and then adjusting the value of these durations based on the context, we are accounting for 'coarticulation' to a certain extent, at the character, word and sentence boundaries. This enhances the speech quality and makes it more natural sounding'.

The contributions of this work are as follows: For the present text-to-speech system, the 'character' set of an Indian language is selected as the basic units. Some guidelines are given for collection of speech data for various basic units of our system. The importance of durational knowledge for speech synthesis in Hindi is examined. The decisions regarding (i) the choice of character as the basic unit, (ii) the guidelines for the collection of speech data for the basic units, and (iii) the coding technique to be employed are taken so that the prosodic features can be easily incorporated in the existing system. Finally some durational rules specific to Hindi are formulated and the results are

CH2766 – 4/89/0000 - 0486 © 1989 IEEE

presented. The paper is organised as follows: Section 2 discusses some issues involved in the design and development of a speech synthesis system for Hindi. Section 3 discusses the importance of durational knowledge for speech synthesis in Hindi. Section 4 presents the conclusions.

2. THE IITM SPEECH SYNTHESIS SYSTEM FOR HINDI

We have adopted the parameter concatenation approach for our text-to-speech system for Hindi. The issues involved in developing a text-to-speech system based on parameter concatenation model are: (i) choosing the basic speech unit, (ii) collecting the signal data of all the basic units of the language, (iii) coding all the basic units in terms of parameters, (iv) identifying the basic units corresponding to the input text, (v) synthesis of speech from the parameters of basic units and (vi) incorporating prosodic features during synthesis [2]. Steps (ii) and (iii) involve the creation of a 'database' of speech units, which will be used by the text-to-speech system. Steps (iv) to (vi) involve conversion of the input text into speech. Each of these steps are discussed in the remaining part of the section.

2.1 Choice of the basic speech unit

The choice of the basic speech unit involves a trade-off between the size of memory needed to store all the units and the computation required during synthesis. This is because, if the size of the unit is large, the number of units in the language increases and hence larger memory is needed to store them. On the other hand, if the size of the unit is smaller, the effects of coarticulation between these units increases and hence more computation is required during synthesis. For Indian languages, which are phonetic in nature, the 'characters' are generally the orthographic representations of speech sounds. In the case of characters, most of the coarticulation effects are preserved and the number of units is not very large. Also, the characters can be extracted from the text by simple parsing, taking into account a few exceptions. Therefore, characters are chosen as the basic units in the present implementation of our text-to-speech system.

A character in most of the Indian languages represents a speech sound in the form of V, C, CV, CCV, or CCCV where V and C refer to a vowel and a consonant respectively. The vowels and consonants for Hindi are 10 and 31, respectively. The total number of characters is about 5000. Out of these, the number of cluster characters (CCV and CCCV) is very large. We have observed that the coarticulation effect between two adjacent consonants in a cluster is not significant. Therefore the cluster characters can be generated from the constituent CV combination and the other consonant(s). This results in a great reduction in the number of basic units (from 5000 to 350) and hence in the storage required. Therefore the basic units in the present implementation of the text-to-system are·

(a) Isolated vowels (V), for example a: (आ)
(b) Isolated consonants (C), for example k (क्)
(c) Combination of a consonant and a vowel (CV), for example ka: (का).

The basic speech units may be stored in the form of raw signal data (waveform concatenation model) or in the form of parameters (parameter concatenation model). The waveform concatenation model takes more memory and also has less flexibility for manipulation to take care of concatenation and coarticulation effects [3]. The parameter concatenation model provides a lot of flexibility at the cost of synthesis quality and computation. Therefore we have chosen a parameter concatenation model based on Linear Prediction Coding (LPC) for our studies [4].

2.2 Collection of speech data for the basic units

Speech signals corresponding to the basic units are collected using an interactive speech digitizer cum editor package with provision to display, edit, save and playback the digitized data [3]. The data is collected using carrier words containing the basic units. The carrier words chosen are mostly nonsense words in order to quickly form words containing the required basic units. The carrier words are selected such that the effects of coarticulation between the adjacent character sounds are minimal. This is necessary from the point of view of incorporating prosodic features during synthesis. For this the desired basic unit is followed by an unvoiced and an unaspirated stop in order to avoid the effect of the following consonant. The duration of the extracted basic 'units can be considered as their base' durations. The base duration is the starting point for application of durational rules during speech synthesis.

Based on the points mentioned above, the guidelines for selecting the carrier words to extract the various basic units in our implementation of the text-to-speech system are as follows. Each carrier word contains 3 characters and has the form C_1 C_2 C_3. For C and CV units, C_1 is any character, C_2 is the desired basic unit and C_3 is a stop consonant. For example, ka ma: t (कमात) is a carrier word for ma: (मा). For V units, C_1 is the desired basic unit, C_2 is the stop character, C_3 is any character. In Hindi, a stand alone vowel can appear in the word beginning position or in a few cases in the word final position, but very rarely in a word medial position. The vowel extracted from word beginning is used for word final as well as word medial units also. This is because a stand alone vowel appearing in word final position merges with the preceding vowel and hence is difficult to extract. For example, au ka t (औकट) is the carrier word for au(औ). There are some exceptions for the basic units involving y(य), r(र), h(ह), and for some of the case markers and clause connectors [2].

2.3 Extraction of parameters

The speech digitized at 10 KHz sampling rate, is pre-emphasized before extracting the parameters (i) pitch, (ii) gain, and (iii) Linear Prediction (LP) coefficients. A 256-sample analysis frame is used for extraction of the parameters. Parameters are extracted for every 64 samples. The pitch is extracted using SIFT (Simplified Inverse Filter Tracking) algorithm [5] and the pitch contour is hand edited wherever necessary. The gain and LP coefficients (14th order) are computed using autocorrelation method [4]. The gain contour is also hand edited wherever necessary.

2.4 Extraction of basic units from input text

The input to the text-to-speech system is a text in the Indian language, Hindi. The text is stored in the form of ISSCII (Indian Script Standard Code for Information Interchange) codes [6]. A preprocessor scans the string of ISSCII codes to locate the abbreviations, numbers, dates, and special symbols and replaces them by their 'spoken forms'. The basic units are extracted from the expanded text using a simple parser [3]. To handle a few exceptions pertaining to 'inherent vowel

suppression' (specific to Hindi only), some rules have been used in this parser. For example, the text, bha: ra t ha ma: ra: de s ha i (भारत हमारा देश है) is parsed to obtain the following basic units: bha:(भा), ra (र), t (त्), blank, ha(ह), ma: (मा), ra: (रा), blank, de (दे), s (श्), blank, and hai (है).

2.5 Synthesis of speech from the basic units

The LP coefficients, pitch and gain contours corresponding to the basic units in the input text are concatenated. The pitch and the gain contours are smoothened at the boundaries between the basic units using three point median smoothing technique. The smoothed pitch and gain contours are then used to generate the excitation signal. For voiced frames the excitation is a periodic signal, while for an unvoiced frame the excitation is assumed to be white noise [5]. The excitation signal and the LPCs are used to generate the speech waveform while taking into account the **durational rules**. These durational rules modify the base duration of each basic unit, depending upon its position and context in the given text to obtain its actual duration which is then used in synthesis. The speech data is fed through a D/A converter to obtain speech output.

3. SIGNIFICANCE OF DURATIONAL KNOWLEDGE FOR SPEECH SYNTHESIS IN HINDI

The present system has the flexibility to change pitch or duration at the basic unit level during synthesis. This flexibility is useful in incorporating rules pertaining to intonation (pitch) and rhythm (duration). We have studied the behaviour of durations of basic units in different contexts From these studies, rules are being formulated to modify the duration of the basic units during synthesis depending on the context in the input text. These studies are presented in this section.

Some durational cues can be obtained from similar research done in other languages since some prosodic features are claimed to be cross linguistic in nature [7,8]. The extent of prosody in Indian languages have been studied by some phoneticians and their statistical results might help to get some clues [9,10]. For the sake of simplicity, the various durational effects can be roughly divided into the following categories (the words enclosed in square brackets below for example, SCL, POS etc will be used to refer to the corresponding effects, later in this section)

(1)[**POS**] A character is longer in word final position than in word beginning position, which in turn is longer than in a word medial position.
(2) [**PPL**] The vocalic duration of the character appearing before a pause is increased. The pause may be due to a phrase boundary or due to a 'breath group' or due to a sentence ending.
(3) [**PVC**] The effect of post vocalic consonant (PVC) on the preceding vowel occurs along various dimensions. The voicing, aspiration, 'sonority', nasality of the PVC, and the position of a syllable boundary with respect to the PVC affect the duration of the preceding vowel. Some specific PVCs also affect the duration of the vowel
(4) [**POA**] If two adjacent characters within and across word boundaries have the same place of articulation (POA), then one or both of the characters are shortened.
(5) [**NOV**] If an unusual word appears in the passage, it is spoken slowly and hence all the characters in the word will have a longer duration than otherwise.
(6) [**SCL**] In case of a cluster character (CCV or CCCV), the durations of the various constituent basic units are reduced.
(7) [**PSS**] If the number of characters in a word is greater than three, then the vocalic durations of the various characters are reduced.

The base duration for each basic unit is taken from the carrier word where it occurs in the word medial position. Let it be denoted as D_i. Since the carrier word has been spoken in isolation, the actual base duration in continuous speech will be less than D_i. It is found that the base duration of a basic unit in continuous speech is about 70% of the corresponding D_i. Therefore, the actual base duration for each basic unit is taken to be 70% of D_i. Let this base duration be denoted as D. Now, depending upon the context of each basic unit, various rules may be applied. Each of these rules modify the duration D for that particular basic unit using Eq. 1.

$$D_s = D + (\alpha * D)/100 \qquad (1)$$

where α is specified in each rule. If more than one rule applies, the rules will combine multiplicatively. After application of all durational rules, a value of α is determined for a particular basic unit. The base duration is then modified by Eq. 1 to obtain the value of duration D_s to be used during synthesis of that unit. The rules obtained so far are mentioned below, under each category

CATEGORY 1: POS effect

RULE 1 Lengthening of word final basic units : For each of the 25 pairs of nonsense words (uttered in isolation), the duration of different basic units in word medial and word final positions are compared. The average percentage increase in the word final position over the word medial position is taken as the value of α. This was verified in case of continuous speech and the value of α was found to be +35.

RULE 2 Lengthening of word beginning basic units: For each of the 30 pairs of nonsense words (uttered in isolation), the duration of different basic units in word beginning and word medial positions are compared. The average percentage increase in the word beginning position over the word medial position is taken as the value of α. This was verified in case of continuous speech and the value of α was found to be +10.

CATEGORY 2: PPL Effect

There is an increase in the final syllable of the word just before a pause. If the last character of the syllable has a vowel then the increase is only in the last character, otherwise the increase is in the vocalic portion of the penultimate character. For the following discussion, assume D_{v1} to be the duration of the basic unit, when the following character is unvoiced and unaspirated. D_{v2} in each case below, is the duration of the appropriate basic unit (as discussed earlier). The value of α is determined as the percentage increase of D_{v2} over D_{v1}. The sample examined for this rule were 30 word pairs.

RULE 3 Phrase boundary : If the pause is due to a phrase boundary, then α is 30.

RULE 4 'Breath group' : If the pause occurs at the end of a breath group, then α is 35.

RULE 5 Sentence ending : If the pause is due to a sentence ending, then α is 40.

CATEGORY 3: PVC Effect

This effect was examined for the vowel a:(आ). The basic unit is a CV combination or a stand alone

vowel. The CV combinations examined for this rule were chosen carefully, so that they cover different combinations of manner of articulation and place of articulation. The various CV combinations examined for this rule were ba:(बा),ka:(का),ta:(ता), gha:(घा),ṭa:(टा),na:(ना),ra:(रा), and sa:(सा). For each CV combination, a number of effects were examined. For the following discussion, assume D_{v1} to be the duration of the vocalic portion of the basic unit (which is one of the above CV combinations), when the PVC is unvoiced and unaspirated. D_{v2} is defined for each rule below, and the value of $\propto$ is obtained in each case as the percentage increase of D_{v2} over D_{v1}.

RULE 6 When the PVC is a voiced stop : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is a voiced stop. The value of $\propto$ was found to be +15. Here, the place of articulation of the PVCs in each word pair must be the same to avoid the influence of other factors. The sample examined for this rule were 24 word pairs. For basic units involving the consonant r(र्), the value of $\propto$ is found to be +11.

RULE 7 When the PVC is an aspirated stop : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is an aspirated stop. The value of $\propto$ was found to be +8.Here, the place of articulation of the PVCs in each word pair must be the same to avoid the influence of other factors. The sample examined for this rule were 24 word pairs. For basic units involving the consonant r(र्), the value of $\propto$ is found to be +5.

RULE 8 When the PVC is the trill, r(र्) : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is r(र्). The value of $\propto$ was found to be +30. The sample examined for this rule were 8 word pairs.

RULE 9 When the PVC is the fricative, h(ह्) : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is h(ह्). The value of $\propto$ was found to be -25. The shortening can also be attributed to the [POA] effect, whereby the durations of the basic units with same POA (glottal in this case) get reduced. The sample examined for this rule were 8 word pairs.

RULE 10 When the PVC is a nasal : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is a nasal n(न्) or m(म्). The value of $\propto$ was found to be -8. The sample examined for this rule were 8 word pairs.

RULE 11 When the PVC is a semivowel : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is y(य्) or v(व्). The value of $\propto$ was found to be +10 for y(य्) and +15 for v(व्). The reason for lengthening can also be attributed to [SCL] effect, as it is difficult to pronounce a semivowel after a basic unit involving a:(आ). The sample examined for this rule were 16 word pairs.

RULE 12 When the PVC is followed by a syllable boundary : Here, D_{v2} is the vocalic duration of the basic unit when the PVC is followed by a syllable boundary. The value of $\propto$ was found to be +10. The sample examined for this rule were 24 word pairs.

Exceptions to PVC Effect:

(i) If C_2 belongs to the 'breathy voiced' category (both voiced and aspirated) i.e., bha(भ), dha(ध), jha(झ), dda(ढ), and gha(घ), then the voicing dominates over aspiration and therefore the effect due to voicing only is valid.

(ii) It does not apply across word boundaries. In other words this category does not apply for basic units in the word final position.

These durational rules have been incorporated in the present text-to-speech system. The quality of output speech has improved with the addition of these rules.

4. CONCLUSIONS

In this paper we have discussed the issues involved in the design and development of a speech synthesis system for the Indian language, Hindi. This speech synthesis system was based on the parameter concatenation (Linear Predictive) model to have the flexibility to incorporate prosodic features. The characters of Hindi have been chosen as the basic units. A framework for collection of the basic units has been evolved, keeping in mind the flexibility needed to incorporate prosodic rules. The importance of incorporating durational knowledge for speech synthesis in Hindi has been demonstrated. In order to improve the quality of speech produced from input text, it is necessary to expand the set of rules to take into account the many variations of durations of segments that occur in natural continuous speech in Hindi.

REFERENCES

[1] I. Lehiste, Suprasegmentals, The MIT Press, Cambridge, 1970, ch. 1, pp. 1-5.

[2] S.R.Rajesh Kumar, R.Sriram and B.Yegnanarayana, 'A New Approach to develop a text-to-speech conversion system for Indian languages', presented at the Regional Workshop on Computer Processing of Asian Languages, Bangkok, Sep,1989.

[3] S. Srikanth, S .R. Rajesh Kumar, R. Sundar and B. Yegnanarayana, 'A text-to-speech conversion system for Indian languages based on waveform concatenation model', Technical Report No. 11, Project VOIS, Dept. of Computer Science and Engineering, I.I.T Madras, March 1989.

[4] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall, Englewood Cliffs, N.J, 1979.

[5] P.E. Papamichalis, Practical Approaches to Speech Coding, Englewood Cliffs, N.J, 1987.

[6] S.P. Mudur, L.S. Wakankar, and P.M. Ghosh, 'Text composition in Devanagari', SESAME bulletin : language automation worldwide, Vol 1, parts 1 and 2, pp. 18-27, 1986.

[7] J. Vaissiere, 'Language independent prosodic features', Prosody - models and measurements, Eds A. Cutler and D.R. Ladd, Springer Verlag, 1983, pp. 53-66.

[8] D.H.Klatt, 'Linguistic uses of segmental duration in English : acoustic and perceptual evidence', JASA, pp. 1208-1221, May 1976.

[9] K. Nagamma Reddy, 'The duration of Telugu speech sounds : an acoustic study', Special issue of JIETE on Speech Processing, pp. 57 - 63, 1988.

[10] S. R. Savithri, 'Durational analysis of Kannada vowels', JASI, pp. 34 - 40, April 1986.