Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 60 – 67

Procedia
Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Duration Modelling Using Neural Networks for Hindi TTS System Considering Position of Syllable in a Word

Shreekanth.T[a,*] , Udayashankara.V[b], Chandrika.M[c]

[a]JSSRF, SJCE Campus, Department of ECE, SJCE, Mysore,570006, India
[b] Department of IT, SJCE, Mysore,570006, India
[c] Department of ECE, SJCE, Mysore,570006, India

Abstract

The main criterion in duration modeling is to model the duration pattern of the natural speech, considering various features that affect the pattern. Proper estimation of segmental durations plays a vital role in natural sounding text-to-speech (TTS) synthesis. The primary reason for choosing the syllable as a basic unit is that the Indian languages are syllable centered. This paper presents a novel text processing and a syllable based data driven modelling of segmental duration for Hindi, using feed forward neural networks. The effectiveness of the system is demonstrated by synthesizing natural sounding speech for Hindi, national language of India.

Keywords: Duration; Neural Networks; Hindi; Mean Opinion Score (MOS); PRAAT; TTS; UNICODE.

1. Introduction

The TTS system is intended to convert an arbitrary input text to corresponding natural sounding speech in a more intelligible way. The two main components of a TTS system are text processing and speech generation. The function of the text processing component is to generate appropriate sequence of phonemic units.

* Corresponding author. Tel.: +91-998-648-3968.
E-mail address: speak2shree@gmail.com

doi:10.1016/j.procs.2015.01.056

These phonemic units are realized by the speech generation component either by synthesis from parameters or by selection of units from a large speech corpus. For natural sounding speech synthesis, it is essential that the text processing component produces an appropriate sequence of phonemic units corresponding to an arbitrary input text[1].

Computation of correct segmental durations is vital for natural sounding text-to-speech (TTS) synthesis. Variation in segmental duration is a key to the identity of a speech sound and helps to segment a continuous flow of sounds into words and phrases thereby increasing the naturalness and intelligibility[2].

The duration models are generally grouped into rule-based models and corpus based. The main difference between rule-based and statistical models is that a rule-based model can be built on relatively less speech data. Rule based methods involve manual analysis of segment durations. The derived rules get better in terms of accuracy, as the amount of speech data for analysis is increased. But with large amount of data the process of manually deriving the rules becomes tedious and time consuming. Hence, rule-based methods are limited to small amount of data. Also, this method depends on linguistic and phonetic literature about the factors that affect duration of the units (segments, syllables or phones). The complex interaction among the linguistic features at various levels makes the rule based methods more difficult to analyse and implement. Statistical data-driven methods are attractive when compared to rule based methods. This method works when large phonetically rich sentences are present in the corpora and is based on either parametric or non-parametric model that uses probability or likelihood functions[2].

One of the early attempts for developing rule-based duration models was in 1970s[4]. The model was based on information present in linguistic and phonetic literature about different factors affecting segmental durations. The rules were derived by analysing a set of phonetically balanced sentences. Following this model, similar models were developed for other languages like French[5] and Mandarin[6]. Of late, a corpus based duration model has been developed for Hindi (the Indian national language) TTS system[2, 3].

This paper intends to design a duration modelled Hindi TTS system considering position of syllable in a word using data-driven method. The syllable is used as a unit in this work as it captures the co-articulation effects and it is also a convenient unit for speech in Indian languages. The syllable used is of the form V and CV. This paper is organized as follows: The section 2 provides the information about the Hindi language. Section 3 describes the developed duration and speech database. Section 4 presents the Methodology. Section 5 depicts the results and discussion, and as a final point section 6 concludes and remarks about some of the future aspects.

2. Overview of Hindi Language

Hindi, the national language of India, spoken by 33 percent of the population has 33 consonants and 13 vowels. Hindi language is having one to one correspondence with the spoken language and the written form. The phonemes are divided into two types: vowels (swaras) and consonants (vyanjanas). Vowels (Swaras): Vowels are the independently existing letters which are called swaras and the sound of Vowels cannot be modified. Consonants (Vyanjanas): Consonants are those which depend on vowels to form their independent letter. Consonants sound can be modified by combining vowels with consonants[7]. For this reason Hindi language is phonemic in nature. Amalgamation of vowels with consonants will form a syllable and it is also called as "Baraha Khadi".

3. Duration Database and Speech Database Building

During the process of speech synthesis the required speech units are fetched from the database, concatenated and further processed using a suitable algorithm. Hence creating an error free database considering syllable as a basic unit is of greatest importance.

In order to carry out this task, a set of about 1540 words were collected from standard Hindi to English dictionary[11]. Later speech recording was done using the utility software for windows operating system called Praat[10]. The syllables were recorded with a sampling frequency of 16 kHz and represented using 16-bits. The pitch and formant frequency of the syllable fluctuate with position of the syllable in uttered speech. Consequently the syllable level speech database is generated for all the possible position of occurrences. The syllable can befall at three possible positions.

- Beginning of the word (Start)