A System for Analysis of Large Scale Speech Data for the Development of Rules of Intonation for Speech Synthesis

Asoke Kumar Datta$^1$ and Arup Saha$^2$

$^1$ BOM Public Charitable Trust
3/3 Girish Ghose Street, Kolkata 700035
dattashoke@yahoo.com
$^2$ C-DAC, Kolkata, Salt Lake, Kolkata
arupmtech@gmail.com

Abstract. For providing naturalness in synthesized speech it is imperative to give appropriate intonation on the synthesized sentences. The problem is not with synthesis engines but with the fact that comprehensive intonation rules of natural intonation are not available for any of the major spoken language of India. The knowledge available in this area is primarily subjective with the risk of unintentional personal bias. It lacks plurality in the sense that these do not reflect the natural intonation of common people. It is imperative to derive intonation rules through analysis of large amount of sentences spoken by common people. Manual processing is time consuming and extremely cumbersome. The present paper describes briefly an automated approach for such a task. A pilot study on about 1000 complex and interrogative sentences spoken by five female and four male native speakers is presented. 93% accuracy is obtained for the desired objective.

Keywords: Concatenative speech synthesis, intonation, epoch synchronous overlap add.

1 Introduction

Intonation plays an important role in providing both intelligibility and naturalness in synthesised speech. In addition to its important function in communicating linguistic information concerning emphasis, sentence structure and discourse structure, it is also important in the transmission of paralinguistic and non-linguistic information such as speaker’s intention, emotion and idiosyncrasy [1]. Since the naturalness of the synthesized speech output of a text-to-speech (TTS) system depends predominantly on its intonation, it is necessary to construct models or rules, which will relate the native intonation of output speech and the linguistic content of the text [1]. For this it is necessary to analyze a large corpus of spoken sentences. For analysis of such a large data, which includes an examination of dynamic pattern of the pitch contour extracted from spoken sentences, correlating them with the syntax, parts of speech and sometimes even emphasis, is so massive that some short of automation appears to be very
welcome. Furthermore the pitch contour of natural speech contains small perturbation without any contribution to the perceived intonation. These need to be removed resulting in a much needed data compression.

One may note that SLP is quite different from NLP. Spoken language is a natural language quite different from the artificial textual language. Analysis in NLP is a top-down process and follows a well specified grammar. The grammar for spoken language, on the other hand, cannot be so well specified. It is so because unlike writing where one has to know the grammar beforehand, when one speaks one does not, generally, have to follow a specified grammar strictly. However it does not mean that speech does not have a grammar. It seems that the grammar of speech evolves through intercourse between the members of the dialect community. Such a natural grammar, as opposed to the artificial grammar formulated by wise men, is likely to be fuzzy in nature. In actual practice we do not have an incorrect spoken sentence. The syntax and order seem to be ill-defined because of the spontaneity of speech communication. Over and above the dimensions of acoustic phonetics (corresponding to the graphemes, the only dimension, in NLP) other dimensions like intonation (related to pitch), stress, focus (related to pitch, duration and loudness), rhythm (related primarily to time and duration) come into play. It makes formal language theoretic approach more complex as the symbols are n-tuples. To build a grammar for SLP one needs the bottom-up approach and for this analysis of large volume of free speech collected from a large number of speakers of the dialect is imperative. The need for automation comes in here.

The emphasis of the present study is not to find appropriate intonation and prosody rules for Bangla but on the development of an automatic procedure to extract relevant parameters for the purpose corresponding to appropriate linguistic segments like syllable, word, clause, phrase and sentences. However to see how the automation helps researcher to find rules some examples with Bangla read sentences is included. The result of such analysis with only average pitch contour along with the correlation of syllable marking reveals existence of prosodic groups and clause/phrase segmentation with about 93% accuracy. It also reveals characteristic signatures of statements as opposed to interrogation.

The present paper describes an algorithmic approach for studying basic intonation patterns of a large number of sentences with the objective of providing a rule base for speech synthesis. The emphasis is on developing a processing scheme for quickly analyzing large audio database from spoken sentences. The analysis process includes extraction of pitch, estimation of syllabic contours using linear or non-linear estimation, automatically selected based on the nature of the data and later associating them with different linguistic elements like emphasis, prosodic groups, etc. The average pitch value of a syllable instead of the actual pitch contour is important parameter for general analysis as it has been reported that the contour patterns in syllables are not cognitively relevant in cognition of intonation [2]. This also reduces avoidable microstructure in pitch contour and consequent complexity of analysis. This pilot study uses complex affirmative sentences and interrogative sentences from the corpus of C-DAC, Kolkata [3], as these are likely to reveal most of the general trends of intonation patterns involving clauses, phrases and prosodic groups. The corpus provides the wave files tagged in terms of phonemes, syllables and words. The number of different complex sentences selected for the purpose is 46. This is read by five female and four