Duration analysis and Modelling for text to speech synthesis system in Malayalam

The duration, pitch and intonation of speech segments vary dynamically during speech production, providing the rhythm or prosody of speech. Synthesis of good quality speech depends on how well the duration and intonation patterns are imposed on speech segments. The variation of prosodic parameters is due to complex interaction of different factors. Hence synthesis of natural sounding speech is the greatest challenge in a text to speech synthesis system. Durational variation is incorporated in text to speech synthesis systems using duration models, which predict the duration of individual segments to be used for synthesizing speech, by considering various factors affecting duration.

The duration patterns being language specific, detailed durational analysis on natural speech has to be performed for each language, based on which suitable models can be built. The objective of the thesis is to study the duration patterns of speech segments in Malayalam Language, and to derive models to predict duration. Duration analysis and modelling for Malayalam has not been reported in any previous work, even though such work has been reported for world languages like English, German, French, Chinese and Japanese and Indian languages like Tamil, Hindi and Telugu. Transcribed database of Malayalam speech including 10 Malayalam news bulletins of Doordarsan TV Channel (Doordarshan is the public television broadcaster of India) was set up for analysis and modelling. Along with developing models for Malayalam speech, investigations were primarily oriented so as to develop novel duration models, which can be adapted for other languages as well.

The durational patterns of the phonemes in Malayalam news was analyzed using statistical tools like boxplots, distribution fitting, Q-Q plots, multiple comparison, confidence interval and ANOVA. The duration values were statistically analyzed to find the significant factors and to fix the number of levels for each factor. Feature vector is formed based on the analysis, combining the different factors. The feature vector framed is used for training the duration models developed. The duration models predict duration of a speech segment, when the feature vector corresponding to it given as input. The probability distribution of duration values was investigated by fitting it into different distribution functions (normal, lognormal, gamma and Weibull) and comparing the distribution plots using QQ

plots. It has been observed that gamma distribution is the best fit for probability density function of duration values for the corpus taken.

The first part of the work on duration modelling includes two models, i) probabilistic model and ii) hybrid model combining CART and HMM, which are new in the literature. The probabilistic duration model was developed based on the conditional probability distribution of the phonemes, given different factor values. The predicted duration is that which gives maximum value for the conditional probability, given the feature vector. The hybrid duration model is based on CART and HMM. It combines the advantages of the non linear regression model CART and the stochastic model based on HMM. The CART predicts the duration by regressing the values, corresponding to the given factor values. The HMM models the deviation of the values from that predicted by CART. The CART predicts the durational variation due to factors that can be derived from the given text, whereas HMM captures the variation due to complex interaction of factors. These models follow the conventional approach of considering each speech unit separately, analyzing the factors affecting duration and framing models based on this.

The second part of modeling, introduces a new model named memory based duration model, which differs from the traditional models, in terms of the basic approach in modeling. The basic premise on which model is built, is the observation that sentences, phrases and words have rhythm as a whole. The rhythm is not imposed on each phoneme or syllable. The memory based model is inspired by exemplar theory in phonology, memory prediction frame work depicting human intelligence, analogical modeling explaining language acquisition and Zipf's law dealing with nonuniform distribution of words in natural utterances . The model predicts the complex durational patterns of natural speech by mimicking the way, human brain captures and reproduces durational variation of speech segments.

In traditional models, it is assumed that, speech is a linear combination of segments occurring one after other and that the duration patterns can be predicted using factors determined from text. But in continuous speech, the speech segments interact with each other and hence the prediction of duration patterns is a complex nonlinear task. Moreover the rhythm or prosody varies with different speaking styles, which means that for the same text the durational patterns may be different for different styles. That is textual information is not sufficient to produce natural

sounding speech. Hence it is very difficult for duration models to capture the duration patterns of a language for a particular style, beyond a certain limit. At the same time, human brain is able to generate any speech construct in a particular style, if we have sufficient exposure to it.

The durational patterns can be captured, if we mimic the way human brain acquires different speaking styles. The new duration model is developed, using the basic concepts of memory prediction framework and exemplar theory. The memory prediction framework explains the hierarchal structure of human brain and cognition. The human neocortex has a hierarchal structure where each layer stores different speech constructs. The memory prediction framework is expected to be the direction in which speech processing research will proceed in the coming decades. The exemplar theory is a psychological model of perception and categorization. According to exemplar theory, the human brain perceives and reproduces speech constructs, by retrieving the stored exemplars of previously perceived examples. A new speech construct, which is not already stored, will be produced analogous to the most similar exemplar stored in memory.

The model is further supported by analogical theory and Zipf′s law from phonology. The analogical theory explains how we create new utterances, based on what we have already perceived and cognized. Zipf′s law deals with the nonuniform distribution of words and phrases in natural languages. These two theories, suggest that the rhythm of any particular style can be captured, by storing the prosodic parameters of most commonly occurring phrases and words. The new duration model captures the duration patterns by storing the durational variation of most commonly occurring syllables, morphemes, words and phrases. The model predicts duration by retrieving a stored pattern for each given speech construct. The memory prediction framework or exemplar theory has not been previously used for duration modeling.

The performance of the duration models are evaluated objectively and perceptually. The different models are compared objectively by calculating the root mean squared error (RMSE) and correlation. The mean opinion score is used for perceptual evaluation of the model. The memory based model gave best results on evaluation.