A PROBABILISTIC APPROACH TO SELECT UNITS BASED ON ACOUSTIC SIMILARITY FOR SPEECH SYNTHESIS

A THESIS

submitted by

ANJANA BABU

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

MARCH 2015

To my parents,

sister and brother

DECLARATION

I hereby declare that the entire work embodied in this thesis is the result of investigations carried out by me in the **School of Computing and Electrical Engineering**, Indian Institute of Technology Mandi, under the supervision of **Dr. Anil Kumar Sao**, and that it has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgements have been made wherever the work described is based on finding of other investigators.

Mandi, 175 001 **Anjana Babu**

THESIS CERTIFICATE

This is to certify that the thesis titled **A Probabilistic Approach to Select Units based on Acoustic Similarity for Speech Synthesis**, submitted by **Anjana Babu**, to the Indian Institute of Technology, Mandi, for the award of the degree of **Master of Science** (by Research), is a bonafide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Mandi, 175 001

Dr. Anil Kumar Sao
(Guide)

Acknowledgements

I would like to express my sincere thanks to my guide Dr. Anil K Sao for the invaluable guidance and help provided. Also, I would like to extend my gratitude to Prof. Hema A. Murthy for guiding me in my work extensively. Most of the research mentioned in this thesis was done at IIT Madras in collaboration with Raghava Krishnan K in Don Lab. So I would like express my gratitude to Raghava Krishnan K and other members of the lab who constantly aided in the work as well as encouraged wholeheartedly.

Most of the work is carried out as part of the consortium project *Development of Text to Speech Systems for Indian Languages* (11(7)2011-HCC(TDIL)). I would like to thank all the consortium members for sharing the data for research and giving valuable suggestions for research. Also I would like to thank all those who are/were part of the TTS project team in IIT Mandi, for the constant help in my research work. Special thanks to Priyanka Kaushal who helped me with some of the experiments and working hand in hand for many experiments on TTS. Also, I would like to thank research scholars Shejin T., Srimanta Mandal, Pulkit Sharma, Vinayak Abrol, Sarvesh Jayaraman and others for their help in research. I would like to thank research scholars and faculty members in the Multimedia Analytics and Systems group for the weekly research meetings, which have been a great platform to discuss research ideas, expand knowledge. Also, I am happy for the enthusiasm shown by the research scholars and students at IIT Mandi for participating in the evaluation of TTS systems. I am grateful to my Academic Progress Committee members for assessing my progress and keeping me on track. Last but not the least, I would like to thank my dear friends for being with me in times of joy and sorrow.

Anjana Babu

ABSTRACT

Keywords: *TTS, Unit selection speech synthesis, Prosody.*

One of the major challenges in Text-to-Speech Synthesis Systems (TTS) is the incorporation of prosody in the synthesised speech. Various techniques based on linguistic characteristics have been proposed in the literature for improving prosody. However, prosody in TTS is still an open problem and this is addressed in this thesis.

In this work, approaches are proposed for improving the naturalness, intelligibility and prosody. It is performed by selecting the sequence of sound units from a large corpus in such a way that acoustic features are made consistent at (a) segmental level, and (b) supra-segmental level. At segmental level, the differences in acoustic features of sound units are reduced over the entire utterance to improve naturalness and intelligibility. At supra-segmental level, units are selected by ensuring the consistency in the differences in acoustic features of adjacent syllables at phrase level. In this method, consistency of acoustic features is also maintained at utterance level. Unlike the existing USS based TTS, which rely mostly on linguistic information for improving prosody, the proposed approach makes use of acoustic information. Probabilistic approaches are proposed for selecting units based on an acoustic framework. Since the context is only specified by acoustic features, the proposed approaches can be applied to any language and perhaps even for multilingual synthesis. The experimental results of the proposed approaches are demonstrated using five Indian languages. It was observed from the subjective evaluation tests that f_0 contributed to the naturalness of the systems whereas duration and energy helped in improving the intelligibility of the systems. Also, ensuring consistency of energy and f_0 across syllables in phrase and duration across syllables in utterance further improved the prosody.

Contents