Universitat Ramon Llull

Ph. D. Thesis

Multi-domain Unit selection based Text-to-Speech Synthesis with subjective weight tuning and robust pitch marking

Author **Francesc Alías Pujol**

School **Enginyeria i Arquitectura La Salle**

Department **Communications and Signal Theory**

Supervisor **Dr. Joan Claudi Socoró Carrié**

July, 2006

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es

Abstract

The final purpose of any Text-to-Speech (TTS) system is the generation of perfectly natural synthetic speech from any input text. Historically, two strategies have been followed in the quest for this goal: the general purpose TTS synthesis (GP-TTS), which strives the flexibility of the application at the expense of the achieved synthetic speech quality; and the limited domain TTS synthesis (LD-TTS), which prioritizes the development of high quality TTS systems by restricting the scope of the input text. At present, the most used strategy to develop TTS systems is the so called corpus-based text-to-speech or unit selection TTS (US-TTS) synthesis. Although the quality of US-TTS synthesis systems is quite good in general, there are still several open issues which are still being investigated.

This PhD thesis introduces different contributions for US-TTS systems in order to improve, by one hand, the naturalness of GP-TTS systems, and by the other hand, the flexibility of LD-TTS systems. To deal with the former problem, a new technique for efficiently incorporating human perception in the unit selection process by means of subjective weight tuning is introduced, which also allows controlling user fatigue and user consistency. Moreover, a new method for improving the reliability of automatic speech corpus labelling is described, particularly, a generic pitch marks filtering algorithm is introduced —an essential issue in corpus-based TTS systems. Moreover, the latter problem is addressed by multi-domain TTS (MD-TTS) synthesis, following the LD-TTS approach, which deals with achieving synthetic speech quality equivalent to that of LD-TTS systems, but improving TTS flexibility by considering different domains (speaking styles, emotions, topics, etc.) for conducting speech synthesis. In this context, the MD-TTS system needs to know, at run time, which domain or domains are the most suitable for synthesizing the input text with the highest synthetic speech quality. To that effect, the MD-TTS system incorporates a text classification module to classic TTS synthesis architecture adapted to the MD-TTS classification particularities. Finally, all the proposals are evaluated in terms of objective experiments —by means of classic or new measures— and/or subjective tests —perceptual tests— in order to validate the improvements achieved by the methods developed in the US-TTS framework, as a step further in our research towards developing high quality and flexible text-to-speech synthesis systems.