

Demo System for NU-MBROLA concatenator

Baris Bozkurt, Michel Bagein, Thierry Dutoit

Multitel-TCTS Lab,
Faculté Polytechnique de Mons, Belgium
{bozkurt,bagein,dutoit}@tcts.fpms.ac.be

Abstract

A simple demo system is prepared for demonstrating the quality of NU-MBROLA concatenator which is able to produce speech from text data. The system is composed of three modules. The first module is the NLP part of EULER system [1] which produces list of phonemes and target prosody from given text. The second module is a non-uniform unit selector and the last module is the NU-MBROLA concatenator. The system is not a full TTS demo, it is prepared just to demonstrate NU-MBROLA concatenation.

1. The EULER Module

The first module in our demo system is the NLP part of EULER (fig.1). EULER(<http://tcts.fpms.ac.be/synthesis/euler>) is a collaborative project aimed at obtaining a set of highly modular TTS synthesizers for as many languages as possible, free for use in non-commercial and non-military applications. It is an open project: many of its components are provided as GNU C++ sources.

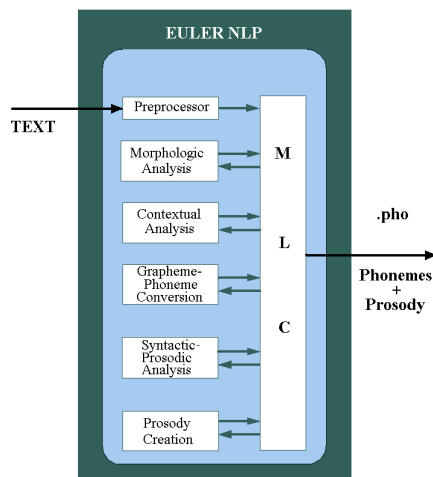


Figure 1. EULER NLP

The backbone of the Euler project is a C++ generic data structure called Multi Layer Container (MLC) which recall the MLDS of Festival [2] and Speech Maker [3]. The MLC is an extension of the C++ Standard Template Library to multi-level data with links across levels (Its code is distributed under the GPL licence).

Our demo system is prepared by using tools of EULER for French. EULER tools for other languages can be found on the project page.

2. The Nu-selection Module

The second module is a simple non-uniform units selector which is used in half-phoneme units mode for this demo system. The module produces a list of half-phoneme units and their prosody from given target phoneme and prosody specifications by performing a search on a speech database segmented in half-phonemes. A phoneme is assumed to be composed of two half-phonemes with equal durations.

The selection system first finds the list of candidates regarding linguistic context. Target and concatenation costs are calculated for the candidates and then a Viterbi search finds the best path to obtain the list of units to be concatenated. No weight training is performed, all the measures have the same weight. Target costs are calculated regarding the linguistic context, average f0 and duration of the units. Concatenation costs are calculated from f0 values at concatenation boundaries. Target prosody is imposed on the units without taking account of actual prosody of units, it is directly taken from prosody generated by the Euler module.

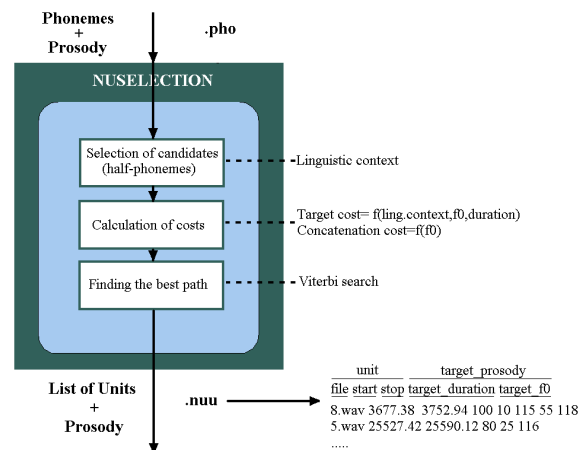


Figure 2. Unit selection module

3. NU-MBROLA Concatenator

The last module of the system is NU-MBROLA. The NU-MBROLA system synthesizes and concatenates units with the standard MBROLA algorithm [4]. The only difference is that, it is no more restricted to use of diphones as units. Units are defined by the file names they exist and their position in the wave files, no other information is needed to define units. The

smallest unit that can be used with NU-MBROLA is an analysis frame length unit and there is no upper-limit for the size of units.

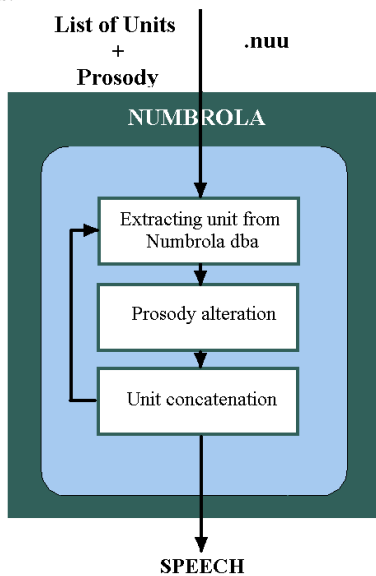


Figure 3. NU-MBROLA concatenator

NU-MBROLA utilizes a database of re-synthesized speech frames obtained by processing the speech corpus by a pitch asynchronous harmonic/noise analyzer and re-synthesizing speech with constant phase envelope and constant pitch. This operation is done only for the voiced frames, unvoiced frames are directly copied. No segmentation information is needed for this procedure.

During synthesis, NU-MBROLA maps the speech corpus unit definitions to NU-MBROLA database unit definitions. Frames are extracted and speech is synthesized by overlap adding these frames. The target prosody is imposed during this operation.

Smoothing is performed at concatenation boundaries as in MBROLA. To achieve real-time spectral smoothing, the operation is performed by distributing the difference boundary frames to neighbour stationary voiced frames in time domain. This operation is performed only when the units are non-consecutive.

4. Conclusions

A simple demo system for NU-MBROLA concatenator is presented. The quality of synthetic speech is limited with the simplicity of unit selection algorithm used and the target prosody imposed on the units. Actually, this system is half-phoneme based other than non-uniform units. Some other examples are presented on the project page of NU-MBROLA including speech synthesis with aligned prosody and direct re-synthesis of speech with original prosody. (<http://tcts.fpms.ac.be/synthesis/numbrola>)

5. References

[1] Dutoit, T., Bagein, M., Malfreire, F., Pagel, V., Ruelle, A., Tounsi, N., and D. Wynsberghe, "EULER : an Open,

Generic, Multi-lingual and Multi-Platform Text-To-Speech System" , *Proc. LREC'00, Athens, May 2000*, p. 563-566.

- [2] Black, A., Taylor, P. and Caley, R., The festival speech synthesis system: System Documentation, University of Edinburg, 1997.
- [3] Van Leeuwen, H.C. and Te Lindert, E., "Speech Maker: a flexible and general framework for text-to-speech synthesis and its application to Dutch", *Computer Speech and Language*, 1993, p 149-167.
- [4] Dutoit, T. and Leich, H. "Text-to-speech synthesis based on a MBE re-synthesis of segments database", *Speech Communication*, Vol.13, 1993, p 435-440.
- [5] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using a large speech database", *Proc. of ICASSP, Atlanta, Georgia, 1996*, p 373-376.