From MBROLA to NU-MBROLA

Baris Bozkurt, Michel Bagein, Thierry Dutoit

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

Abstract

We introduce the NU-MBROLA project as an extension of the MBROLA project for designing large NU-MBROLA databases exhibiting the same properties as MBROLA diphone databases, without the restriction of being composed of diphones. The accompanying NU-MBROLA synthesizer implements the MBROLA algorithm on speech segments only defined by their starting and ending points in natural speech files. It is distributed with the same terms and conditions as in the MBROLA project. The terms and conditions for the creation of NU-MBROLA database are slightly different from those related to the creation of MBROLA databases, mainly in that no speech segmentation is required, and in that we leave it to providers to distribute their NU-MBROLA databases.

1. Introduction

Speech synthesis based on the concatenation of units selected dynamically from a large speech corpus have undoubtedly made an important breakthrough in the past 5 years.

Sessions devoted to speech synthesis in many recent conferences reveal a global quest for unit selection systems providing highly intelligible and natural speech [1-5,19]. More recently, efforts have also been devoted to achieving this goal with manageable computational cost and storage requirements [6,16]. Phonologically grounded [7,17] or acoustically grounded [8], all such selection algorithms also need a synthesis engine, if only for properly concatenating speech segments. In practice, such engines range from almost nothing (as originally proposed by Nick Campbell [9]) to frequency-domain synthesizers (as in proposals by the AT&T speech synthesis group [10]), through the widely used TD-PSOLA algorithm.

Each of these synthesis methods has advantages and drawbacks for corpus-based synthesis:

- Approaches which explicitly banish signal processing are undoubtedly optimal when concatenated units are acoustically very close; if this is not the case, they might not be able to smooth out pitch, phase, and spectral mismatches at unit boundaries. This is in contradiction with recent findings in design of speech coders, which show spectral modifications which preserve smoothness and continuity are better accepted by listeners than those which produce localized distortions [11]. Hence, chances are that such purely sequencing approaches will require larger speech corpora, in order to increase the chances to find units with limited boundary mismatches.
- PSOLA-based synthesis requires some form of pitch marking, which is even harder for large corpora than it was for diphone sets. Assuming this can be done consistently, these approaches will somehow eliminate phase and pitch mismatches at boundaries, but they will not easily handle spectral mismatches.
- Although frequency-domain synthesizers can be used for removing all mismatches at boundaries, this is achieved at the expense of substantial computational load. Additionally, frequency domain synthesizers can be somehow tricky to tune to specific voices. Their design requires specific expertise.

In this paper, we introduce the NU-MBROLA project (section 2) as an extension of the MBROLA [12] project for designing large NU-MBROLA databases (section 2.1) exhibiting the same properties as MBROLA diphone databases, without the restriction to being composed of diphones. The accompanying NU-MBROLA synthesizer (section 2.2) implements the MBROLA algorithm on speech segments only defined by their starting and ending points in speech files. It can therefore be used in many different synthesis paradigms (from diphone-based synthesis to the insertion of words in predefined recorded messages, through non-uniform-unit-based synthesis; section 3). It is distributed with the same terms and conditions (section 4) as in the MBROLA project. The terms and conditions for the creation of NU-MBROLA database are slightly different from those related to the creation of MBROLA databases, mainly in that no speech segmentation is required, and in that we leave it to providers to distribute their NU-MBROLA databases.

2. The NU-MBROLA Project

The NU-MBROLA project implies the collaborative design of NU-MBROLA databases, their distribution, and the distribution of the NU-MBROLA synthesis engine.

2.1. NU-MBROLA databases

Prior to synthesis, the original speech corpus is processed by a constant frame size and shift harmonic/noise analysis and re-synthesis to develop a NU-MBROLA database. The analyser estimates the harmonic parameters (amplitudes and phases of harmonics and the modelling errors), pitch frequency values, and makes voiced/unvoiced and stationary/transient decisions for each frame. The next step, to produce the NU-MBROLA database, is re-synthesizing voiced frames at constant pitch frequency and constant phase envelope (for the low frequency part of the speech spectrum) with a harmonic synthesizer and copying the unvoiced frames directly. To perform constant pitch re-synthesis while preserving the original vocal tract properties, harmonic

amplitudes are re-calculated by re-sampling their envelope spectrum at a constant pitch frequency. These analysis/re-synthesis operations do not require segmentation information of the original speech corpus.

2.2. NU-MBROLA synthesis

The NU-MBROLA synthesis engine performs synthesis with the standard MBROLA algorithm [12] but it is no longer limited to diphone-based synthesis. As opposed to the MBROLA databases and synthesis engine, which embody acoustic and phonetic information, both the NU-MBROLA databases and the NU-MBROLA engine are purely acoustic objects. Units could thus be words, syllables, phonemes, or any other type of speech segment). NU-MBROLA need not know about this. Examples of use are presented in section 3 on unit selector.

For synthesis, the user provides the list of speech segments to be concatenated and produced with some target prosody. Speech segments are defined by their location in the original speech corpus, with their starting and ending points in milliseconds. Target prosody must then be defined as in the MBROLA .pho file format : duration (in milliseconds) followed by optional pitch pattern point definitions (each point being defined by its position in percent in the target duration and target pitch value in Hz). An example of NU-MBROLA input format is:

```
snd01.wav 253 289 45 10 119 21 112
snd01.wav 123 189 56
snb09.wav 5078 5096 100 99 120
```

This will produce 201 ms of synthetic speech, using 3 units taken from 2 separate files, and applying a pitch movement that goes from 112 Hz (at 10% of the 45 ms of the first segment) to 120 Hz (at 99% of the .100 ms of the last segment). Intonation is linearly interpolated from pitch pattern points in a log scale.

During synthesis stage, speech segment descriptions (with references to the original speech files) are translated into NU-MBROLA segment descriptions and are extracted from the NU-MBROLA database. Synthesis is performed by the standard MBROLA algorithm imposing the target prosody features. Duration modification is uniform (i.e., the duration scaling factor is constant) throughout each speech segment.

Some smoothing can advantageously be performed to reduce the spectral differences at segment boundaries. Consecutive segments (regarding their location in the original speech corpus) naturally exhibit some spectral differences but these differences correspond to the natural evolution of speech and need to be preserved. Therefore, smoothing is only performed at stable and voiced boundaries of non-consecutive segments by distributing linearly the difference of boundaries frames in the right and left stationary frames by a fading/fadeout operation. Since the NU-MBROLA corpus is composed of constant length frames with constant phase envelopes (for low ordered harmonics), sample by sample subtraction of boundary frames provides the difference frame. Linear distribution of this difference frame corresponds to distributing the spectral difference linearly.

3. Using NU-MBROLA

For synthesising speech by using units from a corpus, a unit selector can be implemented to select the proper units for a good quality synthetic speech.

Then the unit selector needs to provide the list of units selected in the correct input format of NU-MBROLA.
In the context of non-uniform units, different format of units have been previously explored by researchers: word [13,18], syllable [14], phonemes, diphones, half-phonemes, etc. This issue is clearly still open. All those units can be defined as speech segments located by files with their boundaries.

We have implemented a unit-selector for testing different units types and some audio examples are available on this web page. http://tcts.fpms.ac.be/synthesis/mbrola/numbrola.html .
NU-MBROLA can also be used for other aims than TTS, i.e. for speech-to-speech re-synthesis to modify prosody.

4. Terms and Conditions

NU-MBROLA has been designed for TTS developers and it is intended to be a speech synthesizer to be used in mainly corpus based speech synthesis.

An important feature of NU-MBROLA is that it will be freely available for non-commercial and non-military applications. Since NU-MBROLA utilizes a NU-MBROLA database other than just recorded speech we provide the database processing service after a contractual agreement with the provider of a speech corpus. The speech corpus needs to be composed of only the speech recordings and no segmentation information is needed. The NU-MBROLA database obtained by processing the data will be supplied to its provider. We take the right to publish speech synthesis examples obtain by re-synthesizing a small part of corpus with its original prosody (This is indeed the best quality which can be expected from NU-MBROLA synthesis using this database). These files will be presented on a web page with additional pointers to provider who may distribute the NU-MBROLA database (or their full TTS system).

Although we encourage corpus providers to publish their NU-MBROLA databases on the web or via official corpora distribution bodies (ELRA or LDC), including the corresponding acoustic-phonetic information (i.e., text, phonetic segmentation, and intonation), so as to make it possible for other people to design NU selection systems based on identical speech corpora and synthesis technique, we do not impose this on corpus providers.

As an example, we have ourselves made available a large NU-MBROLA database for French (LeSoir97) on the official web site of NU-MBROLA, http://tcts.fpms.ac.be/synthesis/mbrola/numbrola.html .

5. Conclusions

NU-MBROLA can handle most of the mismatches at segment boundaries; it does not consume computational power; it allows important database compression ratios (typically 1/5

without noticeable effect on speech quality [15], and at small computational cost); and the NU-MBROLA project makes it possible for TTS developers to use these advantages without having to deal with the design of a complex frequency-domain synthesizer.

Examples given in an accompanying web-site tend to show that the speech degradation produced by the underlying algorithm is not important, provided segments are re-synthesized with synthetic prosody close (if not identical) to the prosody they had in the original speech corpus.

Our hope is that, given its smoothing capabilities, NU-MBROLA will make it possible to decrease the number of units required to produce a given speech quality. If this is the case, the slight distortions introduced could be seen as the price to pay for fast and compact non-uniform-unit-based synthesis.

We therefore believe that, although NU-MBROLA does not solve all the problems related to the use of signal processing in corpus-based speech synthesis, it has its place in the landscape of usable synthesis techniques.

6. References

[1] 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.*

[2] Balestri M., Paechiotti, A., Quazza, S., Salza, P. L., Sandri, S. "Choose the best to modify the least: a new generation concatenative synthesis system", *Proc. of EUROSPEECH, Budapest, Hungary, Sept. 1999.*

[3] Black, A., Taylor, P. "Automatically clustering similar units for unit selection in speech synthesis", *Proceedings of EUROSPEECH, Rhodes, Greece, Sept. 1997.*

[4] Donovan, R. Ed. "Segment preselection in decision tree based speech synthesis systems", *Proc. of ICASSP, Istanbul, Turkey, June, 2000.*

[5] Coorman, G., Fackrell, J., Rutten, P., Van Coile, B. "Segment selection in the L&H Realspeak laboratory TTS system", *Proc. of ICSLP, 2000.*

[6] Beutnagel, M., Mohri, M. and Riley, M., "Rapid unit selection from a large speech corpus for concatenative speech synthesis", *Proc. of EUROSPEECH, Budapest, Hungary, Sept. 1999.*

[7] Taylor, P. and Black, A.W. "Speech Synthesis by Phonological Structure Matching", *Proc. of EUROSPEECH, Budapest, Hungary, Sept. 1999.*

[8] Beutnagel, M., Conkie, A. and Schroeter, J., Stylianou, Y., and Syrdal, A. "The AT&T NextGen TTS system", *Proc. of the Joint Meeting of ASA, EAA and DAGA , Berlin, Germany, 1999.*

[9] Fujisawa, K., and Campbell, N. "Prosody based unit-selection for Japanese speech synthesis", *Proc. of 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov. 1998.*

[10] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Trans. Acous., Speech, Signal Processing, Vol.9, Jan.2001, p 21-29.*

[11] Hedelin P., Kang, H.-G and Eriksson, T. "Low-Rate Quantization of Spectrum Parameters," *Proc. IEEE ICASSP, Istanbul Turkey, June, 2000.*

[12] 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.*

[13] Chou, Tseng, Lee, "A Chinese TTS system based on part of speech analysis, prosodic modeling and non uniform units", *Proc. of ICASSP 1997, p 923-926.*

[14] Tanaka, K., Mizuno, H., Abe, M. and Nakajima, S. "A Japanese Test-to-speech system based on multi form units with consideration of frequency distribution in Japanese", *Proc. of EUROSPEECH, Budapest, Hungary, Sept. 1999, p 839-842.*

[15] Van Der Vreken, O., Pierret, N., Dutoit, T., Pagel, V., Malfrere, F. "A Simple and Efficient Algorithm for the Compression of MBROLA SegmentDatabases", *Proc.of ICSPAT, San Diego, pp. 241-245,1997.*

[16] Campbell, W.N. "Reducing the size of a speech corpora for concatenation waveform synthesis", *Technical Publications, ATR Interpreting Telecommunications Research Laboratories, p. 90-91.*

[17] Breen, A.P. and Jackson, P. "Non-unit selection and the similarity metric within BT's Laureate TTS system" *Proc. of Third ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, p.373-376,1998.*

[18] Stöber, K., Portele, T., Wagner, P., Hess, W. "Synthesis by word concatenation" *Proc. of the 6th European Conf. on Speech Communication and Technology, Budapest, Hungary, 1999, vol. II, p. 619-622.*

[19] Möbius, B. "Corpus-based speech synthesis: methods and challenges" *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), 87-116, 2000.*