THE MBROLA PROJECT: TOWARDS A SET OF HIGH QUALITY SPEECH SYNTHESIZERS FREE OF USE FOR NON COMMERCIAL PURPOSES

T. Dutoit, V. Pagel, N. Pierret, F. Bataille, O. van der Vrecken

Faculté Polytechnique de Mons, TCTS Lab, 31 Bvd Dolez, B-7000 Mons, Belgium
Phone: +32-65-374133
Fax: +32-65-374129
Email: mbrola@tcts.fpms.ac.be

ABSTRACT

The aim of the MBROLA project, recently initiated by the Faculté Polytechnique de Mons (Belgium), is to obtain a set of speech synthesizers for as many voices, languages and dialects as possible, free of use for non-commercial and non-military applications. The ultimate goal is to boost up academic research on speech synthesis, and particularly on prosody generation, known as one of the biggest challenges taken up by Text-to-Speech synthesizers for the years to come.

Central to the MBROLA project is MBROLA 2.00, a speech synthesizer based on the concatenation of diphones. Executable files of this synthesizer have been made freely available for many computers/operating systems, as well as a first diphone database for a French male voice.

We describe here the terms of participation to the project, as a user, as an associated developer, or as a database provider.

1. INTRODUCTION

Designing a high quality TTS system is an increasingly complex task, which typically requires the cooperation of various specialists. As a result of this multi-disciplinarity, however, it has become more and more difficult to achieve it with public funding only.

Hence this paradox, recently denoted by Cole et al. [1]: *"As for industry funding, the results are not generally in the public domain, and consequently speech research has suffered. In contrast to a decade ago, it is difficult to gain access to a state-of-the-art synthesis system that will allow full control of the parameters necessary for conducting speech research"*.

The MBROLA project, recently initiated by the TCTS Lab of the Faculté Polytechnique de Mons (Belgium), precisely aims at changing this situation by gathering a set of speech synthesizers free of use for non commercial purposes.

Central to this project is MBROLA 2.00, a speech synthesizer based on the concatenation of diphones. Diphone databases tailored to the MBROLA format are needed to run the synthesizer, and the MBROLA project is organized so as to incite other research labs or companies to share their diphone databases.

MBROLA 2.00 takes as an input a list of allophones associated with prosodic information (duration and piecewise linear pitch curve which implies no prosodic model) and outputs 16 bit linear speech samples. Hence, it is *not* a Text-to-Speech system.

A first French diphone database is provided with the software (FR1: a French male voice), as well as command files intended to demonstrate the quality of the resulting synthesis [SOUND A920S02.WAV] [SOUND A920S03.WAV]. As a result of the sharing policy of this project, we expect a growing number of voices and languages to become available soon.

We describe in the following paragraphs the advantages of the MBROLA algorithm, and the command file format to control the related program. Then, we explain how to join the project as a user, a program developer, or a database provider. We end by describing typical steps to follow when recording a diphone database intended for the MBROLA project.

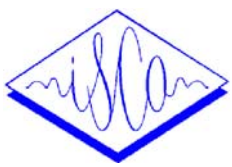
2. MBROLA ALGORITHM

The MBROLA 2.00 program uses a technique known as Multi Band Resynthesis OverLap Add [1] which produces speech by diphone (triphone or polyphone will be available in future versions) concatenation (for an introduction to concatenative approaches to TTS synthesis, refer to [2]).
Like the well-known PSOLA methods (TD-PSOLA[2] for Time Domain-Pitch Synchronous OverLap Add [3], or PIOLA [4], standing for Pitch Inflected Overlap Add, or MBR-PSOLA[5] standing for Multi Band Resynthesis Pitch Synchronous OverLap Add) it adds overlapping frames directly in the time domain. MBROLA shares MBR-PSOLA's ability to smooth spectral discontinuities in the time domain, which enhance the *fluidity* cue (comparison with other synthesis methods available in [6]), while allowing efficient database coding facilities.

As a result, MBROLA cumulates the flexibility and data compression ratio of parametric speech models while keeping the computational simplicity of time-domain synthesizers:

1. Computational complexity remains as low as an average 7 op-

[1] Patent pending accounting for Faculté Polytechnique de Mons
[2] PSOLA-TD^{TM} is a trademark of France Telecom



4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

```
; bonjour
_ 51 25 114
b 62
o  127 48 170
Z 110 53 116
u 211
R 150 50 91
_ 9
```

Figure 1: Bonjour.pho [SOUND A920S01.WAV]

eration per sample. For a 16Khz sampling rate this lead to real time synthesis on a Intel486 processor.

2. MBROLA is able to smooth spectral discontinuities arising at diphone junction points. As a result diphones (as opposed to, e.g., triphones) produce high quality synthetic speech.
3. Diphone database development is simplified thanks to spectral smoothing. There is no need for trial and error operation and successive refinements to eliminate "bad" sounding diphones which introduce the biggest spectral discontinuities. For example, the speech quality of the French FR1 diphone database accompanying MBROLA was directly obtained after recording and automatic spectral transformation to the MBROLA format.
4. At last, due to their peculiar nature, MBROLA diphone databases can be efficiently coded at the cost of a minor percentage of the computational load required for synthesis (databases can be coded with less than 40kbps, for an original sound at 16Khz, i.e. 256kbps for 16 bits linear coding).

Point three is the main feature which makes MBROLA a good candidate for gathering many diphone databases developed all over the world, as we wish to do in this project.

3. MBROLA 2.00 PROGRAM

MBROLA 2.00 takes as an input a file or a pipe giving the list of phonemes with some prosodic information (durations/pitch), and outputs an audio file or pipe containing 16 bit linear samples at the sample rate of the diphone database (16Khz for the French FR1 database). The audio output file format depends on the name extension. Supported format are Raw (no header), Wav (Riff-Wav Microsoft format), Au (Sun), Aiff (Macintosh, Silicon Graphics).

The command file format is quite simple, as in the example file *bonjour.pho* listed Fig. 1. Each line begins with the name of an allophone (in the example given , FR1 database uses the SAM Phonetic Alphabet) followed by a duration in milliseconds, and optionally followed by one or more pitch points.

Each pitch point is a pair of two values: relative position of the pitch point in percent of the phoneme duration, and pitch value in Hertz. So, the first line of *bonjour.pho* triggers the synthesis of a 51 ms long silence, and begins with a first pitch point at 144 Hz at 25 percent of 51 ms. Pitch points let the user draw a piecewise linear intonation curve regardless of the voiced or unvoiced nature of the segment. Voicing is actually coded in the database, and voiceless sounds are not modified by pitch points at synthesis time.

MBROLA 2.00 is *not* a demonstration program, it is a fully functional version of the MBROLA synthesizer, though it (still) has some limitations:

- Fundamental frequency is limited to a range of 2 octaves from the base F0 of the diphone database. Many synthesizers share this limitation since artifacts arise outside this interval.
- Although three pitch points per phone are enough to produce good quality speech, MBROLA accepts up to 20 pitch point which allows the reproduction of vibrato on singing voices.
- Maximum phone length depends on the fundamental frequency with which they are produced. The higher the frequency, the lower the duration. For a 133 Hz frequency, the maximum duration is 7.5 sec, for a frequency of 66.5 Hz, it is 15 sec, for a frequency of 266 Hz, it is 3.75 sec.
- Although pitch point are optional, the program does not accept more than 250 allophones without any pitch point.

Binaries are available for :

SUN/Solaris 2.4,
SUN4/ Sunos4,
HPUX9.0 and 10.0,
VAX/VMS 6.2,
DECALPHA(AXP)/VMS6.2,
PC/Dos6,
PC/Win3.1,
PC/Linux,
PC/Solaris2.4,
NEXT/Nextstep,
SGI/IRIX 5.3

Utilities accompanying the program let the user listen to the result on many architectures. When possible, using the pipe facility allows to enter command lines like:

mbrola fr1 bonjour.pho -.au | audioplay

which plays the command file given as an example Fig. 1.

4. HOW TO JOIN THE PROJECT

There are three types of participations to the MBROLA project: as a user, as an associated developer, or as a database provider.

4.1. Join as a user

The MBROLA 2.00 program can be freely used for non commercial and non military applications. When no charge is made, the program may be copied and distributed freely. In return, the authors ask users to mention the present paper in any scientific publication referring to work for which this program has been used. Those points are precisely described in a license file coming along with MBROLA.

A typical example of scientific use, published in these proceedings [7] is the *Prosodic Karaoke* program [SOUND A920S04.WAV] [SOUND A441S01.WAV], an application designed for speech perception experiments.

For convenience, we have defined a mailing list dedicated to users:

mbrola-interest@tcts.fpms.ac.be

This is a forum for MBROLA questions and issues. The maintainers of the mbrola project use it to announce new releases, bug fixes, new voices and languages, and other information of interest to all MBROLA users. Users who have a question, comment, think they have found a bug, or simply who want to share phonemic command files or free applications running on top of mbrola should send an e-mail to mbrola-interest. To register, simply send an e-mail to mbrola-interest-request@tcts.fpms.ac.be with the word "subscribe" in the subject.

4.2. Join as a developer

The MBROLA license allows the user to develop programs on top of the synthesizer in the framework of the MBROLA project. The availability of talking apps built on top of MBROLA (especially scientific tools for testing prosody and talking tools for handicapped persons, but also talking clocks, talking calendar, etc), is announced via the mbrola-interest mailing list.

We encourage developers to distribute these applications in the same way as MBROLA, so that the mbrola archives become a speech tool library available for free to the community.

Applications working on top of MBROLA may not be sold or incorporated into any product which is sold without prior permission from the author.

4.3. Join as a diphone database provider

One of the biggest interests of the MBROLA project (and definitely its most original aspect) lies in its ability to provide an ever growing set of languages/voices to users. To achieve this goal, the MBROLA project has been organized so as to incite other research labs or companies to share their diphone databases.

The terms of this sharing policy can be summarized as follows:

1. We shall only use the shared database to adapt it to the mbrola format, and destroy the copy when this is done.
2. The resulting mbrola diphone database will be copyright Faculté Polytechnique de Mons - T.DUTOIT. Non-commercial use of the database in the framework of the MBROLA project will be automatically granted to Internet users. In return, we send to the database provider a license agreement which transfers all our commercial rights on the newly created database to him, provided the database is used with and only with the MBROLA program.

The MBROLA project allows us to easily build diphone databases, since the spectral analysis-resynthesis on which the synthesizer depends is completely automatic. The next section gives advice to help users prepare a diphone database for its processing according to the MBROLA format.

5. CREATION OF A DIPHONE DATABASE

Creating a database is typically achieved in four steps : Creating a text corpus, Recording the corpus, Segmenting the speech corpus, Equalizing diphones.

In the following paragraphs we concentrate on the first three points, since equalization facilities can be provided in the processing of MBROLA databases.

5.1. Creating a text corpus

Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. Their main interest in synthesis is that they minimize concatenation problems, since they involve most of the transitions and co-articulations between phones, while requiring an affordable amount of memory, as their number remains relatively small (as opposed to other synthesis units such as half-syllables or triphones).

Hence, the first step to build a diphone database consists of fixing a list of all the phones of a language. Notice that phones are acoustic instances of phonemes. Phonemes are themselves defined on a functional, linguistic level.

Obtaining a list of phones from a list of phonemes requires to number allophones, i.e. acoustic versions of some phonemes that significantly differ from the standard one, mostly due to co-articulation constraints. Although it is not necessary to account for all allophonic variations to build an intelligible synthesizer, the naturalness of synthetic speech may be affected if too few allophones are considered. In FR1, for example, we did not consider allophones at all. As a result, some allophonic phenomena, such as devoicing of /R/ when followed or preceded by unvoiced plosives, is only partially accounted for.

When a complete list of phones has emerged, including allophones if possible, a corresponding list of diphones is immediately obtained, and a list of words is carefully completed, in such a way that each diphones appears at least once (twice is better, for security). Unfavorable positions, like inside *stressed* syllables or in strongly reduced (i.e. over-co-articulated) contexts, should be excluded. One typically uses carrier sentences in which the word with the diphone considered is inserted. Notice that many diphones only appear in the association of words (i.e. not in single words). A number of diphones even never appear at all. Hence, the task of creating a text corpus which contains all existing ones is not trivial.

5.2. Recording the corpus

The corpus is then read, by a professional speaker if possible, digitally recorded, and stored in digital format.

In order for the MBROLA resynthesis operation to achieve best re-

sults, it is important to note that the corpus should be read with the most monotonic intonation possible (just like when reading a long and boring enumeration). Even the end of words should maintain their fundamental frequency constant. Since this is a totally unnatural way of reading a text, the speaker should train before starting the recording session.

Database provider who already have a diphone database which they want to adapt to the MBROLA format should contact the author (mbrola@tcts.fpms.ac.be), even if it has not been recorded with constant pitch. It is very likely that the database can be used anyway.

It is best to use high quality audio devices (microphone, pre-amp, A/D converter). The sound recording tools provided with many low-price commercial boards, for example, should be avoided, as they produce undesired recording noise. Better still, speech should be recorded with a DAT audio tape recorder and transferred in numerical format to the computer.
To roughly test the quality of the recording system, short circuit your microphone (caution, this should not be done with phantom plugs) and adjust the recording level, you can then evaluate the recording noise. In the case of FR1, the noise level only corrupted the last three bits of our data, leaving thirteen significant bits.

Another important type of noise to avoid is ambient noise and reverberation. Particularly, the recording should be free of low frequency noises carried by building substructures. For instance, you generally won't notice low frequencies due to trucks passing in the neighborhood, but your high quality microphone will hardly fail to detect them. The best way to avoid them is to install your recording system inside a professional soundproof room, as we did for the FR1 database.

5.3. Segmenting the speech corpus

Once the corpus has been recorded, all diphones must be spotted, either manually with the help of signal visualization tools, or automatically thanks to segmentation algorithms, the decisions of which should be checked and corrected interactively. A diphone database is finally created, which summarizes the results, in the form of : the name of diphones, the related waveforms, their duration, and internal sub-splittings. As a matter of fact, the position of the border between phones should be stored, so as to be able to modify the duration of one half-phone without affecting the length of the other one.

For optimal results with MBROLA, it is best to keep diphones in context. The resynthesis operation, indeed, includes some pitch analysis, which itself achieves more accurate results when, say, 100 ms of speech are kept at the left and right of each diphone.

6. CONCLUSIONS

We hope that the widespread dissemination of scientific tools for phonetization and transcription, and for the testing of prosody generation algorithms, will give life to new collaboration efforts in the field of speech synthesis.

An increasing number of scientists working in the area of natural language processing and speech synthesis developers working on prosody generation won't be unwelcome, as the production of natural or at least credible intonation and rhythm variations will likely remain one of the weakest points of speech synthesis for the years to come.

For more details, and to download MBROLA and its diphone databases, you are invited to report to the MBROLA WWW Home page:

http://tcts.fpms.ac.be/synthesis/mbrola.html

This WWW page includes mirror addresses in France, Switzerland and in the USA.

7. ACKNOWLEDGMENT

We would like to thank Vincent Fontaine (TCTS Lab,Mons, B), Arnaud Gaudinat (University of Geneva, CH), and Sam Przyswa (Paris, F) for their help in the compilation of MBROLA. We also thank Jeff Bilmes (ICSI, Berkeley, US) and Mark Liberman (University of Pennsylvania, US) who have arranged mirror sites, and last but not least, the many individuals who have manifested their support to this project.

8. REFERENCES

1. R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D.G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue. The challenge of spoken language systems: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3:1–21, 1995.
2. T. Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Boston, 1996. Forthcoming textbook.
3. E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:5–6, 1989.
4. P. Meyer, H. W. Rühl, R. Krüger, M. Kugler L. L. M. Vogten, A. Dirksen, and K. Belhoula. PHRITTS: A text-to-speech synthesizer for the german language. In *Eurospeech'93*, pages 877–890, Berlin, 1993.
5. T. Dutoit and H. Leich. MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 1993.
6. T. Dutoit. High quality text-to-speech synthesis: A comparison of four candidate algorithms. In *ICASSP'94*, volume 1, pages 565–568, Adelaide, Australia, April 1994.
7. V. Pagel, N. Carbonell, and Y. Laprie. A new method for speech delexicalization, and its application to the perception of french prosody. In *ICSLP'96*, Philadelphia, 1996.