

Efficient diphone database creation for MBROLA, a multilingual speech synthesiser

Jolanta Bachan, *Institute of Linguistics, Adam Mickiewicz University*

Abstract

Diphone synthesis is a convenient way for testing phonetic models of human speech. It allows easy manipulation of duration and pitch, therefore it is used not only for general intonation contour evaluation, but also for expressive speech synthesis. The main advantage of using MBROLA [11][9],[12],[13] is the fact that not all the diphones need to be contained in the voice to test speech models. For experimental work with specific tasks such as expressive speech [3],[6],[7], or underresourced languages¹, it is enough to create micro-voices which contain just a handful of diphones [3].

In the present paper tools created for efficient diphone database building for MBROLA speech synthesiser are presented. First, having a large corpus of transcribed texts, it is possible to select a small set of phonetically rich sentences which contain all the diphones for a given language. Second, having annotated a speech corpus on the phone level, it is easy to extract all the present diphones. Both tools were tested and evaluated on large text and speech corpora of the Polish language and a demo-MBROLA voices was created.

With phonetically rich sentence extractor and diphone extractor tools, building new MBROLA voices for any language is very efficient and encourages work on evaluating speech models which can be later used in more advanced speech synthesis systems or applied in spoken dialogue systems.

1. Introduction

The aim of the Ph.D. thesis is to model different speech styles which will align with the speaker in a consultation situation, particularly where the interlocutor is in a stress situation, and, based on the phonetic and linguistic characteristics of the speaker's speech, to design and build a speech synthesis component and style selection module for an adaptive dialogue system. The dialogue system is intended to adapt its speech by selecting a speech

style appropriate for the speaker's level of speech arousal. Adaptive dialogue systems of this kind are intended to improve human-computer interaction at emergency unit control centres and the help desks of call centres, by making the dialogue more natural.

The present contribution is concerned with issues in the efficient development of different voices for the speech synthesis component.

MBROLA, a multilingual speech synthesiser, although a bit outdated in technology, is very useful for testing speech models in linguistic work. This diphone concatenative synthesiser does not only allow easy manipulation of duration and pitch values, but it is also easy to create new synthetic voices which can be used in experimental work. MBROLA voices do not need to contain all the diphones for a given language. For certain tasks it is enough to create micro-voices with just a handful of diphones. Recently, new MBROLA micro-voices have been created for evaluation of expressive speech [3], synthesised dialogues, voice quality [23] and underresourced languages such as Thadou, a Sino-Tibetan language [14], and 11 Nigerian languages¹. Additionally, MBROLA was used for large speech corpora evaluation by the means of the Automatic Close Copy Speech (ACCS) synthesis [2],[14].

The interest in using MBROLA invoked the need to automatise the linguistic work. Two tools have been created to make the MBROLA voice creation faster. First, the phonetically rich sentence extractor picks up the smallest possible number of sentences with the largest number of new diphones from a very large text corpus of transcribed sentences. These sentences can be then recorded and annotated either manually or automatically using available software for a given language, e.g. SALIAN for Polish [20], Aligner for German [1]. Second, the diphone extractor selects diphones from the recorded speech corpus annotated on the phone level in the format needed for creating a new MBROLA voice with Mbrolator², the voice creation software.

¹ Fieldwork carried out by Dafydd Gibbon in Abuja, Nigeria in 2010.

² Mbrolator – the mbrolation software – was provided under license to Dafydd Gibbon, Universität Bielefeld, Germany, by Thierry Dutoit and Vincent Pagel.

2. General MBROLA voice creation procedure

The core of a diphone synthesiser is the diphone database (also called the voice). Creating the database must be well thought-out, because it has to include all the possible diphones in the language of concern. Creation of the diphone database is mainly achieved in four steps [13]:

1. Creating a text corpus:
 1. A list of phones (PL), including allophones if possible, for a given language is prepared.
 2. Out of the list of phones a list of diphones (DL) is generated; in general $|DL| \leq |PL|^2$.
 3. A list of words containing all the diphones is created. Each diphone should appear at least once; diphones in such positions as inside stressed syllables or in strongly reduced (i.e. overcoarticulated) contexts should be excluded.
 4. The key words are put in a carrier sentence.
2. Recording the corpus:
 1. The corpus is read by a professional speaker with monotonous intonation.
 2. The speech is digitally recorded and stored in a digital format.
3. Segmenting the corpus:
 1. The diphones must be found in the corpus and annotated either manually or automatically by the means of automatic segmentator.
 2. The position of the border between the phones is marked.
4. Equalising the corpus:
 1. The energy levels at the beginning and at the end of a segment are modified in order to eliminate amplitude mismatches – the energy of all the phones of a given phoneme is set to phones' average value [11].
 2. Pitch is normalised.

When the corpus is created it contains information in the following format: the name of the diphones, the corresponding waveforms, their durations and internal sub-splitting. Such created diphone database allows to modify the duration of one half-phone without affecting the length of the other [13].

The present paper is addressed at the points #1 and #3. First, the aim is to reduce the number of sentences to be read. – This affects point #2, as instead of dozens of sentence carrying one target diphone, a few phonetically rich sentences may be recorded. Second, the presented tool is to extract the diphones from an already annotated speech corpus automatically, regardless the original purpose of the corpus. Point #4 is performed by Mbrolator and its output is an MBROLA voice.

2.1. Mbrolation

The Mbrolator, is a software suite for MBROLA voice creation. The requirements of the system are

diphone files in the WAV format and diphone database file in the SEG format.

The restrictions put on the diphone files are:

1. the diphone WAV files need to be at 16kHz sampling rate;
2. the diphone WAV file cannot be longer than 10000 samples;
3. for each diphone a context of 500 samples needs to be left on the left and on the right sides [12].

If the rules #1 or #2 are violated, the Mbrolator will exit and no voice will be created. The rule #3 is set to give the MBE analysis and pitch extraction algorithm a window which is large enough for the correct analysis.

The overall process of MBROLA voice creation with Mbrolator is shown on Fig. 1.

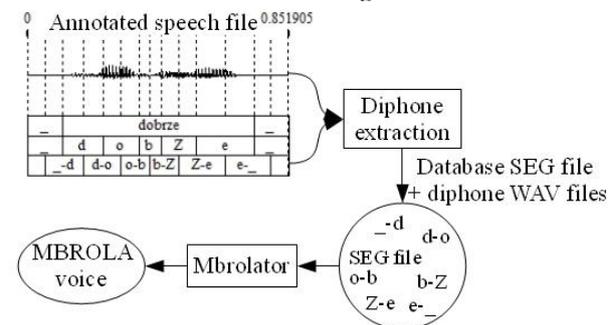


Fig. 1. MBROLA voice creation with Mbrolator software.

3. Phonetically rich sentence extractor

The objective is to select the smallest possible set of sentences from a text corpus which will contain the largest number of diphones. Such a sentence set may be then recorded and annotated, the diphones extracted and a new MBROLA voice may be created. Usually, in order to collect diphones, researchers create a large set of sentences aimed at containing one target diphone in a carrier sentence. The phonetically rich sentence extractor searches for diphones in a text corpus and selects only the sentences which contain the most new diphones in relation to those chosen before.

3.1. Diphone set creation

A full MBROLA voice should contain all the diphones of a given language. A diphone (or a dyad) is a unit that begins in the middle of the stable state of a phone and ends in the middle of the following one [11]. The theoretical number of diphones in a given language is $|DL| \leq |PL|^2$, where DL represents a diphone list and PL represents a phone list. However, not all the phones appear in all phone contexts, therefore the actual number of diphones for a natural language may be smaller.

The calculation of the diphones was done to find out the optimal number of diphones in Polish. For the Polish language, two sets of phones are accepted,

both being applied to the SAMPA alphabet 1.[21]: the Polish SAMPA [18] and the Polish Extended-SAMPA [8], [16]. The Polish SAMPA contains 37 labels. Adding the pause to it, it gives a maximum of $38 \times 38 = 1444$ diphones. The Polish Extended-SAMPA contains 40 labels, with the pause it gives $41 \times 41 = 1681$ diphones.

3.2. Available text resources

For the present study, a set of sentences created for speech technology purposes were used. 1623 sentences were taken from the Bonn Open Synthesis System (BOSS) corpus [5],[10] and 8828 sentences came from the Jurisic database [9], a database made for a speech recognition system creation. Altogether 10451 sentences were written orthographically and saved in a text document in the ANSI encoding, which is required by the automatic transcription software. The sentences were then transcribed using PolPhone 1.[22][20].

3.3. Sentence extraction procedure

On Fig. 2 the phonetically rich sentence extraction procedure is presented. First, the sentences are automatically transcribed using PolPhone [22]. Then, the sentences are divided into diphones and the number of diphones for each sentence is calculated. Next, the sentences are sorted according to the descending number of diphones. Having the sorted list, the program creates an empty set of diphones and adds to it the diphones which occur in the selected sentences: The algorithm compares the diphones in the processed sentence. If it contains new diphones which have not occurred before, the sentence is selected and the new diphones are added to the diphone set.

The algorithm is designed in such a way that first it can select only the sentences which contain 10 new diphones or more. Then the number may be decreased by 1 in a loop until 1. In this way all the sentences are checked and selected even if they contain only 1 new diphone. However, the precedence is given to the diphone richest sentences.

3.4. Results of sentence extraction

The phonetically rich sentence extractor was run on the available text resources. According to the Polish SAMPA, 1008 diphones were found in the corpus and 211 sentences were selected out of 10451. When the Polish Extended-SAMPA phone set was applied, 1095 diphones were found and 201 sentences were selected from the text corpus.

4. Diphone extractor

4.1. Available audio resources

A corpus of recordings for a male voice was available from a BOSS speech synthesis development scenario [5]. All the recordings, i.e. 1580 sentences, were annotated automatically and then checked manually by trained phoneticians. The annotated phone set contained 40 labels, in accordance with the Extended-Polish SAMPA phoneme set. Some additional marks were added during the automatic and manual annotation procedure, but they are not the point of this study.

4.2. Design and workflow

For automatic diphone extraction, a system is designed to automatically cut out diphones from the recordings based on the annotations of those recordings. In the first stage the system searches for the diphones in the annotations on the phone level and cuts them out from the recordings, creating at the same time a database file in the SEG format with information about those diphones. The information in the database includes the labels of the diphones which have been extracted, the names of the files in which the individual diphones are, the beginning and the end of the diphones and the placement of the boundary between the half-phones.

The extracted diphones are then put forward to the evaluation stage in which both, the database file and the diphone files are evaluated.

Based on the database file, annotation files for separate diphones are created. These diphone annotation files allow manual comparison of the annotations with the signal in the diphone files to see

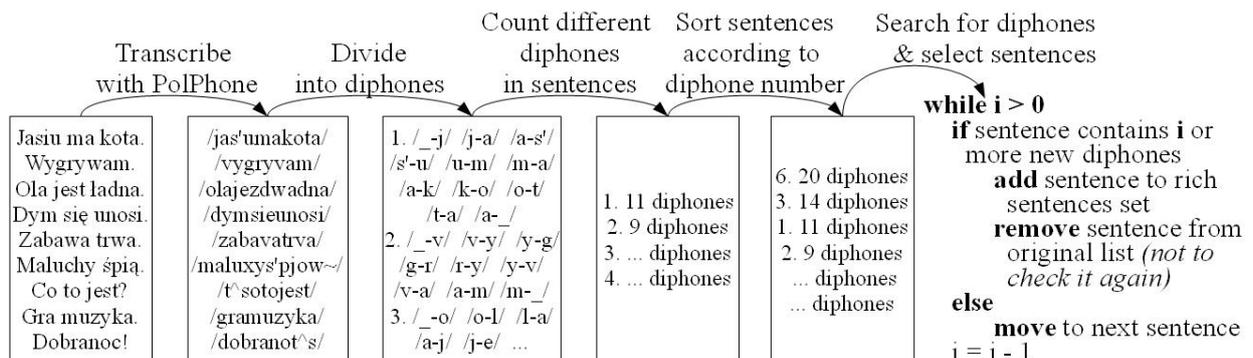


Fig. 2. Phonetically rich sentences extraction procedure

if the annotation is correct with the signal. Additionally, manual investigation of the diphones with their annotations allows evaluation of the quality of the diphone and selection of diphones of the best quality for the synthetic voice creation.

The next step in diphone extraction evaluation is the concatenation of the extracted diphones. The extracted diphones from one utterance are glued back together. In the ideal case, the concatenated diphones extracted from one recording with preserved order of the diphones' occurrence should give a re-synthesised speech output identical with the original recordings, without any glitches or repetitions. If the extraction of diphones was not correct, the synthesised speech signal would be disrupted.

The overall architecture of the automatic diphone extraction system is presented on Fig. 3.

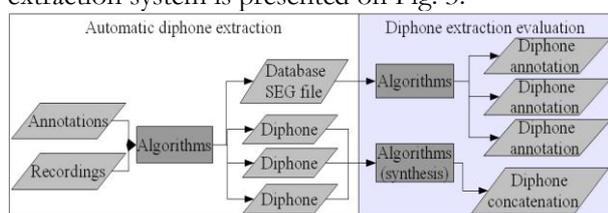


Fig. 3. Architecture of the automatic diphone extraction system

4.3. Automatic diphone extraction system implementation

The automatic diphone extraction system was implemented in Python [17] on the Linux platform, with support of the SOX software [19] for the processing of the recording files. The diphone extraction includes the following steps:

- BLF to TextGrid conversion - annotation file format conversion from BOSS BLF format to Praat TextGrid format³ [4];
- Extended-SAMPA TextGrid to SAMPA TextGrid conversion (optional);
- Search for all diphones in TextGrid files and DIPH file creation – DIPH file contain information about the diphone IDs and diphone labels, the clean phone labels without additional annotation marks and information about time boundaries;
- Diphone extraction and database SEG file creation – either all instances of all diphones are extracted, or only one instance of each diphone is extracted.

The architecture of the automatic diphone extractor is presented on Fig. 4. The dotted arrows correspond to the optional conversion of the SAMPA alphabet.

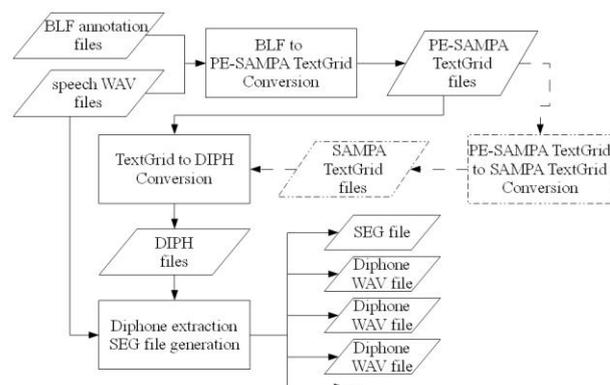


Fig. 4. Architecture of the automatic diphone extractor. PE-SAMPA stands for the Polish Extended-SAMPA

Evaluation of the automatically extracted diphones is based on:

- Generation of annotation TextGrid files for diphones for manual investigation;
- Concatenation of diphones in order to assess the re-synthesis of pre-processed diphones.

4.4. Diphone extraction evaluation and mbrolation

In 1580 recorded and annotated sentences, the diphone extractor found and extracted 1039 diphones when the Polish SAMPA phone set was applied and 1058 diphones in accordance with the Extended-Polish SAMPA. Most of the missing diphones do not occur in the Polish language.

Selected diphones were examined manually together with the TextGrid annotation files created from the SEG files. Additionally, the diphones were concatenated and evaluated perceptually by a phonetician with positive results.

The extracted diphones were mbrolated and 2 MBROLA voices were created. The voices were then assessed by the means of the Automatic Close Copy Speech (ACCS) synthesis [2],[14].

5. Conclusions and future work

In the present paper, two tools for automatizing MBROLA voice creation were presented. The phonetically rich sentence extractor drastically minimises the volume of the text corpus, at the same time decreasing the amount of recording and annotation to be carried out. The procedure of creating carrier sentences for the target sounds is replaced by the automatic diphone search in an existing corpus, as in unit selection speech synthesis systems. Additionally, in order to extract diphones from a corpus of annotated speech, a tool was presented which cuts out diphones and creates a database file which can be then mbrolated and a new MBROLA voice created. In the future, both methods are to be combined, i.e. recordings of the phonetically rich sentences performed and an

³ In the BLF file the information about the final pause is not included. To reconstruct it, it is necessary to measure the length of the recording.

MBROLA voice created. The sentences may be collected on the Internet sources, such as news or thematic portals, do not need to come from the text corpora created for speech technology purposes.

Having only just 200 sentences to be recorded and annotated, it should be efficient to create an MBROLA voice for any language. This should enhance the experimental work with speech synthesis in different environments, in small student groups in linguistic, phonetic, psycholinguistic and technical studies in technologically less well equipped countries.

Acknowledgements

This work was partly funded by the research supervisor project grant to Prof. Grażyna Demenko and Jolanta Bachan No. N N104 119838 titled "Communicative alignment of synthetic speech" and by an international cooperation scholarship funded by the Bielefeld University, Germany, and by a scholarship for scientific achievements funded by the Kulczyk Family Foundation.

The author is very grateful to Prof. Grażyna Demenko for providing the text and speech corpora and to Prof. Dafydd Gibbon for his invaluable advice on the system design and implementation.

Bibliography

- [1] Aligner - automatische Segmentierung von Sprachsignalen <<http://www.ims.uni-stuttgart.de/phonetik/helps/aligner.html>>, accessed on 2010-09-19
- [2] Bachan, J. 2007. Automatic Close Copy Speech Synthesis. In: *Speech and Language Technology. Volume 9/10*. Ed. Grażyna Demenko. Poznań: Polish Phonetic Association. 2006/2007, pp. 107-121
- [3] Bachan J. & Surmanowicz, B. 2008. Preliminary results of expressive speech synthesis in Polish. In: G. Demenko, K. Jassem, S. Szpakowicz (Eds.) *Speech and Language Technology, Vol 11*. Poznań: Polish Phonetic Association, pp. 103-112
- [4] Boersma, P. & Weenink, D. 2001. PRAAT, a system for doing phonetics by computer. In: *Glott International 5(9/10)*, pp. 341-345
- [5] BOSS, the Bonn Open Synthesis System. <<http://www.ikp.uni-bonn.de/forschung/phonetik/sprachsynthese/boss/>>, accessed on 2010-09-19
- [6] Burkhardt, F. 2005. Emofilt: the Simulation of Emotional Speech by Prosody-Transformation. In: *Proceedings of Interspeech 2005*, pp. 509-512
- [7] Burkhardt, F. Emofilt. <<http://emofilt.syntheticspeech.de/>>, accessed on 2010-09-19
- [8] Demenko, G., Wypych, M. & Baranowska, E. 2003. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. In: G. Demenko & M. Karpiński (Eds.) *Speech and Language Technology, Vol. 7*. Poznań: Polish Phonetic Association, pp. 79-95
- [9] Demenko, G., Grocholewski, S., Klessa, K., Ogórkiewicz, J., Wagner, A. Lange, M., Śledziński, D. & Cylwik, N. 2008. JURISDIC: Polish Speech Database for Taking Dictation of Legal Texts. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. 28-30 May 2008, Marrakech, Morocco
- [10] Demenko, G., Klessa, K. Szymański, M. & Bachan, J. 2007. The design of Polish speech corpora for speech synthesis in BOSS system. In: *Proceedings of XII Sympozjum Podstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki (PPEEm 2007)*. Wisła, Poland, pp. 253-258
- [11] Dutoit, T. 1997. *An Introduction To Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers
- [12] Dutoit, T. 2005. The MBROLA project. <<http://www.tcts.fpms.ac.be/synthesis/mbrola.html>>, accessed on 2010-09-19
- [13] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken O. 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: *Proceedings of ICSLP 96, vol. 3*. Philadelphia, pp.1393-1396
- [14] Gibbon, D. & Bachan, J. 2008. Automatic Close Copy Speech Synthesis Tool for Large-Scale Speech Corpus Evaluation. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. 28-30 May 2008, Marrakech, Morocco
- [15] Gibbon, D. Pandey, P., Haokip, D. M. K. & Bachan, J. 2009. Prosodic issues in synthesising Thadou, a Tibeto-Burman tone language", In *INTERSPEECH-2009*, 500-503
- [16] Jassem, W. 2003. Polish. In: *Journal of the International Phonetic Association, Vol. 33*. Cambridge: Cambridge University Press, pp. 103-107
- [17] Python Programming Language. <<http://www.python.org/>>, accessed on 2010-09-20
- [18] SAMPA - computer readable phonetic alphabet. Polish. Maintained by J.C. Wells. Created on 1996-09-06. <<http://www.phon.ucl.ac.uk/home/sampa/polish.htm>>, accessed on 2010-09-20

- [19] SoX - Sound eXchange
<<http://sox.sourceforge.net/>>, accessed on 2010-09-20
- [20] Szymański, M. & Grochowski, S. 2005. Transcription-based automatic segmentation of speech. In: Proceedings of 2nd Language and Technology Conference, Poznań, pp. 11–14
- [21] Wells, J.C. 1997. SAMPA computer readable phonetic alphabet. In: Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B., <<http://www.phon.ucl.ac.uk/home/sampa/>>, accessed on 2010-09-20
- [22] Wypych, M., Demenko, G., Baranowska, E. & Szymański, M. 2002-2006. PolPhone – Polish phonetizing filter. Based on phonetization rules from M.S. Batóg and W. Jassem
- [23] Windmann, A. 2009. Perspectives of Corpus-Based Implementation of Voice Quality in Concatenative Speech Synthesis. University of Bielefeld: MA thesis

Author:



Jolanta Bachan, M.A.
Institute of Linguistics
Adam Mickiewicz University
ul. Aleja Niepodległości 4
61-874 Poznań, Poland
tel. +48 61-829-36-63
fax +48 61-829-36-62

email: jolabachan@gmail.com
