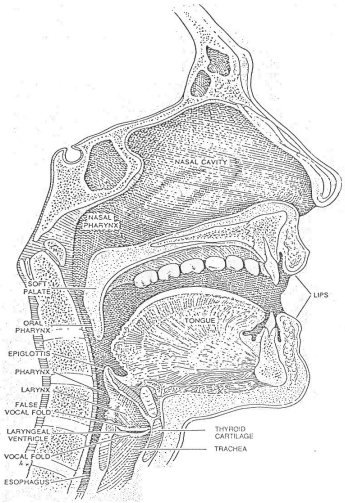
Polish-Japanese Institute of Information Technology

MBROLA.
Creating the Polish diphone database for speech synthesis.
Summary of Master's thesis.



Krzysztof Szklanny

Warsaw 2002

1. Preparing the Polish diphone database

The goal of my work was to create the Polish diphone database. Additional aim of the project was to obtain high quality speech synthesis for the Polish language. The whole process included several stages. First of all I had to prepare a phoneme list and create the corpus of diphones. Then next was to make the recordings of the corpus and the segmentation of diphones. One of the last processes was to test the database then export it and sent it to the MBROLA team who conducted the normalization process.

2. Preparing the corpus

The most important thing at this stage was to create such a context of nonsense words including diphones that, the diphone could not be the stressed syllab. Secondly it's neighborhood should not influence the co-articulation of the diphone.
Sometimes there were no rules that how to find the appropriate context of the diphone and that was the most difficult problem while creating the corpus.

The corpus includes such information as :
the name of the diphone, context of the diphone, number of wave file where it is placed, and the three numbers which are :

- The beginning
- The middle
- The end of the diphone

For example:

i i w0.wav ani imadwo 41658 42577 43144

These numbers are presented in samples. In order to get the real time of the diphone it is necessary to divide the number by 16000 which is the frequency of recorded corpus. Such a frequency allowed me to obtain the best quality of speech synthesis and preserve a good quality/size ratio.

As I mentioned, the next thing to do was to make recordings. These took place in the Polish-Japanese Institute of Information Technology in it's recording studio.
Four microphones were used. A so called close-talk microphone and three table stand microphones. In the segmentation process I used the close-talk microphone, which guaranteed the best quality signal and made no distortion.

The next stage was to prepare the segmentation process which was the most difficult and time-consuming stage.

3. Segmentation

For ***manual*** segmentation, the sound elements are taken from the speech material "by hand". Generally I used Praat[1] as a tool for segmentation. It has a built-in spectrogram and uses a graphic-acoustic display for marking the parts of the signal to be cut out and for acoustical control. During manual segmentation, considerable difficulties might occur. The main difficulties arise, if there are no sharp boundaries between the individual sounds (flowing sound transitions). In such cases, marking of the boundaries is often arbitrary.

Manual segmentation is very lavish and time consuming. That's why, ***automatic*** segmenting procedures have been developed. These procedures do not achieve the reliability of a good expert. However, the advantage of an automatic procedure is that it can make suggestions which facilitate the work for the user.

[1] Praat is a program for doing acoustic phonetic.

4. Characteristics of sound classes

For the segmentation, knowledge of the characteristics of the sound classes is necessary. The sound classes (subdivided according to the type of articulation) are specified in the table below with their characteristics in the time and frequency domain:

type of sound	example	time domain	frequency domain
vowels	/a/ /e/ /i/	- quasi-periodic - high energy	- 3 - 6 clearly visible formants - significant pitch peaks - main energy in the low frequency range
voiced plosives	/b/ /d/ /g/	- sudden and high rise of the amplitude	- formant at low frequencies - rise of the formants after closure opening
unvoiced plosives	/p/ /t/ /k/	- sudden and high rise of the amplitude - no signal before closure opening	- no formant structure - energy in high frequency area
Voiced fricatives	/v/ /z/ /j/	- broad quasi-stationary area - low energy	- existing formants - lower first formant compared to vowels - energy in high frequency area
unvoiced fricatives	/f/ /s/	- noise-like form - low energy	- broad spectrum
Nasals	/m/ /n/	- quasi-periodic - similar to vowels - lower energy than vowels	- minima within the spectrum - formants similar to vowels
liquids	/l/ /r/	- short or missing stationary area - weak transition to a following vowel	- lower first formant

Table 4.1 Characteristics of sound classes

5. Conventions

The segmentation of the sound elements should be executed very carefully because the quality of a speech synthesis system strongly depends on correct segmentation. The following fundamental conventions have to be considered.

Cutting must be done always in a positive zero crossover. Thus, errors at concatenation boundaries can be avoided or reduced when assembling the sound elements.

The cut has to be made at the beginning of a new period. The figure below shows this principle.

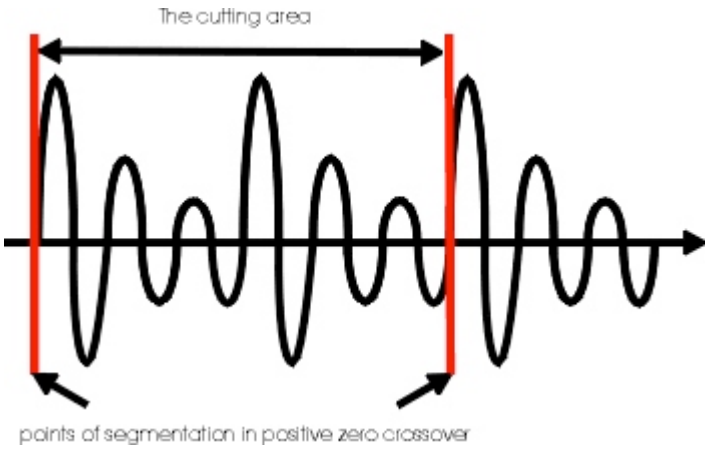


Figure 5.1 Conventions for the segmentation of sound elements

The segmentation of ***vowels***, ***nasals*** and ***laterals*** should be made in the zero crossover before the pitch mark (The figure 5.2).

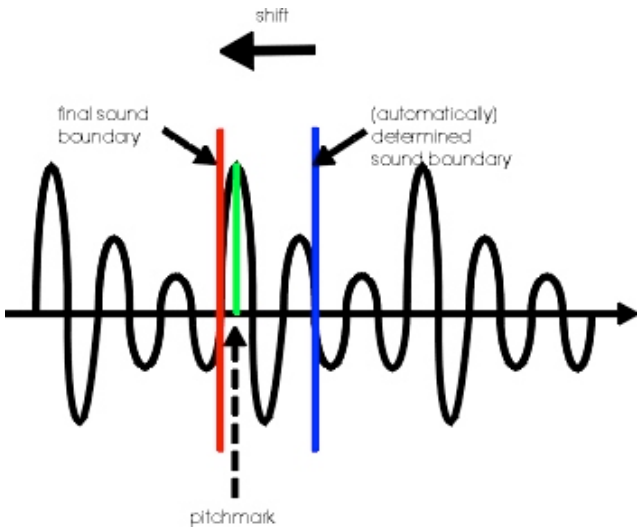


Figure 5.2: Conventions for the segmentation of vowels, nasals and laterals

The segmentation of ***fricatives***, ***affricates*** and ***intermittends*** should be selected as follows (The figure 5.3):

- If the boundary proposed by the system shows positive values, it has to be shifted to negative time direction.
- If the boundary has negative values, it has to be shifted in positive time direction to the next positive zero crossing.

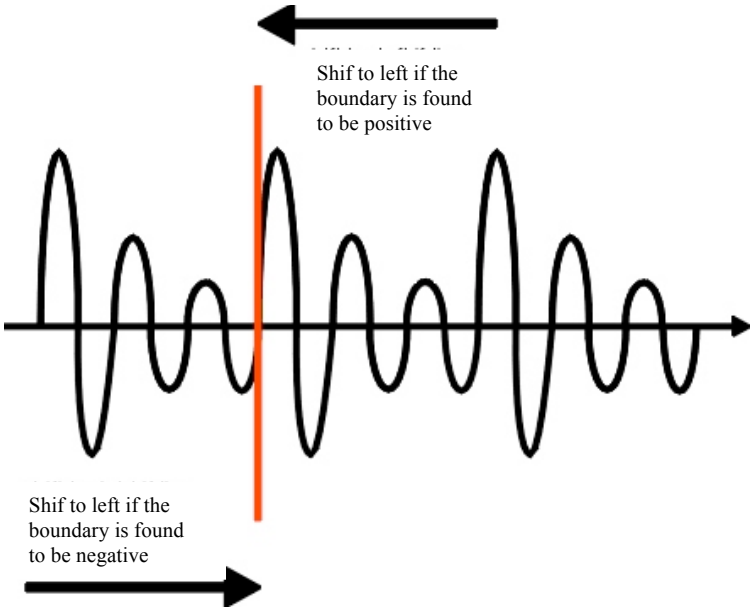


Figure 5.3: Conventions for the segmentation of fricatives, affricates and intermittends

Plosives are segmented similarly. The boundary should be shifted always temporally forward. The sudden rise of the amplitude at the start of the sound is particularly critical for the heard impression of the sound.

6. Problems

After the concatenation of the segmented sound elements the result is often unsatisfactory, because discontinuities at the sound boundaries cause clearly audible disturbances. Discontinuities can occur on the following items:

- amplitude

The ***discontinuities of amplitude*** are already visible in the time domain. They are produced, if the amplitudes at the end of a sound and at the beginning of the following sound are strongly different. It is clearly audible as “cracks”. The figure 6.1 shows a discontinuity of the amplitude in the time domain.

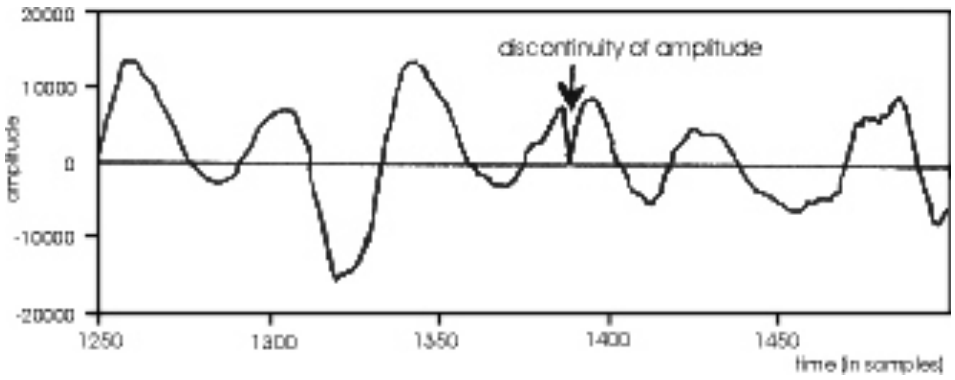


Figure 6.1: Discontinuity of the amplitude

- energy

The ***discontinuities of energy*** are produced by different volumes of the speech material. Great changes usually exist over time. The figure 6.2 shows a discontinuity of the energy in the time domain and the figure 6.3 the power over time.

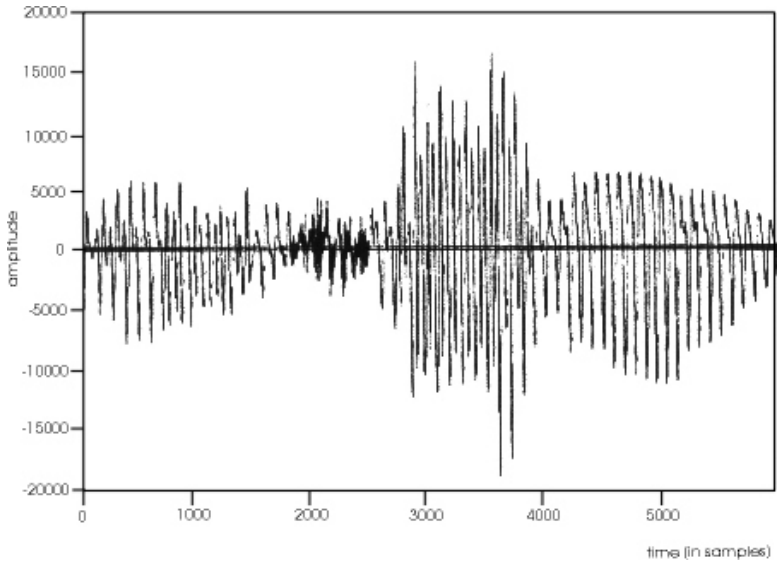


Figure 6.2: Discontinuity of the energy (time domain)

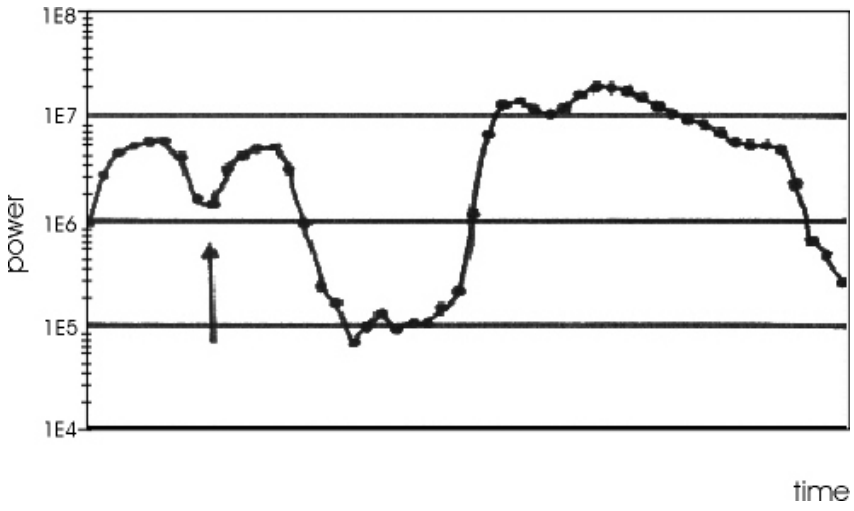


Figure 6.3: Discontinuity of the energy (power over time)

- frequency

The ***discontinuities of the frequency*** are very short, but they are clearly audible as cracks. The figure 6.4 shows an example.

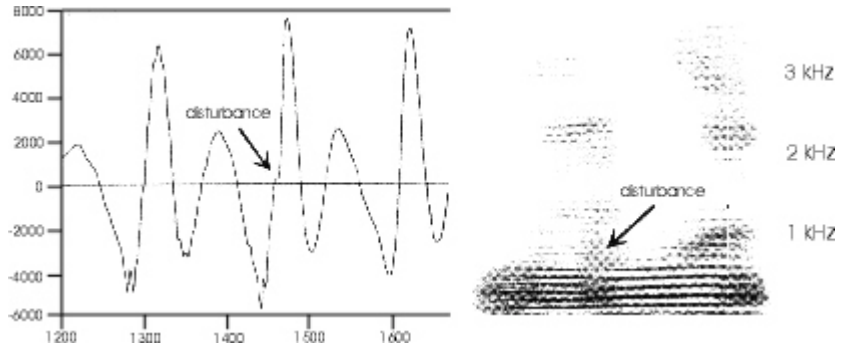


Figure 6.4: Discontinuity of the frequency (time and frequency domain)

- phase

Discontinuties of phases occur, if the boundaries are not set at the beginning of a new pitch period. They are also clearly audible as cracks. The figure 6.5 shows the effect of such a discontinuity ("additional" area).

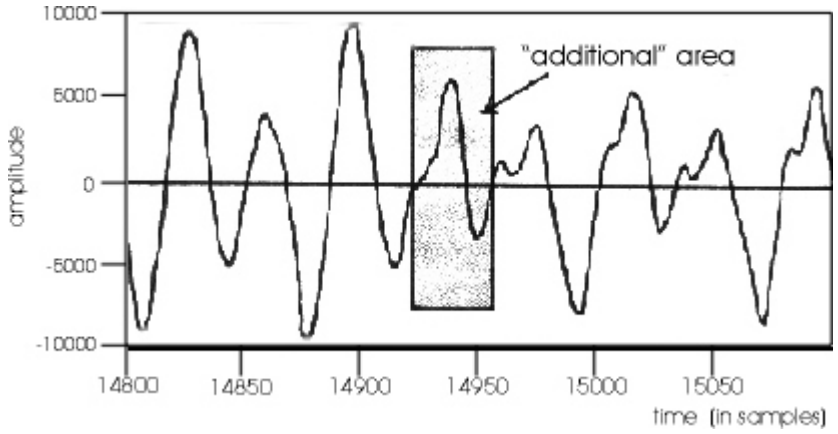


Figure 6.5 Discontinuity of the phase

These disturbances can be widely reduced or avoided by careful segmentation.

7. Testing

The last stage was to test the quality of the diphone database in the Diphone Studio program (the figure 7.1) and send the database to MBROLA in order to normalize it. So I used the test made by Krzysztof Marasek and Ryszard Gubrynowicz. It consists of all the possible connections that exist in the Polish language and then sent to Belgium.

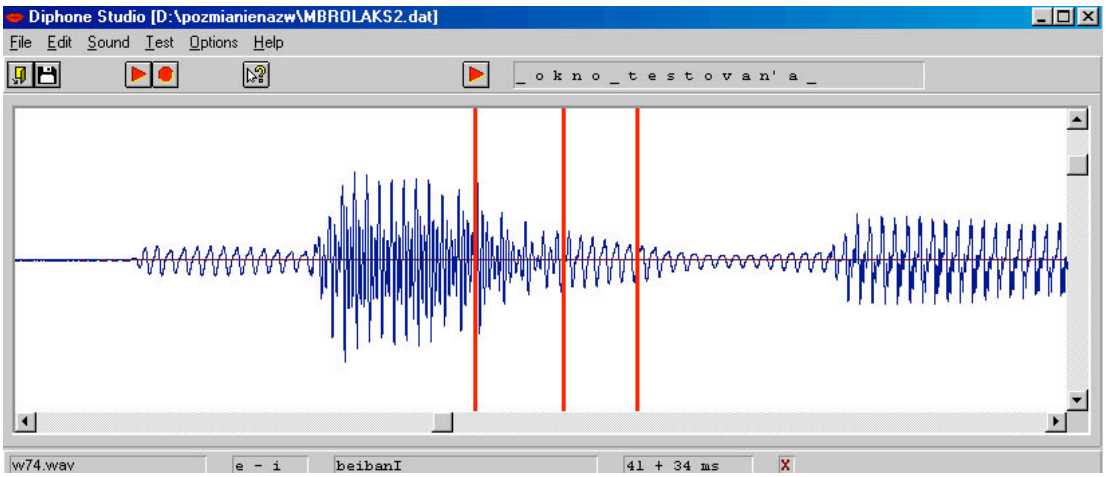


Figure 7.1 Diphone Studio window

8. Conclusions

The goal of my work was to create the Polish speech synthesis based on diphones. This is for the realization of speech synthesis for the MBROLA system. The requirement was to create the diphone database that generates speech synthesis in the best possible way.

The quality of speech synthesis must be intelligible, and it has to sound natural so that the acoustic model could be further used in public applications, which use speech synthesis. By public use I mean the using of the database for education, voice portals, Enhanced Eyes-Free Access to Critical Information While Driving, and also an aid to the people with speech disabilities.

There were a few stages that had to be done. First of all I had to prepare the corpus that could be used further in the MBROLA project, which was additional work. Next I had to conduct the recordings. The most sophisticated stage was the realization of the

segmentation stage. It required accuracy and precision. The last stage was to test the quality of speech synthesis by using the most popular connections of diphones in the Polish language. Finalization of the work was the normalization of the database by MBROLA in Polytechnic in Mons.

Now the speech synthesis system works for the input data as phonetic transcription. The next stage will be the creation of the prosodic model and natural language processing in order to create the full TTS system.

The polish database is available since May this year on the MBROLA web site (http://tcts.fpms.ac.be/synthesis/mbrola) as new voice model of polish language.

Sources

http://www.ias.et.tu-dresden.de/kom/lehre/tutorial/selection.htm - Corpus segmentation
korpusu