# 1 Lab Book S3

In this lab book you will work with the two tools, MBROLA diphone voices with front end realized by **eSpeak**:
Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996, October). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 3, pp. 1393-1396). IEEE.
https://espeak.sourceforge.net/

## 1.1 Assignment 2

**[10pts]** *Manually managing code-switching with eSpeak and MBROLA*:

**Intro:**

eSpeak is a compact open source software speech synthesizer for English and other languages, that uses a "formant synthesis" method. This allows many languages to be provided in a small size. We will use eSpeak-NG, a fork of original eSpeak (and call it just "eSpeak" for brevity) as a front end for MBROLA diphone voices (which are cost-free but not open source). We will be working with the `us1` voice.

### Installing and trying out MBROLA and eSpeak

Both MBROLA and eSpeak-NG can be installed out of the box. For Debian-based Linuxes you can use

```
apt install mbrola-us1
apt install espeak-ng
```

(if you are using Google colab, prefix the shell commands with an exclamation mark, e.g. `!apt install mbrola-us1`).

You can find different commands for eSpeak here: https://espeak.sourceforge.net/commands.html. For the following tasks we will need the transcription that eSpeak generates based on the available diphones for the US1 voice [a]. Here are commands that we will use:

- To examine a transcription you can call

  ```
  espeak-ng -v mb-us1 -q --pho 'Hey! Enter your text here'
  ```

  where

  - `-v mb-us1` asks to use the "us1" voice from mbrola
  - `-q` disables sound output
  - `-pho` prints the transcription (in this case – in mbrola format)

  The transcription consists of multiple lines similar to `EI 60 0 232 80 215 100 215`:

  - `EI` is the phoneme to be generated
  - `60` is its duration (60ms)
  - the rest is a sequence of (time-F0) pairs, where time is in percents of the phoneme's duration. In this example, it starts with F0=232, changes to 215 at T=80% (thus, 48ms) and then to 251 Hz at the end of the phoneme.

- To save this transcription to a file "testmb":

```
espeak-ng -v mb-us1 -q --pho 'Hey there' > testmb
```

- To synthesize speech based on the transcription in "testmb" file using US1 voice and to write it into a "test.wav" file:

```
mbrola /usr/share/mbrola/us1/us1 testmb test.wav
```

**Task:**

Here is a short text in English that uses Dutch names:
`Characters from Dutch children's books - Jip and Janneke - has recently become part of an expression "jip-en-janneketaal" meaning "simple language" or layman's terms. It is most often used in context of politics and software engineering.`

You will need to:

- Transcribe the text with eSpeak.
- Synthesize spoken version of this text using MBROLA's US1 voice.
- Edit the transcription manually to adjust to pronunciation of Dutch words in English text so they will sound as Dutch as possible given the US1 set of diphones [b].
- Synthesize the spoken version of this text based on the adjusted transcription using MBROLA's US1 voice.
- Reflect on the results.
- Describe the steps you had to take: why you changed any of diphones (if any) and why you adjusted duration and pitch settings for certain diphones (if any).

## Submission

Alongside with this LaTeX file with your steps description and reflections, commit the adjusted transcription and the new voice sample(s).

---

[a] There are problems with phoneme “02”, equivalent to /ɒ/ in “often” and “software” in MBROLA, based on the documentation there are at least 2 similar sounds you can consider instead: /ʌ/ (MBROLA symbol “V”) as in “nut” (in the original MBROLA transcription this sound is used to synthesize the vowel in “of”), or /ɑ/ (MBROLA symbol “A”) as in "Arthur"

[b] "jip-en-janneketaal" in IPA is [jɪp ɛn jɑnəkətɑːl]

### Answer

Initially, I utilized the eSpeak us1 package to transcribe the text and subsequently synthesized it for the first attempt. Upon reviewing the synthesized audio, I observed that the Dutch names were pronounced entirely in English.

Subsequently, I consulted the official manual from eSpeak and Kirshenbaum's Wikipedia page to establish a mapping between Dutch and English transcriptions. Based on this information, I made adjustments to the phonemes of the Dutch names in the English transcription. Additionally, I noted from the official description of the us1 diphone database on GitHub that the diphone matrix is square, meaning there are no missing diphones. This implies that I didn't need to overly concern myself with phoneme pairings, as each phoneme in this database is mutually paired to form diphones.

The adjustments made to the transcription of Dutch names for a more Dutch-like pronunciation were as

follows:

- As 'Jip' is the first word in a compound phrase, I modified the pitch of 'Jip' to gradually rise, indicating a conjunction.
- After consulting 'jip-en-janneketaal' online and Dutch language resources, I discovered that Dutch does not employ the English-style "A and B's something" construction for joint possession. Instead, Dutch expresses shared possession only in the compound words close to the possessor, potentially leading to ambiguity. Consequently, I introduced a pause after 'Janneke' to indicate that 'taal' is shared by both 'Jip' and 'Janneke.'
- In Dutch, 'aa' represents a long vowel sound. Consequently, I extended the duration of the corresponding /A/ sound.

I believe these adjustments contribute to a more authentic Dutch pronunciation, aligning with the nuances of the US1 set of diphones. And the adjusted transcription is as below:

```
j 65
I 35 0 232 80 215 100 215
p 84

E 40 0 200 80 183 100 183
n 81 100 166

j 65
A 70 0 204 80 149 100 149
n 65
@ 19 0 158 80 152 100 152
k 102
@ 40 0 157 80 147 100 147
_ 100

t 65
A 300 0 151 80 143 100 143
l 115 100 131
```

## 1.2 Assignment 2 (Bonus)

**[3 bpts]** *Automatic mapping of transcriptions*:

**Task:**

Imagine you have a Dutch transcription made for Dutch MBROLA voices. Unfortunately, the diphone database is no longer available for you, so you need to make an automatic mapping of Dutch eSpeak transcription to English eSpeak transcription so MBROLA can read it with US1 voice.

- Take the following text "Ik wil wel graag werken aan dit project, maar ik weet niet of het gaat lukken. Ik zie veel beren op de weg."[a]
- To generate Dutch transcription, use `nl2` voice [b] (the `nl1` one seems to be broken).
- Similar to the previous task, describe the steps you had to take: what are the changes you decided to introduce and why.

## Submission

Alongside with this LaTeX file with your reflections, commit the complete code you used, the transcriptions - original and resulting, and the voice sample that you get by using the new transcription. Don't forget to write clear comments in your code.

[a] Translation: "I would like to work on this project, but I don't know if it will work. I see bumps in the road ahead" (literal translation of "ik zie veel beren op de weg" is "I see a lot of bears on the road").

[b] Don't forget to install the `mbrola-nl2` package

**Answer**

I began by transcribing the Dutch text using the nl2 voice and subsequently faced the challenge of generating an English eSpeak transcription for MBROLA's US1 voice. Since the diphone database for Dutch was unavailable, I opted for an automatic mapping approach.

Here are the steps I took:

- Identification of Non-English Phonemes:
  By comparing the phoneme descriptions in the official documentation on GitHub for both the us1 (English) and nl2 (Dutch) voices, I identified Dutch phonemes that did not have counterparts in the English phoneme database.

- Creation of a Phoneme Dictionary:
  I then created a dictionary mapping each non-English Dutch phoneme to its closest English equivalent based on the International Phonetic Alphabet (IPA). Given the inherent differences between Dutch and English phonetics, a perfect one-to-one mapping was challenging, but the dictionary provided reasonable substitutes.

- Automated Transcription Conversion:
  Utilizing a loop, I systematically examined each non-English Dutch phoneme in the transcription and replaced it with its corresponding English equivalent using the created dictionary. This automated process aimed to facilitate a smoother reading by the English eSpeak US1 voice.

- Adjustment for Dutch Vowel Length:
  Recognizing the difference in vowel length between Dutch and English, I realized that adjusting the duration of English phonemes based on Dutch vowel length could enhance the naturalness of the synthesized speech. However, due to time constraints, this aspect of fine-tuning was deferred for further exploration after this assignment.

In summary, the approach involved creating a phoneme dictionary for Dutch-to-English mapping, automating the conversion process, and acknowledging the potential impact of vowel length differences on the overall synthesis quality.