SHS homework 8: text-to-speech

Today we start the module Speech Technology by obtaining impressions of text-to-speech systems. Upload a report that describes your experiences.

Homework 8.1. Explicit acoustic synthesis via a KlattGrid in Praat.

Use the KlattGrid (Weenink chapter 12) to synthesize the twelve Dutch vowels with a duration of 0.2 seconds (short) or 0.4 seconds (long), as well as the three Dutch diphthongs with a duration of 0.4 seconds, assuming that the diphthongs go from [ɛ] to [i] (*sijs*), from [a] to [y] (*suis*), and from [ɑ] to [u] (*saus*). Synthesize for a male or female speaker. You can use the formant frequencies from homework 7, or others. Take bandwidths as 10% of the formant frequency values. Because of the many steps, it will be easiest to do everything in a script; this also makes your vowel generation reproducible. Upload the script.

The remaining assignments concern eSpeak.

Homework 8.2. The eSpeak speech synthesizer in Praat.

Praat's text-to-speech synthesis is based on the open source text-to-speech system *eSpeak-ng*. This supports 101 languages (or varieties), and one of them is Dutch. In Praat you create a synthesizer that is fixed for a particular language and a voice variant. Use the **Create SpeechSynthesizer** command in the **New** > **Sound** menu.

Homework 8.3. Playing text.

Synthesize some sentences in the languages that you know and compare their "quality". Use **SpeechSynthesizer: Play text** (if you try to synthesize multiple lines of text at the same time, it might be comfortable to do this via the script window). If the text is spoken too fast, you can lower the speaking rate (in words per minute) via **Speech output settings**.

Homework 8.4. Playing different voices.

Compare for one language, perhaps Dutch, a number of voice variants. You can create multiple synthesizers in the object list, and easily have them pronounce the same sentence by keeping the **Play text** window open (by clicking **Apply** instead of **OK**).

Homework 8.5. Phoneme output.

With **To Sound**, you can get phoneme information in a TextGrid if you switch on **Create TextGrid with annotation**. View the result.

Homework 8.6. Phonetic input.

Try to improve the output of the synthesizer by adding phoneme information, i.e. try to replace one or more words by their phoneme representation. In the current version of the synthesizer, phoneme info can only be given in so-called *Kirshenbaum notation*, which uses ASCII characters instead of IPA. In **Set text input settings**, choose **Mixed with tags** for the **Input text format**. For example, the sentence "This is some phonetic text input" is in

Kirshenbaum notation “[[D,Is Iz sVm f@n'EtIk t'Ekst 'InpUt]]” (the accents (, and ') mark intonation accents). You can mix text and phoneme representations for complete words. For example, you can input text as “Het [[#v]]as me wel een dag zeg”.

Homework 8.7. SSML input.

If **Mixed with tags** is on, your SpeechSythesizer will understand Speech Synthesis Markup Language (SSML). Copy–paste the following text into a text editor, then from there into the **Play text** or **To Sound** window:

```
<speak version="1.0" xmlns="http ://www.w3.org/2001/10/synthesis"
xml:lang="en-US"> <sentence>
Your order for <prosody pitch="+0.5st" rate="80%">
<say-as type="number"> 8 </say-as>
books </prosody> will be shipped tomorrow. </sentence>
```

The pitch on “eight books” is expressed in semitones; check that you get something really different if you make this value very high or very negative. Also check that changing the rate works as you expect.

Then mark a short text of several sentences, preferably a whole paragraph, with SSML, and have the synthesizer speak it. Because of the long text, you should probably copy–paste from a text editor, or use the Praat script window in this way (make sure to double the string-internal double quotes):

```
Play text:
... "<speak version=""1.0"" xmlns=""http ://www.w3.org/2001/10/synthesis""
... xml:lang=""en-US""> <sentence>
... Your order for <prosody pitch=""+0.5st"" rate=""80%"">
... <say-as type=""number""> 8 </say-as>
... books </prosody> will be shipped tomorrow. </sentence>"
```

The SSML tags are described at http://espeak.sourceforge.net/ssml.html. The SSML standard can be found at http://www.w3.org/TR/speech-synthesis/. Please come up with something nice and copy it into your report.

Homework 8.8. Evaluation

Now evaluate the eSpeak SpeechSynthesizer in Praat. What are its strong points (if any)? What could be improved?

Homework 8.9. Other text-to-speech systems

Find a number of online state-of-the-art text-to-speech (TTS) systems (perhaps AT&T, Festival, Nuance) that use other synthesis methods than the acoustic synthesis of eSpeak, and briefly evaluate these systems. Evaluation criteria could include number of languages, whether they have Dutch, the number of voices to choose from, the intelligibity of the synthesis, its naturalness, the intonation, and whatever more you can think of.