APPLICATIONS OF PHONETICS: SPEECH SYNTHESIS

PRELIMINARIES

For this assignment, you will be listening to and comparing speech produced by three different speech synthesis programs that have demos available on the web. To do this, you will need access to a computer from which you can play sound files. You will get the best results if you can play the sounds from a fast connection (cable, DSL, or ethernet), rather than a dial-up connection.

If you do not have access to a computer with such a connection, you can use a computer in the Instructional Computing Lab (Social Sciences 224), where we’ve held two of our classes. Unless reserved for a class, the lab is open Monday-Friday 8AM-6PM (except Thanksgiving weekend). If you go there to do your work, be sure to choose a computer that has headphones (although I’m pretty sure they all do). If you have problems with the sound, ask the attendant for help (that’s what they’re there for).

You are getting this assignment two weeks in advance, so you can start early if you want. That way, if you run into any technical problems that slow you down, you have plenty of time to overcome them and/or to ask for help.

THE PROGRAMS

The three speech synthesis programs you will be using allow you to type in whatever English text you want and then hear the program’s synthesis of it. As you will see, they differ considerably in the quality of the synthesis they produce.

The programs and their web addresses are:

University of Twente (UOT): http://wwwtios.cs.utwente.nl/say/form
OGI (Oregon Graduate Institute):* http://cslu.cse.ogi.edu/tts/demos/index.html#onlinedemo

*The OGI site says the demo (Festival) is out of order, but only certain parts don’t work; what does work is sufficient for you to complete this assignment.

Each program allows you to select different speakers or voices to produce whatever text you type. On the next page is a chart of the (working) voice choices you will find at each site.
**Site** | **Voice Choices** | **# / Gender/ ID of Voices**
---|---|---
UOT | British, American (select via "dictionary") | 2 female: as, tll; 3 male: mwm5, jph, aec
| | American | 1 male: rab |
| | British English | 1 female: hvs; 1 male: abc |
| | Mexican Spanish | 1 female: bcs; 1 male: axk |
| | German | |
| | AT&T | 3 female: Crystal, Lauren, Claire; 2 male: Mike, Rich |
| | US English | 1 female: Audrey; 1 male: Charles |
| | UK English | 1 female: Rosa; 1 male: Alberto |
| | Latin American Spanish | 1 female: Klara; 1 male: Reiner |
| | German | 1 female: Juliette; 1 male: Alain |
| | French | |

**The Texts**

The chart below provides the text you will type in and which speakers you will use for each text at the different sites.

| Text 1 | Dr. Johnson works at St. Francis Hospital, which is on Magnolia St., but he lives at 2234 Elm Dr. in Tucson. He treats 2234 patients per year. | UOT/American  
OGI/mwm5  
AT&T/Mike |
|---|---|---|
| Text 2 | A bee buzzes, but people ride busses. Would you like a little butter? | UOT/American  
OGI/mwm5  
AT&T/Mike |
| Text 3 | To subject it to rigorous tests is important. To subject one, the experiment was easy. The subject found the experiment easy. | UOT/American, British  
OGI/mwm5, tll, rab  
AT&T/Mike, Crystal, Charles, Audrey |
| Text 4 | Doctor Johnson works at Saint Francis Hospital, which is on Magnolia Street, but he lives at twenty-two thirty-four Elm Drive in Tucson. | OGI/all voices listed in previous chart  
AT&T/all voices |

**The Tasks**

1. According to the chart above, type the texts (one at a time) into the relevant synthesis programs. You should type the texts exactly as they appear, punctuation and all. Then play the speech it produces (see “troubleshooting” on last page of this assignment if you have trouble playing a file).
2. Listen to the synthesized speech for each program and try to notice what sounds good (like a human speaker) or bad (wrong pronunciations, difficult to understand, not human-like) for each text and each program.

3. Take detailed notes on which sounds and words come out good or bad, and what’s wrong when they’re bad.

4. For Text 3, also notice what sounds better and worse for the male and female and British and American voices. For Text 4, notice what sounds different for each voice.

Questions to Answer

5. (5 points). Overall, which synthesizer sounds most human-like? Which sounds least human-like?

6. (15 points). Text 1 contains several abbreviations and numbers. For Text 1, please answer the following questions:
   - In general, what difficulties do the abbreviations and numbers create for a synthesis program?
   - More specifically, which program or programs made errors on the abbreviations and/or numbers, and what were the errors? Be specific about which programs made what errors.
   - What information might a synthesis program have to have to get the abbreviations and numbers right?

7. (10 points). In Text 2, there were voiced and voiceless fricatives (/z/ in “buzzes” and /s/ in “busses”) and flaps (in “little” and “butter”). Both types of sounds are typically as tricky to synthesize as they are acoustically and aerodynamically tricky to produce. For Text 2, please answer the following questions:
   - In general, how well did each program do with the fricatives and the flaps? Do the words sound natural? If they didn’t, what sounded wrong?
   - Within a single program, do “buzzes” and “busses” sound different from each other? Do the flaps in “little” and “butter” sound different within the same program? Describe what you hear for /z/ vs. /s/ and the two flaps in detail (e.g., is one more voiced than the other? Anything else you notice?).
8. (20 points). Text 3 has the word “subject” used three times. In two instances the word is used as a noun, “subject”, with stress on the first syllable. In one instance the word is used as a verb, “subjéct”, with stress on the second syllable. With respect to Text 3, please answer the following questions:

- Which of the three sentences in Text 3 did each of the synthesizers get right?
- Why do you think the synthesizers make the mistakes they do on these sentences? That is, what rule might each synthesizer be using to determine where the stress should be in the word “subject,” and why do the rules some of them use fail on some of the sentences? (Hint: what rule, if any, does each program seem to use to decide what’s a noun and what’s a verb?)

9. (20 points) Also for Text 3, please compare the specified male, female, American and British voices for each synthesizer, and answer the questions below.

- AT&T and OGI both have male and female American voices. For each site, which sounds better - the male or female American voice? Which sounds better overall?
- Do you think any of the three programs succeed in achieving really British-sounding speech? Really American-sounding speech? Which is more convincing, the British or the American?
- Do any of the synthesizers produce mistakes in one voice that they don't produce in another? If so, (and there's at least one) list these mistakes.

10. (20 points). In Text 4, all the abbreviations and numbers from Text 1 are spelled out (and the final bit of Text 1 is removed). You also used a number of different voices in Text 4. For Text 4, please compare the various voices of AT&T and OGI and answer the following questions:

- What differences do you notice in general voice quality or synthesis quality among the various AT&T English voices? Among the various OGI English voices?
- Several people who have listened to “Mike” and “Rich” at the AT&T site, think that “Mike” is White, and “Rich” is African American. Do you share this impression? If you think the two synthetic speakers represent different racial backgrounds, what do you think you are hearing that leads you to that conclusion? (I just want you to speculate about this; I don't know the real answer.)
- What happens when you run the English text through a synthesizer intended to synthesize a different language (French, Spanish, or German)? How would you describe the resulting speech? Do you get the same impression from AT&T non-English voices and OGI non-English voices?
TRoubleshooting

- If the synthesized speech won't play, it could be that the text is too long, in which case you can break it up into smaller parts. (I did not have this problem when I did it, but some of the programs are buggy and this may be one cause).

- Another reason that the synthesized speech may not play is if you have a program on your computer that blocks pop-up windows. This may prevent some of the necessary plug-ins from automatically engaging and so the sound won't play. You will need to temporarily disengage that software for these sites in order for the files to play.

- If the synthesized speech is played over a network (or worse, over a modem), you may get pauses in the middle, making it sound choppy. This is not a problem with the synthesizer, but with the connection.

- At the AT&T site, if you get the sound file but when you click the “play” button you don’t hear anything or only a stutter, you may want to try going back to where you typed the text and click the box underneath that says “display a link for the audio file”. Then resubmit the text. You should be directed to a page that says “your audio file can be found at (website address)”. If you play it from the link listed, it sometimes works better.

- If you hear consistent little gaps or pauses, it’s probably an audio problem, not a problem with the synthesizer.

- Each of the programs can be a little buggy, so you may have to try a couple of times to complete the whole assignment, possibly rebooting (or switching computers, if possible) as you do it.