Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Preliminary observations on discontinuities
Lindh, J.

journal: Proceedings of Fonetik, TMH-QPSR
volume: 44
number: 1
year: 2002
pages: 113-116

http://www.speech.kth.se/qpsr
Preliminary observations on discontinuities

Jonas Lindh
Department of Linguistics, Gothenburg University

Abstract
One of the problems with concatenating speech units are the discontinuities or ‘jumps’ that appear when a not so successful concatenation between units has been made. These preliminary observations deal with that problem by acoustically investigating the discontinuities in two TTS systems. The major question is what parameter it is that mainly fails when a discontinuity is perceived and whether it is possible to integrate this acoustic/phonetic knowledge into systems to avoid discontinuities.

The observations made so far, has shown that discovering discontinuities in formant frequencies (especially F2 and amplitude) seems to be a problem that can be solved by assigning more importance to acoustic parameters when selecting units to connect.

1 Introduction
Lately commercial unit selection systems have started to get a more natural prosody. With large lexicons and considerable work put into recordings and labelling there has been some success for systems like L&H’s RealSpeak and AT&T’s NextGen. However, the transitions between units connected often contain discontinuities. These discontinuities are often perceived as ‘jumps’ in the output.

Normally in unit selection systems, a large speech database is used containing several thousand units. The selection is usually made by a dynamic programming search to optimise a cost function given a target specification (how the spectra of a given sentence should look), a concatenation cost, which usually include some kind of spectral continuity cost and sometimes a pitch continuity cost. Measures used in these kinds of systems are often based on distances between cepstra or mel-cepstra, but not much work seems to have been done on investigating the usefulness of the measures selected. (Donovan, 1998)

In order to develop unit selection systems further phonetic/acoustic knowledge needs to be applied. An investigation of the discontinuities in the systems output might indicate which acoustic parameters that seems to fail mainly. If these parameters can be identified it might also be possible to integrate this knowledge into the unit selection process.

The Investigation
I have chosen to work with the two commercial systems RealSpeak and AT&T NextGen, because of the particularly natural sounding output (according to the authors subjective opinion). For both systems the female voice for American English has been used for the analysis.

By using the interactive demos (type-and-talk) that are available for each system on their official websites (www.naturalvoices.att.com/-demos/www.lhs1.com/realspeak/demo) several sentences has been tested. The sentences where obvious discontinuities were perceived were then downloaded with a sample rate at 22 kHz and then analysed with the programs Praat and KTH Wavesurfer (www.praat.org www.speech-kth.se/wavesurfer).

So far only a few sentences have been analysed for which F0, F1, F2, F3 and amplitude have been measured. The values have been measured through the sounds where the discontinuities were perceived, and some of the results are presented further on in this paper.

The Two Systems
NextGen is a system developed within the Festival framework (CSTR, Univ. Edinburgh, Scotland). Text normalization, linguistic processing such as syntactic analysis, word pronunciation, prosodic prediction (phrasing and accentuation), and prosody generation (translation between a symbolic representation to numerical values of fundamental frequency F0, duration, and amplitude) is done by a Flextalk object that borrows heavily from Bell Labs’ previous TTS system,
Flextalk. From ATR’s CHATR system, the online unit selection (with modifications) is adopted. Speech is synthesized using one of many possible back-ends, including AT&T Harmonics plus Noise Mod (HNM) synthesizer. (Beutnagel et al., 1998)

CHATR uses phonemes as units, but the NextGen team has modified the CHATR system and use half phones instead. (Conkie, 1999)

In RealSpeak, the units (diphones) are scored with a cost according to their prosodic/phonetic mismatch with the target description of the utterance to be synthesized. The prosodic/phonetic cost is computed on the basis of a combination of symbolic and numeric features. The candidate units from the speech database are then evaluated for the ease with which they can be concatenated. By using additive costs to a dynamic programming algorithm the optimal path of candidates are chosen that best represent the spoken utterance. (Coormat et al., 2000)

2 Results and discussion

F0 does not often seem to be the problem of discontinuities, even though many people working with TTS systems (including the author), seem to believe that that is the largest problem in the unit selection process. In the sentences analysed so far, only the short sentence “Ha ha” had a discontinuity mainly depending on the F0 parameter. The reason for this is that both systems have a target prosody which is based on high F0 in the beginning of sentences and low in the end. This seems logical, but in the mentioned example it gives a discontinuity in F0 leading to a difference from 216 Hz to 158 Hz over 60 ms for RealSpeak.

Figure 1. F0 of “Ha ha” NextGen

Figure 2. F0 of “Ha ha” RealSpeak

While listening to the sentence “I hardly hear you Heather” the first thing discovered was that both systems seemed to have discontinuities in the exact same positions, [ɪ. ju] and [hɛ]. When analysing it in spectrogram, the discontinuity in RealSpeak’s output was clear in F2 (for [hɛ]) where F2 goes from 1251 Hz to 1818 Hz within 15 ms, and amplitude (for [ɪ. ju]). In NextGen, though, there were no discontinuities found. After a second listening no discontinuities were perceived in NextGen’s output either, which shows that perception can fool even a trained ear (the author has at least three years of experience in listening to discontinuities in speech synthesis).

Figure 3. Formant speckle diagram of [hɛ] RealSpeak

Figure 4. Formant speckle diagram of [hɛ] NextGen
NextGen, however, has discontinuities in F2. In the sentence “Can you please show me the way to the toilet sir?” F2 drops from 2440 Hz to 1400 Hz in less then 20 ms.

RealSpeak does not have the same problem in that sentence (see figure 8). The RealSpeak system uses a different target prosody for questions, though, which results in a strange F0 raise in the end of the same sentence. (Coormat et al., 2000)

These are just a few examples analysed so far, but it presents the way the analysis is moving forward and how it is done. There are still a lot of sentences to be analysed, but these examples show that the systems still have difficulties in finding the right segments for connecting cer-
tain parameters (especially F2 and amplitude) properly.

3 Conclusions and ongoing work

The analysis described in this paper shows just a few examples of what has been analysed so far.

A conclusion that would seem obvious is that if more phonetic knowledge is put into the unit selection procedure there will be fewer discontinuities in systems like these. However, the only conclusion I can make so far is that discovering discontinuities in formant frequencies (especially F2 and amplitude) seems to be a problem that can be solved by assigning more importance to acoustic parameters when selecting units to connect.

One of the next steps will be to look into how discontinuities in F3 are perceived, which seems to be common, often depending on neighbouring lip protrusion in segments connected to units selected from a different environment.

Amplitude seems to be a bigger problem than I initially thought, but to draw any conclusions out of this small amount of data would not be correct.

This paper is only a short presentation of an analysis that has just started. Further on more data will be collected and suitable tables made up on which to draw further conclusions. Nevertheless, this pilot study shows that the failing parameters are easily detectable. By comparing results from a large amount of data, it is possible to determine which parameters that fail most frequently. Thus, it might be possible to create smoother concatenations and to minimise discontinuities.

References


