

Acoustic and perceptual analysis of discontinuities in two TTS concatenation systems

Jonas Lindh

Göteborg University

Department of Linguistics

Background Discontinuities

It is fair to say that L&H's (now Scansoft's) RealSpeak and AT&T's NextGen are two of the most natural sounding unit selection systems.

The transitions between connected units sometimes contain discontinuities, thus creating one of the greatest problems concerning the output in these kinds of systems. The discontinuities are often perceived as 'jumps', i.e. a disturbance. The analyses in this paper investigate the acoustic properties of the 'jumps', if they are perceived as disturbing and in that case how disturbing.

The results show that the selection criteria do not include enough information on single acoustic parameters, such as formants. Since listeners perceive discontinuities in formants, especially F2, as disturbing, one of the conclusions is that the next step in developing these systems must be to include more information on these parameters separately (especially formants 2 and 3) to improve the selection process. Of course other things like increasing database size and better structuring of data etc. can also improve the selection process as well as better grapheme to phoneme conversion, but those aspects are not dealt with here.

Introduction

As a follow-up to the author's earlier published "Preliminary Observations on Discontinuities" (Lindh, 2002), a presentation of the outcome from the complete analyses is presented here.

The paper is a report from an investigation of the problems that appear when trying to concatenate units to create a natural sounding Text-to-Speech system. The focus is on finding cues to where, acoustically, discontinuities occur and how well test subjects perceive them and how disturbing they find discontinuities to be as a function of various acoustic features.

The most recent research on the subject has been concerned with finding new spectral distance measures for the unit selection process or finding new ways to smooth the fundamental frequency to avoid discontinuities. The analy-

ses show that the factor that is perceived as most disturbing is a discontinuity in F2 and that most discontinuities in F0 are neglected by subjects.

Some improvements have been discovered concerning new distance measures though (Donovan 2001, Vepa et al., 2002, Stöber et al., 2001, Plumpe et al., 1998), but most of them are shown to be insufficient in handling all kinds of discontinuities (Klabbers and Veldhuis 2001).

The Two Systems

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 in a dynamic programming algorithm the path of candidates is chosen that best represent the spoken utterance (Coorman et al., 2000).

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 and numerical values for 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 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 uses half phones instead (Conkie, 1999 and Black et. al. 1994 and 1997).

Discontinuities

A simple way of describing a discontinuity in a speech synthesis system would be to say that there is something in the output signal that could not easily or naturally have been produced by a human speaker. Another definition is used however for the purposes of the following analyses: A 'spectral' discontinuity was defined as an abrupt change in one or more of the acoustic parameters where such changes are not expected in normal speech. Two ways of quantifying discontinuities were tried – the absolute magnitude of the change and the mean differential of the parameter during the change. Figure 1 shows an example of a typical F0 discontinuity. F0 changes direction abruptly and rises from 177 Hz to 183 Hz within a short time span (10 ms). The absolute change ($\Delta F0$) is thus 6 Hz and the differential ($dF0/dt$) 0.6 Hz/ms.

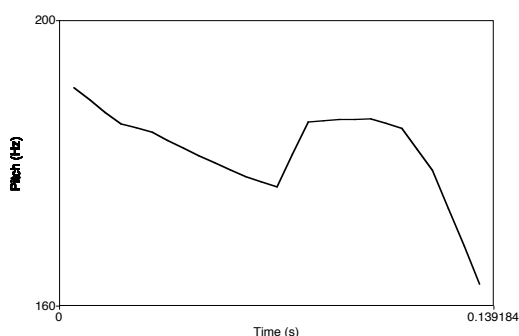


Figure 1. The diagram shows a typical F0 discontinuity.

Method

The method adopted is based on finding the source of the discontinuities perceived while testing the two systems acoustically and perceptually. At first, large sets of test sentences were synthesized and then analysed acoustically to discover discontinuities. The same sentences were then used in a listening test (described below) where test subjects had to detect discontinuities and give a score representing how disturbing they found each discontinuity. The sentences were then reanalysed acoustically measuring F0 (fundamental frequency), formants (1, 2, 3 and 4) and intensity (SPL) in the areas where discontinuities were perceived. All positions where subjects detected something were analysed and when it was possible

measurements for change per unit were calculated.

Material

For both systems the female voices for American English were used.

The sentences were chosen randomly at first and then picked from several different evaluation tests (Allen et al., 1987) to get a reasonable amount of material to work with and avoid semantically predictable sentences.

Sentences were downloaded at a sample rate of 22 kHz and then analyzed with the programs Praat and KTH Wavesurfer. (<www.praat.org> <www.speech.kth.se/wavesurfer>)

Listening Test

Ten subjects were presented a paper copy with information, one example and the test. The sentences were written on the sheet in random order. The subjects were then told to read each sentence first and then listen to it and underline the part where they could perceive some kind of discontinuity. To be able to evaluate how severe the perceived discontinuity was the subjects had to give a score, similar to the MOS scale (Goldstein, 1995), underneath the underlined part to describe how disturbing they found each discontinuity.

In the test 38 sentences were presented. Each sentence was only played once and the subjects were told that it was better to underline more than less if he/she was uncertain about the location of the perceived discontinuity.

Subjects all had an academic background in linguistics/phonetics and a good knowledge of English, even though their first language is Swedish. None of the subjects have any known hearing disorders and ages ranged between 25 and 55.

Results and Discussion

While there are several problems involved in the acoustic measurements there are interesting figures concerning the validity of specific parameters involvement in the disturbance of the synthetic speech output.

There were many discontinuities in the F0 parameter acoustically, but most of them went unnoticed by the listeners or were given low disturbance scores. In everyday speech fundamental frequency changes rapidly depending on context, syllable and phrase position and stress.

Statistical analysis

Several statistical tests were performed in order to evaluate the contributions to perceived disturbance by the various acoustic discontinuity measures. Multiple regression analysis was used with the mean disturbance scores as the dependent variable and discontinuity functions of F0, F1, F2, F3, F4 and SPL (as described above) as independent variables. In half of the tests, F0–F4 was expressed in Hz and in the other test F0 was expressed in semitones and F1–F4 in Bark. The space available in this report does not allow a presentation of all the results of these tests.

Table 1. Model Summary. The result of a regression analysis. Explained variance using these parameters is 73%.

	Unstandardized Coefficients	Standardized Coefficients	t	Sig.
	Raw	Beta		
(Const.)	.455		8.570	.000
$\Delta F0$	8.593E-03	.121	1.701	.094
$\Delta F2$	1.662E-03	.474	5.955	.000
$\Delta F3$	2.409E-03	.300	4.049	.000
$\Delta F4$	1.692E-03	.287	3.405	.001
ΔSPL	1.692E-02	.090	1.295	.201

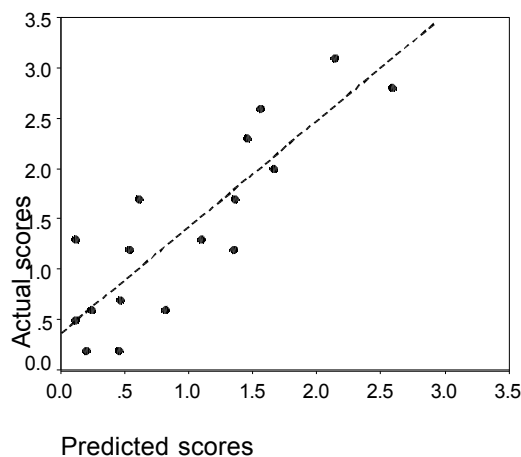


Figure 2. Visual representation of the prediction by given parameters compares to the actual scores given by subjects.

Only the combination of parameters and units that best fit the data will be presented. It turned out that using the absolute values and expressing the frequencies in Hz gave the best fit to the data. The results of one such analysis is presented in Table 1. The explained variance of the model

presented in the table is 73%. The reason for including ΔSPL although its contribution is not statistically significant is that it was used in an attempt to model the data that gave slightly better results with this parameter included. Model and data are presented in Figure 2. All factors do not contribute equally to the disturbance scores. In a stepwise regression analysis model 1, $\Delta F2$ alone, explains 49% of the variance. Adding $\Delta F3$ to the model raises the explained variance 64%. Including $\Delta F4$ raises explained variance to 71%. Adding $\Delta F0$ and ΔSPL only adds another 2%.

Conclusions

The next step in developing these systems must then be to differ among the parameters, and develop tools for how to include them into the different algorithms used to choose good concatenation segments when not adjacent in the speech database. This probably includes an increase of the database size, but also development of better formant trackers and new ways to handle prosody as something more than stylized pitch movements.

Formant frequencies are connected to the phonological processes surrounding them. Introducing them into a different environment than from where they were taken will be perceived as unnatural.

Another approach could be to increase the database size for vowels and decrease it for several voiceless segments which are more or less constant and much more difficult to perceive discontinuities in. One could start with developing costs where concatenation within vowels are more or less forbidden, unless absolutely necessary, and at the same time decrease importance of F0 continuity, since listeners obviously do not take much notice of discontinuities in this parameter.

Klabbers and Veldhuis (2001) reported some success in decreasing the amount of spectral discontinuities by using results from a listening test to detect discontinuities and then “extending the diphone database with context-sensitive diphones to reduce the occurrence of audible discontinuities”. That is an option, but it does not deal with the problem of increasing database size into infinity, which only can be eliminated by finding appropriate measurements for each sensitive parameter individually. The two authors also made tests with all of the best performing distance measures, including:

- Euclidean distance between (F1, F2) pairs, or the Euclidean formant distance, which is often used in phonetics.
- Symmetrical Kullback-Leibler distance, which originates from statistics.
- Partial Loudness, which comes from the area of sound perception.
- Euclidean distance between Mel-frequency cepstral coefficients, which comes from automatic speech recognition.
- Likelihood ratio, which is used in speech coding and automatic speech recognition.
- The mean-squared log-spectral distance, which also comes from automatic speech recognition.

Kullback-Leibler showed to be best at predicting audible discontinuities. (Klabbers and Veldhuis 2001)

None of the distance measures will solve the problem though, since the problem has reached a dead end. At this point efforts should be made to divide the different parameters instead of squeezing them into one single distance measure. All of them are discovered to produce discontinuities. To avoid that, proper formant trackers must be developed (for a lot of other reasons as well) and the distance measures evaluated to find out how and how well they correlate with human auditory perception.

References

- Beutnagel M., Conkie A., Schroeter J., Stylianou Y. and Syrdal A. (1998) The AT&T Next-Gen TTS System. Joint Meeting of ASA, EAA and DAGA, 18-24
- Black A., Taylor P. (1994) CHATR: A Generic Speech Synthesis System. COLING94, Japan.
- Black A., Taylor P. (1997). Festival Speech Synthesis System: System Documentation (1.1.1). Human Communication Research Center Technical Report HCRC/TR-83.
- Conkie A. (1999) A Robust Unit Selection System for Speech Synthesis. Proceedings of 137th meet. ASA/Forum Acusticum, Berlin
- Coorman G., Fackrell J., Rutten P. and Van Coile B. (2000) Segment Selection in the L&H RealSpeak Laboratory TTS System. Proceedings of ICSLP, 2:395-398
- Donovan R.E. (2001) A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesizers. Proc. 4th

ESCA Tutorial and Research Workshop on Speech Synthesis, Atholl Palace Hotel, Scotland, UK.

Klabbers E. and Veldhuis R. (2001) Reducing Audible Spectral Discontinuities. IEEE Transactions on Speech and Audio Processing, 2001, 9(1), 39-51

Lindh J. (2002) Preliminary Observations on Discontinuities. TMH-QPSR Vol. 44 – Fonetik 2002

Plumpe M., Acero A., Hon H. and Huang X. (1998) HMM-based smoothing for concatenative speech synthesis. Proc. ICSLP, 6:2751-2754

Stöber K., Wagner P., Klabbers E., Hess W. (2001) Definition of a training set for unit selection-based speech synthesis. In SSW4-2001, paper 118.

Vepa J., King S. and Taylor P. (2002) Objective Distance Measures for Spectral Discontinuities in Concatenative Speech Synthesis, in Proc. ICSLP 2002, Denver, USA.