Comparing Spectral Distance Measures for Join Cost Optimization in Concatenative Speech Synthesis

Ingmund Bjørkan, Torbjørn Svendsen, and Snorre Farner

Department of electronics and telecommunication, NTNU
O. S. Bragstads plass, N-7491, Norway
{ingmund,torbjorn,farner}@iet.ntnu.no

Abstract

In concatenative synthesis the join cost function can be related to the probability of a perceived discontinuity at the join. Therefore it is important that the distance measures in the cost function correlate highly with human perceived discontinuities. In this paper the results of a listening test on joins in two Norwegian long vowels: /A:/ and /e:/, is presented. Five spectral distance measures and the F0 difference are compared as predictors of the human perceived discontinuities using Receiver Operating Characteristic (ROC) curves. In addition, a linear join cost function is optimized by means of stepwise linear regression.

1. Introduction

Unit selection systems are considered state of the art in text to speech synthesis, with capability of producing highly natural-sounding speech. Unit selection synthesis is based on concatenating small units of speech, selected from a large database containing multiple candidates for each unit. The search for the optimal unit sequence is normally based on a combination of two cost functions: target cost and concatenation cost [1]. Target cost measures how well a unit matches prosodic and phonetic features of the target, while the concatenation cost is a measure of how well two neighboring units can be concatenated. The optimal unit sequence with respect to these two cost functions can then be found by a Viterbi search for the lowest cost path through a lattice, where the database units are the nodes with an associate target cost, and the concatenation cost is the cost of the path between two nodes.

One problem with unit selection synthesis system is the large variability in quality, varying from almost perfect speech to very poor quality speech with many disturbing discontinuities. To improve quality we want the concatenation cost to have a high correlation with human perception of discontinuities at concatenation points. The concatenation cost has to be measured with distance measures on physical measurable properties such as spectral parameters, F0 (pitch) and power. The probability of perceived discontinuity at a concatenation point can be estimated as a composite function of the distance measures. Defining the concatenation cost as a function proportional to the probability of perceived discontinuity then gives a cost function $C(d_1(x_1), d_2(x_2), \ldots, d_n(x_n))$, where $d_i(x_i)$ is a distance measure on the i'th feature. A common approach is to use a linear cost function $C = \sum_i w_i \cdot d_i(x_i)$, where the weights w_i are the relative importance of each feature [1]. How well the cost function works would mostly depend on how many orthogonal features that are relevant and of how well the distances on the features discriminate good and bad joins.

In this paper we describe a preliminary perceptual experiment on the detection of discontinuities in two Norwegian vowels, and compare how well different objective distance measures correlate with the human perceived discontinuities. In addition to spectral distance measures, F0 difference was also considered as some stimuli had relatively high F0 difference at the concatenation point.

The paper follows earlier studies ([3], [4], [5]) of comparing different distance measures of spectral distance. As in [4] and [5] the problem is approached as a binary signal detection problem using ROC curves [2] to compare the distance measures.

2. Perceptual experiment

A perceptual experiment was conducted on listeners detection of concatenation discontinuities in vowel joins generated by concatenative synthesis from a Norwegian female speaker. The stimuli consisted of one vowel join inside a test word which was wrapped in the middle of a relatively short sentence. Joins in two Norwegian long vowels, A and e, were tested (SAMPA */A:/* and */e:/*)

The listening test was conducted as a binary forced choice experiment, where listeners had to make a forced decision whether a join were discontinuous or not. 20 adult volunteer listeners, most of them employees at the Signal Processing Group at NTNU, participated in the test.

2.1. The Prosdata Synthesizer

The synthetic voice in the experiment was generated from the speech database PROSDATA [6], which is a collection of 502 Norwegian sentences read by a female speaker. The data have been manually segmented in terms of phonemes, syllables and words, and does also include F0 values and root mean square energy values. A simple automatic demiphone segmentation was conducted to allow for a demisyllable synthesizer, which was used to generate the test stimuli.

The synthesizer is intended for experimental use and is implemented as a waveform-synthesizer in Matlab using java to communicate with a mysql database. The synthesizer uses crosscorelation of two single period length windows at each side of the concatenation boundary to estimate phase lag between two voiced units. This estimate is then used to move the concatenation point so that phase discontinuities are avoided. For voiced sounds the synthesizer also performs a crossfade over two periods around the concatenation point to avoid clicks and to smooth the join. The idea is that testing joins should be done under the same conditions as the synthesizer normally uses.

2.2. Test stimuli

Test stimuli were generated by selecting sentences from the database containing the wanted vowel. The words containing the vowel were restricted to not being in sentence initial or final position, as the original sentence or a part of the original sentence was used as a wrapper sentence around the testwords. In order to make a join at the middle of the vowel, the demisyllable starting at the middle of the vowel was replaced by other instances of this demisyllable from the database.

This gave two joins: The first join at the middle of the vowel, and a second (unwanted) join at the phoneme border between the syllable containing the vowel and the next syllable. A bad second join could influence the listeners choice, and is a noise source in the experiment which we attempted to minimize by restricting which types of sounds that were allowed in the second join. Testwords were restricted to words with at least one consonant at each side of the join and as the main rule, the second join should have an unvoiced consonant on one side.

Candidates for replacing a demisyllable were found by using the synthesizer's target cost function, using the prosodic parameters of the original unit as target values, and then choosing candidates with target score higher than a chosen threshold. The stimuli was then generated by using the synthesizer's concatenation methods described above.

To avoid discontinuities due to rms difference at the join, the unit candidate to the right of the concatenation border was scaled so that the mean rms energy in the vowel parts were equal.

2.3. Test Procedure

The test consisted of 210 stimuli with /e:/ and 248 stimuli with /A:/. These stimuli were split in 4 blocks with 5 listeners on each block. Each block consisted of two sessions, One session with /e:/ stimuli and one with /A:/ stimuli. All listeners at the same block listened to the same stimuli, but with different randomization of the stimuli within each session. Each session consisted of 65 stimuli. The ten first stimuli of the /e:/ sessions was a familarization phase, where examples of two good joins, three bad joins, and five practice stimuli were presented to the listener. Two or three original sentences was also added to each session. Neither the ten first stimuli of the /e:/ session nor the original sentences was used when analyzing results. The listening test was presented using a graphical interface where listeners could press a button to play or repeat a stimulus. The listener could repeat the stimulus as many times as they wanted, before they had to make a forced decision whether the join was dicontinuous or not.

All participants in the test were given an instruction before the test started. Listeners were instructed to focus their listening on the word containing the join, and ignore possible discontinuities elsewhere in the sentence. For each stimulus the testword containing the join was shown in the graphical interface with its SAMPA transcription in paranthesis, so that the listeners could know where to concentrate their listening.

The test was performed on the same computer with the same headphones by all listeners, and took normally about 20 minutes for completion. The consistency between listeners at the same block was on average 72.7% for /e:/ and 68.6% for /A:/. This indicates that the test was a difficult task, although many of the inconsistencies could be explained by different critical levels among listeners. Totally there were 4200 observations on /e:/ stimuli and 4960 observations on /A:/ stimuli.

3. Distance Measures

3.1. Spectral distance measures

The spectral distance measures used were:

1. Symmetrical Kullback-Leibler distance evaluated on LPC power spectra, [7]. Given two LPC spectra $P(\theta)$ og $Q(\theta)$ the measure is defined as:

$$D_{skl} = \frac{1}{4\pi} \int_0^{2\pi} (P(\theta) - Q(\theta)) \log \frac{P(\theta)}{Q(\theta)} d\theta \quad (1)$$

2. Cepstral distance, euclidian distance of a truncated series of cepstral coefficients evaluated from LPC power spectra

$$D_{cep} = \frac{1}{N} \sum_{i=1}^{N} (c_P(i) - c_Q(i))^2 \quad (2)$$

D_{cep} is an approximation of the rms log spectral measure [8].

$$D_{cep} \approx \frac{1}{2\pi} \int_0^{2\pi} |\log P(\theta) - \log Q(\theta)|^2 d\theta \quad (3)$$

3. D_{LR}: This measure is the mean of the nonsymmetrical likelihood ratios calculated from power normalized LPC spectra. The measure is highly correlated with D_{cep} with a correlation coefficient of 0.97 [8].

4. D_{mfcc}: Euclidian distance between Mel-frequency cepstral coefficients [9], using an implementation with 13 coefficients. The first coefficient representing energy was not used.

5. D_{mpsc}: Modified pitch synchronous crosscorrelation, described in more detail in the next section

The spectral distance measures above, except D_{mpsc}, were calculated from parameterizations of speech using two period length windows. All LPC-spectra were calculated using 16 LPC coefficients.

3.2. Modified pitch synchronous crosscorrelation

The distance measure is calculated using exactly one energy normalized period from each side of the concatenation point. The last period of the unit to left of the concatenation point, $x(n)$, and the first period of the unit to the right of the concatenation point, $y(n)$. If the length of $x(n)$ and $y(n)$ are equal, the distance measure is identical to the measure suggested in [10] defined as the Euclidian distance between two pitch synchronous energy normalized signals $x(n)$ and $y(n)$. If $x(n)$ and $y(n)$ are of different lengths, a modification is done by interpolating $y(n)$ to the length of $x(n)$. Denoting the interpolated version of $y(n)$ for $y'(n)$, D_{mpsc} is defined as

$$D_{mpsc} = \frac{1}{N} \sum_{n=1}^{N} (x(n) - y'(n))^2 = 2(1 - C_{xy'}(0)), \quad (4)$$

where $C_{xy'}(0)$ is the crosscorrelation of x and y' at lag zero. The intention of the modification was to make the distance measure more orthogonal to the F0 feature. The method introduces

some distortion of frequencies, but if othogonality to the F0 feature is wanted some modifcation is necessary. Looking at vowel joins in natural sentences, this modification of the distance measure reduced the mean of the distance measure by 68%. Intuitively this distance measure will depend mostly on differences in low frequencies. From calculations of correlation with the other distance measures, it was most correlated with D_{skl} with an estimated correlation of 0.70.

4. Results

Letting class ω_0 represent perceived discontinuities and class ω_1 represent perceptually good joins, we have a two-class pattern recognition problem with several feature vectors x_i representing the distance measures. Although distance measures are not Gaussian distributed. ROC curves are informative describing the different distance measure's separability of the two classes. The ROC curves plot the hit rate $P(hit) = P(x > x^*|\omega_0)$ on the y-axis and the false alarm rate $P(false\ alarm) = P(x > x^*|\omega_1)$ on the x-axis. From this definition $P(hit)$ is defined as the rate of successfully detecting discontinuities, and $P(false\ alarm)$ as the rejection rate of good candidates.

The results were analyzed in two steps:

- In the first step all data in the listening test are analyzed.

- In the second step a minimum error rate Bayes classifier [2], $\max_i p(\omega_i|D_{F0})$, was used to remove stimuli with probability of discontinuity due to F0 difference higher than 0.5. Then the different spectral measures were studied from the remaining data. The classifier removed the data where the log F0 difference was higher than 0.067 for /e:/ and higher than 0.062 for /A:/. This implied that 36.7% of the /e:/ stimuli and 30.2% of the /A:/ stimuli were removed from the original data.

D_{LR} was not plotted as it correlates highly with D_{cep}, and the ROC-curve for this distance measure would therefore approximately follow the same ROC-curve as for D_{cep}. Root mean square energy difference was also tested, but had no correlation with perceived discontinuities, indicating that the smoothing of energy at joins was successful. In addition to the distance measures, a weighted linear combination of distance measures C_{lin} was plotted. The model and weights of the linear combination was optimized by using stepwise linear regression as explained in section 4.1.

From the ROC curves in the figure 1, we see that the F0 difference is the best predictor of discontinuities, except for C_{lin}. Although detection with the F0 difference for the vowel /A:/ is lower, we cannot necessarily state that F0 differences are less important in this case. The detection rate would also be dependent of the number of stimuli with a relatively high F0 difference in the data set. For the vowel /A:/ there were more possible candidates to choose from in the database when generating the stimuli, and therefore the mean F0 difference in the /A:/ became lower, leading to fewer stimuli with high F0 difference in the test. This describes that a feature's discriminability of the two classes will be correlated with the number of audible discontinuities that were caused by this feature. Intuitively the results then could be highly dependent of the synthesis system and the test design.

Although the whole ROC curve is of interest, the hit rate at high false-alarm rates are especially interesting as it corresponds to small distances x^* in the definition of $P(hit)$. In a unit selection point of view a high false alarm-rate corresponds to the case of choosing the assumed best unit out of many good candidates.

An interesting observation is that when high F0 differences were removed, see figure 2, the detection rate at high false-alarm rates was constant or increased for D_{cep} and D_{mfcc} for both /e:/ and /A:/, while the hit rate of D_{skl} and D_{mpsc} were reduced. This indicates that the conditional probability of discontinuity $P(D|D_{cep})$ and $P(D|D_{mfcc})$ may be more orthogonal to $P(D|D_{F0})$ than the other distance measures. Overall, D_{mfcc} seems to be the best performing spectral measure, while D_{cep} is equally good or better for high false alarm rates.

4.1. Stepwise linear regression

The detection rate is not the only important factor for choosing a good cost function based on the distance measures. Also orthogonality of the distance measures in the cost function would be important. Deciding the weights and which distance measures to include in the cost functions, could be done by stepwise linear regression [11]. This method assumes a linear model of the input parameters and enters parameters one by one as long as new parameters significantly reduce the error variance. A hypothesis test on significance level α is used to decide if a new parameter should be entered into the model. Using stepwise linear regression on all the data resulted in a join cost function C_{lin}, containing D_{F0} and D_{LR} for /e:/ on a significance level $\alpha = 0.05$. Both D_{mfcc} and obviously D_{cep} could be used with almost the same performance. When increasing the significance level to $\alpha = 0.10$, also D_{mpsc} was entered into the model. Although D_{mpsc} was the worst performing distance measure in this test, probably the orthogonality to the two other measures could give a small improvement. For /A:/ D_{F0} and D_{mfcc} was chosen at $\alpha = 0.05$, while D_{skl} was added at $\alpha = 0.10$.

Another join cost function C'_{lin} was caluated for the data set where high F0 differences where removed. For this cost function only two parameters were significant at $\alpha = 0.10$: D_{mfcc} and D_{F0} for /e:/, and D_{mfcc} and D_{skl} for /A:/. The linear combinations were most successful when there were many discontinuities with high F0 differences, while a linear combination of the spectral distance measures didn't give much improvement.

4.2. Transformation of distances

In general the join cost function is dependent of the conditional probabilities $P(D|d_i(x_i))$, so if the probabilities are nonlinear functions of the distances d_i and they can be estimated reliably, a better performance can be achieved by doing a transformation of the distances before optimization. Probably a join cost function would also generalize better if transformed distances are used. For such a transformation to work, it would be important to estimate $P(D|d_i(x_i))$ reliably. As an experiment the conditional probabilties were estimated from the data by using Bayes formula [2]. The method gave only small changes, due to that estimates $\hat{P}(D|d_i(x_i))$ were almost linear. With more reliable estimation of the probabilities the method may give more improvement.

5. Conclusions and Further Work

In this paper five different spectral distance measures and F0 difference was tested as detectors of human perceived distcontinuities in two Norwegian vowels. The results were analyzed in two steps where data with high F0 difference where removed before analyzing the spectral distance measures. Overall, F0
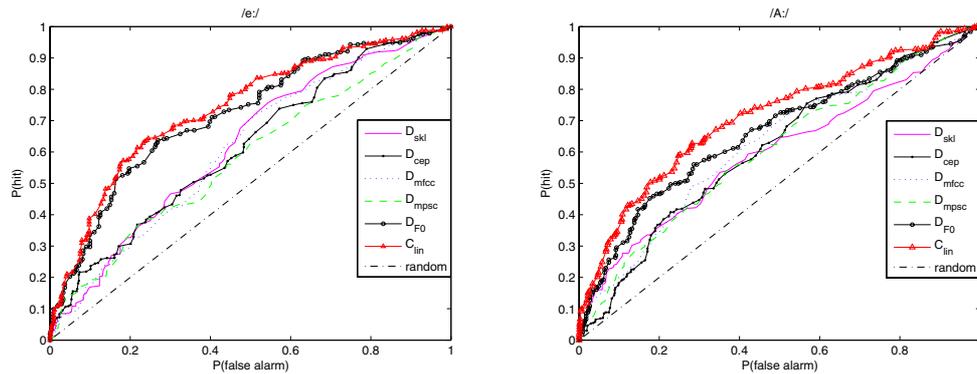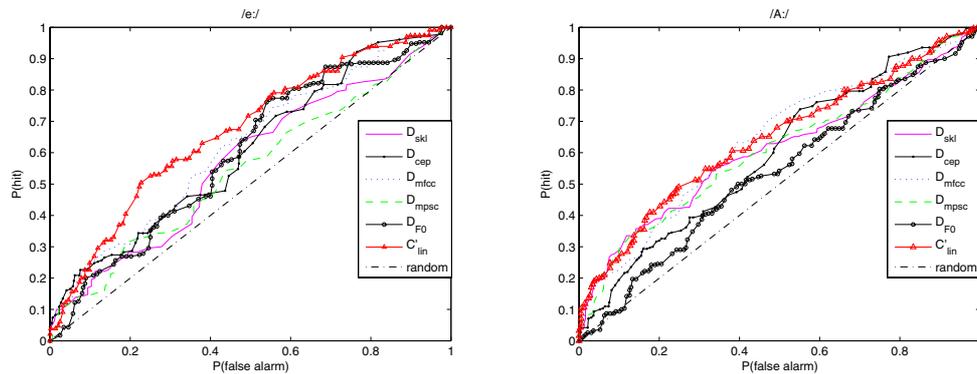
Figure 1: ROC curves using all observations

Figure 2: ROC curves after data with high F0 differences were removed

difference was the best detector, showing that this types of experiments also reflects the frequencies of the different types of discontinuities present in the test, which could be one explanation for the variability of results in such tests between different test systems and test stimuli. Of the Spectral Distance measures the D_{mfcc}, D_{LR} and D_{cep} was the most promising distance measures, somewhat better than D_{skl}, while D_{mpsc} was the worst detector. More experiments are planned, both for trying more types of sounds and to get more data to conclude on. Important issues would be to reduce possible noise sources in the experiments, find methods for significance testing between different distance measures, and to look at more distance measures and speech parameterizations with respect to both detection rate and orthogonality in a join cost setting.

6. Acknowledgements

7. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP-96*, pp. 373-376, 1996

[2] R. Duda, R. E. Hart, D. G. Stork, *Pattern Classification* , John Wiley & Sons, 2001

[3] J. Vepa, S. King, "Join cost for unit selection synthesis", in *Text to Speech Synthis*, S. Naranyan, A. Alwan, Eds., Prentice Hall, 2004

[4] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities", *IEEE Trans. SAP*, pp. 39-51, Jan. 2001

[5] Y. Stylianou and A.K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis", *Proc. ICASSP-01*, pp.837-840, Salt Lake City, 2001

[6] J.E. Natvig, P.O. Heggtveit, *PROSDATA - A speech database for study of Norwegian prosody v2.0*, Telenor R&D, N 20/2000, Kjeller 2000

[7] E. Klabbers and R. Veldhuis, "On the computation of Kullback Leibler measure for spectral distances", *IEEE Trans. SAP*, pp. 100-103, Jan. 2003

[8] A.H. Gray and J.D. Markel "Distance Measures for Speech Processing" , *IEEE Trans. on ASSP*, pp. 380-391, Oct 1976

[9] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing*, Prentice Hall, 2001

[10] N. Nukaga, R. Kamoshida and K. Nagamatsu "Unit selection using pitch synchronous cross correlastion for japanese concatenative speech synthesis," *PROC. 5th ISCA Workshop on Speech synthesis*, Pittsburgh, 2004

[11] R.E. Walpole, R.H. Myers, S.L. Myers, *Probability and statistics.*, Prentice Hall, 1998.