Collection and Analysis of Data for Evaluation of Concatenation Cost Functions*

Milan Legát and Jindřich Matoušek

University of West Bohemia in Pilsen, Faculty of Applied Sciences,
Department of Cybernetics, Univerzitní 8, 306 14, Plzeň, Czech Republic
{legatm,jmatouse}@kky.zcu.cz

Abstract. This paper describes the collection and analysis of data, which are planned to be used for the evaluation and development of concatenation cost functions for unit selection based TTS systems. Data, collected via listening tests following the recommendations given in [1], were analyzed in a variety of ways to identify and possibly exclude "malicious" listeners as well as to demonstrate their sufficient "richness" for the aimed utilization. This study was limited to five Czech vowels as these sounds are characterized by being highly energetic and having rich spectral content, which induces complexity and wide range of possible discontinuities at concatenation points.

Keywords: TTS, unit selection, concatenation cost, listening tests.

1 Introduction

Unit selection based concatenative speech synthesis still represents an approach that, without question, produces synthetic speech of the highest naturalness. The idea of this method is to have more than one instance of each unit stored in a large speech database and to search at runtime for the best sequence of units to generate the desired utterance.

In order to select the best sequence of units, two cost functions are typically calculated – *target cost* and *concatenation (join) cost*. While the task of the target cost function is to estimate the perceptual difference between a target and a candidate unit, the concatenation cost function should reflect a level of perceived discontinuity between two consecutive units.

The concatenation cost consists mostly of a set of sub-components associated with a difference in pitch, energy and spectra of adjacent segments of concatenated units. A weak point of the concatenation cost functions is the spectral component as no objective measure seems to correlate well with human perception of discontinuities in spectra.

A large number of methods have been proposed in last decade, but none of them proved to be comparatively better than others across all languages and recording conditions. The presented results have sometimes even been in contradiction. Thus, the design of the concatenation cost functions is still an open issue, and there is a lot of work remaining to be done.

* Support for this work was provided by the Grant Agency of the Czech Republic, project No. GACR 102/09/0989, by the Ministry of Education of the Czech Republic, project No. 2C06020, and partly also by the University of West Bohemia in Pilsen, project No. SGC-2010-054.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 345–352, 2010.

Generally, there are two ways of evaluating the concatenation cost functions. One can have candidates for a concatenation cost function, synthesize a set of sentences using each of them separately, and then ask listeners to choose the best version. This method is however quite laborious and costly. The other and mostly preferred option is to simply concatenate some units, let a group of listeners assess the quality of the concatenation points and then calculate correlations between values obtained by an objective measure and the listeners' scores.

The crucial point of the latter approach is to have appropriate test stimuli for the listening tests and to collect reliable results based on the listeners' answers. The purpose of this paper is to present an analysis of data we have collected following the recommendations given in [1].

2 Data Collection

2.1 Stimuli

For our experiments we have used recordings covering five Czech short vowels in all consonantal contexts made by two professional speakers (male and female) in an anechoic room. These speakers had been previously found to be appropriate candidates as they had recorded the corpora for the unit selection based TTS system [2] as well as corpora intended for the first experiments with a limited domain Czech emotional speech synthesis [3].

Recorded data were re-synthesized using the "half sentence" method described in [1] resulting in a huge set of sentences, containing only one concatenation point in the middle each. Our preliminary analyses revealed that a difference in pitch and energy at concatenation points is an important factor to be taken into consideration in order to obtain data exploitable for the design and evaluation of the concatenation cost functions, strictly speaking, the spectral component of these functions.

Note that in our preliminary listening test, all perceptually discontinuous concatenation points were clearly separable from the continuous ones in the $F0$ difference $\times$ Energy difference plane, resulting in a database which would not be feasible for an evaluation of the spectral component of concatenation cost functions. In most related studies, the standard procedure is to smooth the concatenation points with respect to differences in pitch and energy to ensure that any perceived discontinuity is not due to pitch or energy "jumps" at the concatenation points. In our study, we have decided to do concatenations without any post-processing to limit a risk of causing audible signal degradations. As the set of synthesized sentences was quite huge, we have rather measured these differences and taken them into account when selecting candidates for our listening test stimuli.

The selection of sentences included into the listening test stimuli was based on methods summarized in Tab. 1. Subsets `f0B`, `enB` and `efB` were included to confirm that large differences in pitch and energy at concatenation points are a significant source of perceived discontinuities. When selecting the candidates for the `f0B` set, a limit for the difference in energy was set to 1dB. For the selection of candidates for the `enB` set, only sentences where the difference in pitch was less than 10 mels were used. The subset `efB` consists of sentences with the largest Euclidean distance from the origin in

the $F0$ difference $\times$ Energy difference plane. The pitch and energy differences were calculated pitch synchronously having all recordings previously pitch marked using the Multi-Phase Pitch-Mark Detection Algorithm [4]. Since large measured differences in pitch and energy at concatenation points very often appear due to phonetic segmentation and/or pitch marking errors, all candidate sentences have been checked manually, and erroneous sentences were excluded.

Based on the results presented in [1], we put more stress on sentences with smooth pitch and energy transitions at concatenation points, i.e. the subsets `efS`, `mfS`, `beS`, `mfB`, `beB`. The subset `efS` consists of sentences with the smallest Euclidean distance from the origin in the $F0$ difference $\times$ Energy difference plane. For the selection of candidates for the other four subsets, we ranked all sentences according to the Euclidean distances from the origin in the $F0$ difference $\times$ Energy difference plane and took into consideration only one third of the best ones. The MFCC based distance was calculated as the Euclidean distance between two standard 39-dimensional MFCC vectors characterizing the left and right one pitch period long segments of the boundary region, respectively. The calculation of the LSM based distance was done in line with the method presented in [5] — the dimension of the SVD was set to 10 and the length of the extraction window was set as K=3.

The total number of sentences presented to listeners in each listening test was 1310. Note that the sentences themselves were not the same for both voices as the selection depended on the actual values measured on the synthesized candidates. However, the words containing the concatenation points (three words per vowel) and the number of sentences in each subset were the same (see Tab. 1).

Table 1. Subsets of sentences contained in the listening test stimuli

Set	Description	Num.
`f0B`	large pitch discontinuity and continuous energy transition	150
`enB`	large energy discontinuity and continuous pitch transition	150
`efB`	large pitch discontinuity and large energy discontinuity	150
`efS`	continuous energy and pitch transition	75
`mfS`	small pitch and energy difference + small MFCC based distance	75
`beS`	small pitch and energy difference + small LSM based distance	75
`mfB`	small pitch and energy difference + large MFCC based distance	225
`beB`	small pitch and energy difference + large LSM based distance	225
`ran`	random selection	135
`nat`	original recordings	15
`rev`	revision sentences	15
`dbl`	same source and target left and right consonantal contexts	20

2.2 Subjects

The subjects were university students, all native speakers of Czech. Some listeners stated that they had some background in phonetics. There were 29 subjects who finished

the first listening test (male voice) and 27 subjects in the second one (female voice). Approximately half of the subjects were the same across the 2 tests. All subjects were paid upon completion of the listening tests.

2.3 Procedure

The task of the listeners was to assess the concatenations in the middle vowel of the central word of each sentence on both a five-point scale (*no join at all* – 1, *unnatural but not disturbing* – 2, *slightly perceived join* – 3, *highly perceived join* – 4, and *highly disturbing join* – 5) and a binary scale (*perceived join* or *not perceived join*). To make the task easier, natural versions of the middle words containing the concatenation points were played to the listeners prior to synthesized sentences.

Both listening tests were conducted using a web interface allowing the listeners to work from home. It was, however, stressed in the test instructions that the tests shall be done in silent environment and using headphones. It is, of course, clear that organizing the tests in a laboratory would provide us with more consistent testing conditions, but taking into account the number of listeners and the length of the whole listening test, it would be unacceptably time-costly. To gain more control over the listeners, we have not only analyzed logs from our test server but also included some control mechanisms into the tests themselves (see the analysis below).

Based on the lessons learned presented in [6] and also on the feedback we had collected in our preliminary listening test, we have provided the listeners with some examples of discontinuities to help them with calibration for the more fine grained scale. They were instructed to undergo the calibration phase before starting the listening test itself and after each break they made. The logs from the test server confirmed that they indeed did so. It was allowed to listen to the calibration sentences at any time during the listening test. There were no restrictions on how many times a listener played a sentence before assessing it.

3 Analysis of Listeners' Answers

Generally speaking, listening tests are still the most reliable way of assessing the quality of synthetic speech. Nevertheless, the key factor which is not completely under control are the listeners themselves. In order to estimate consistency of their assessments and to identify possibly non-reliable participants, the detailed analysis described in the following sections has been undertaken.

3.1 Checking the Listeners' Consistency and Reliability

As found in [1] and also in [6], the inclusion of natural speech samples is a valuable resource for identifying "malicious" listeners. The test stimuli for both of our listening tests contained 15 – three words per vowel – completely natural sentences (`nat` subset). Using these sentences, we have estimated ability of the listeners to identify natural speech. Those, who assessed a natural sentence as containing an audible join, were given one penalty point for each such decision.

In addition, the ratings based on the five-point scale were also checked and each listener was given a score using a penalization scheme shown in Tab. 2. We have ranked all listeners according to their performance on this set and identified a small group of deviating listeners.

Table 2. Penalization scheme based on the listeners' answers using the five-point scale. "Diff" stands for a difference in listener's scores given to a `rev` sentence or a difference from zero in the case of `nat` sentences.

Diff	Penalty
0	0
1	−0.001
2	−0.01
3	−0.1
4	−1

Another useful measure was based on the `rev` subset, which provided us with an indicator of a listener's consistency. The method of scoring the listeners' assessments given to the sentences contained in this subset was based on a method similar to the scoring based on the `nat` subset. Here, the scores given to two instances of a same sentence were compared, meaning that an inconsistent decision on the binary scale was penalized one point, and any difference in the assessments on the five-point scale was penalized using the penalization scheme shown in Tab. 2. Upon rating the listeners we proceeded in the same way as for the `nat` subset to identify possibly non-serious participants.

It is worth noting at this point that we found some sentences which were quite ambiguous in their nature as they contained hardly perceivable joins. This was actually very likely the explanation for why the performance of the listeners in terms of binary scale assessments was generally worse on the `rev` sentences than on the `nat` subset.

3.2 Distribution of Assessments

When we more closely inspected the listeners' binary scale assessments, we realized that there was a trend in both listening tests – for the male voice approximately 60% of the sentences were assessed as containing an audible join, for the female voice the number was even higher reaching 75%. By contrast, there was a small group of listeners in both tests whose scores were equally distributed or even inversed. We have decided to penalize these listeners as it could suggest that they were not serious enough when completing the listening tests or even providing random answers.

All listeners were instructed to use the full range of categories when assessing the joins on the five-point scale. Thus, it was also interesting to analyze the distribution of their five-point scale answers. The results of the analysis are given in Fig. 1 showing the box plots of ratings distributions. Again, there was a small group of listeners in both tests whose answers were outlying.

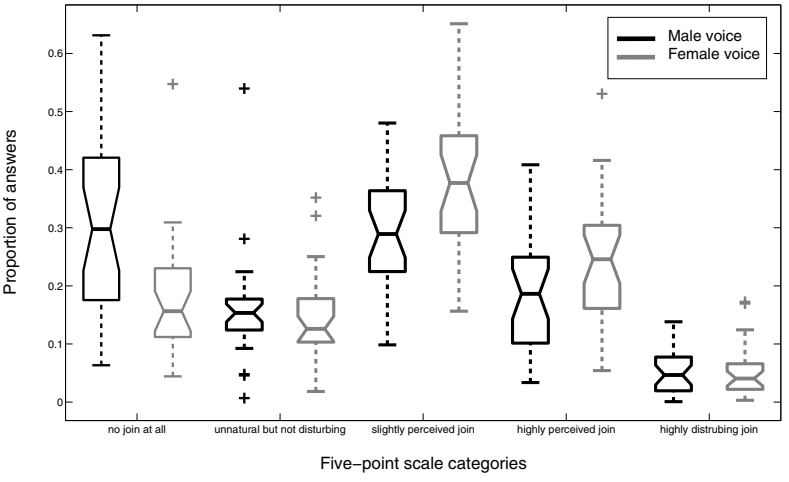

Fig. 1. Distribution of listeners' assessments on the five-point scale

3.3 (Dis)agreement with Facts

The next step of the analysis was to collect "facts", i.e. sentences which were assessed by a majority of listeners in the same way on the binary scale, either as containing an audible join or being completely natural. In our study we have set an ad hoc majority threshold to 80%. After the collection of the "facts", we have started iterating and calculating a disagreement score for each listener using the following formula:

$$\mathrm{DISAGR}_{ij} = \frac{\mathrm{NUM\_DIFF}_{ij}}{\mathrm{FACT\_COUNT}_j} \times 100\,[\%]\,, \tag{1}$$

where DISAGR_{ij} is the disagreement score of the i-th listener in the j-th iteration, $\mathrm{NUM\_DIFF}_{ij}$ is a number of assessments of the i-th listener different from "facts" in the j-th iteration and $\mathrm{FACT\_COUNT}_j$ is the number of "facts" collected in the j-th iteration.

Starting with all listeners, we have performed a few iterations to see the course of a metric based on the DISAGR measure for the worst listener in each iteration (see Fig. 2). In each iteration the worst listener was excluded and a new set of "facts" was collected. It is obvious that this metric allowed us to isolate a group of suspicious listeners in both listening tests (dashed ellipses).

3.4 Correlation with MOS

While the disagreement score (1) was defined to evaluate the listeners' answers based on the binary scale, we turn next to the analysis of the five-point scale ratings. In order to see how consistent the listeners were using this scale, we have performed an analysis of correlations of particular listeners with the group Mean Opinion Score (MOS). We

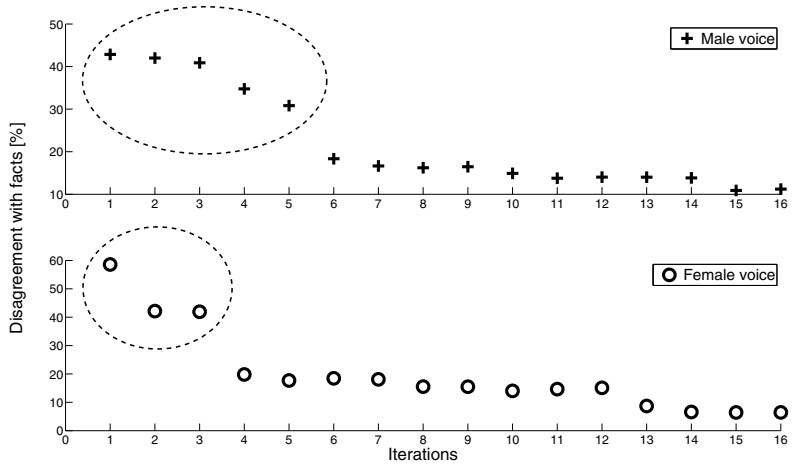

Fig. 2. Metric based on listeners' disagreements with created "facts" (see eq. (1)). The circles and crosses represent the DISAGR value of the worst listener in each iteration.

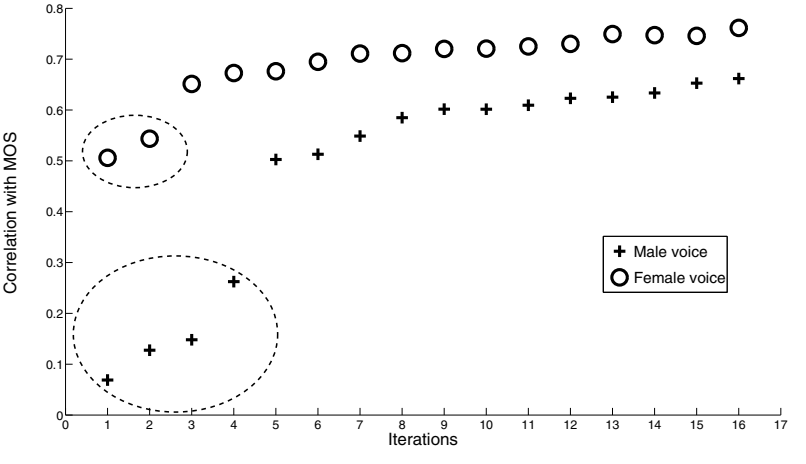

Fig. 3. Metric based on listeners' correlations with the group MOS. The circles and crosses represent the correlation with MOS of the worst listener in each iteration.

have again performed couple of iterations identifying and removing the worst listener from the set, and recalculating the MOS value in each (see Fig. 3).

4 Results

Upon scoring the listeners based on the performed analysis, we have ranked them according to the obtained scores and decided to remove 9 and 6 participants of the

male and female voice listening test, respectively. All excluded listeners obtained two or more penalty points, meaning that they were identified as deviating in at least two analysis steps. In the resulting sets the correlation with group MOS of the worst listener was 0.51 for the male voice and 0.67 for the female voice, the disagreement scores of the worst listeners were 15.99% (male voice) and 17.81% (female voice), and the total numbers of collected "facts" were 494 (male voice) and 887 (female voice). For the female voice we have unfortunately collected small number of continuous "facts" (only 11.16%), but these were still mixed with discontinuous "facts" in the small $F0$ difference $\times$ small energy difference area suggesting that there was another source of perceived discontinuity (likely the spectrum), which was our goal.

5 Conclusions and Future Work

Based on the analysis steps we have proposed in this paper, we have been able to identify and exclude "malicious" listeners who participated in the listening tests we have conducted to collect data for the design and evaluation of concatenation cost functions for unit selection based TTS systems. Despite collecting only a small number of continuous ratings for the female voice — 11.16% in contrast to 32.79% for the male voice — we still believe that we have collected a feasible dataset as the continuous and discontinuous "facts" were not separable on the pitch and energy difference basis in contrast to study [1].

The future work will focus on the development and evaluation of distance measures for costing of spectral discontinuities at concatenation points in five Czech vowels using the created database. It was also interesting to analyze the distributions of "facts" with respect to the different methods (Tab. 1) used for selecting the candidates for the listening test stimuli, and we plan to report on those results in another paper too.

References

1. Legát, M., Matoušek, J.: Design of the Test Stimuli for the Evaluation of Concatenation Cost Functions. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 339–346. Springer, Heidelberg (2009)
2. Matoušek, J., et al.: Recent Improvements on ARTIC: Czech Text-to-Speech System. In: Proceedings of the 8th International Conference on Spoken Language Processing Interspeech 2004 – ICSLP, Jeju, Korea, vol. 3, pp. 1933–1936 (2004)
3. Grůber, M., Legát, M., Ircing, P., Romportl, J., Psutka, J.: Czech Senior COMPANION: Wizard of Oz Data Collection and Expressive Speech Corpus Recording. In: Human Language Technologies as a Challenge for Computer Science and Linguistics, Wydawnictvo Poznanskie Sp. z o.o., Poznan, pp. 266–269 (2009)
4. Legát, M., Matoušek, J., Tihelka, D.: A Robust Multi-Phase Pitch-Mark Detection Algorithm. In: Interspeech 2007, Antwerp, vol. 1, pp. 1641–1644 (2007)
5. Bellegarda, J.R.: A Novel Discontinuity Metric for Unit Selection Text-to-Speech Synthesis. In: SSW5 2004, pp. 133–138 (2004)
6. Bennett, C.L.: Large Scale Evaluation of Corpus-Based Synthesizers: Results and Lessons from the Blizzard Challenge 2005. In: Interspeech 2005, pp. 105–108. Carnegie Mellon University, Pittsburgh (2005)