Design of the Test Stimuli for the 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. A large number of methods for measuring of audible discontinuities, which occur at concatenation points in synthesized speech, have been proposed in recent years. However, none of them proved to be comparatively better than others across all languages and recording conditions and the presented results have sometimes even been in contradiction. What is more, none of the tested concatenation cost functions seem to be reliably reflecting the human perception of such discontinuities. Thus, the design of the concatenation cost functions is still an open issue, and there is a lot of work remaining to be done. In this paper, we deal with the problem of preparing the test stimuli for evaluating the performance of these functions, which is, in our opinion, one of the key aspects in this field.

1 Introduction

Unit selection based concatenative speech synthesis currently 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. Ideally, no smoothing is required, which results in the high naturalness and intelligibility of the synthesized speech. In order to select the best sequence of units, two cost functions are calculated – *target cost* and *concatenation (join) cost* [1]. The task of the target cost function is to estimate the perceptual difference between the target and the candidate unit, and the concatenation cost function should reflect the level of the perceived discontinuity between two consecutive units. While the problem of the designing of the target cost functions has been more or less solved, many concatenation cost functions have been tested in the last decade with results often being in contradiction [2], [3].

The concatenation cost consists mostly of a set of sub–components associated with the difference in pitch, energy and spectra of adjacent segments of the concatenated units. The 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. Many spectral parametrizations in combination with various distance measures have been tested, including mel–frequency cepstral coefficients (MFCCs),

* This research was supported by the Ministry of Education of the Czech Republic, project No. 2C06020 and the Grant Agency of the Czech Republic, project No. GACR 102/09/0989.

V. Matoušek and P. Mautner (Eds.): TSD 2009, LNAI 5729, pp. 339–346, 2009.

linear prediction coefficients, line-spectrum frequencies, bispectrum [2], Wigner–Ville distribution–based cepstrum [5], FFT–based cepstra, perceptual linear prediction coefficients, multiple–centroid analysis coefficients [6] or formant frequencies, in combination with the Euclidean, Mahalanobis or symmetrical Kullback–Leibler distances, to name but a few. Besides these traditional approaches, some other techniques have been proposed, e.g. the application of the Latent Semantic Mapping [7], Kalman–filters [6], or Auditory Modelling [8].

Generally, there are two ways of evaluating the concatenation cost functions. The first one is to have a set of concatenation cost functions, synthesize the same sentences using each of them separately, and then ask listeners to choose the best version or to compare the synthesized versions with the natural forms of the same sentences. The other and more preferred option is to simply concatenate some units, let the listeners assess the quality of the concatenation points and then calculate the correlation between the values obtained by discontinuity measure and listeners' scores. Since the listeners' responses may vary, the Mean Opinion Score (MOS) is typically used as reference.

The crucial point for the latter approach is to have appropriate test stimuli for the listening tests and to collect the reliable results based on the listeners' answers. One of the issues related to the designing of the test stimuli for the evaluation of the naturalness of the synthetic speech is the length of the stimuli presented to listeners. It can vary from isolated phonemes [9] to whole sentences [5]. Shorter stimuli (e.g. isolated phonemes) could seem to be appropriate for evaluation of the concatenation cost functions as they can be synthesized containing only one concatenation point, which is important for avoiding the influence of other joins on listeners' assessments. On the other hand, if the test stimuli are too short, it is difficult for the listeners to perceive discontinuities [9]. Thus, monosyllabic words containing only one concatenation point in the middle seem to be an optimal choice.

However, there is another point of view. The general task of TTS is to synthesize sentences or longer texts (in most cases). If we use isolated words for evaluation of join cost functions, we can expect lower variability of the quality of the concatenated sounds than in real cases. In addition, in isolated words the discontinuities may be masked or difficult to perceive, compared to whole sentences. What we mean is that if the listener is presented with a whole sentence which is absolutely natural except one concatenation point, a possible discontinuity may be more salient than in the case of being presented in a short monosyllabic word.

In this work, we propose a framework for the design of the test stimuli for the listening tests for evaluating the concatenation cost functions. Note that the initial focus of our experiments is on vowels as they are the sounds which can be characterized as highly energetic and having rich spectral content.

2 Preliminary Listening Test Preparation

We have recorded a large database of short Czech sentences containing three words each, e.g. `/kra:lofski:` **`kat`** `konal/` (SAMPA notation), where the middle word is of special interest as it is a mono–syllabic word containing a vowel in the middle surrounded by consonants, i.e. the word in the form CVC. All five short Czech vowels were

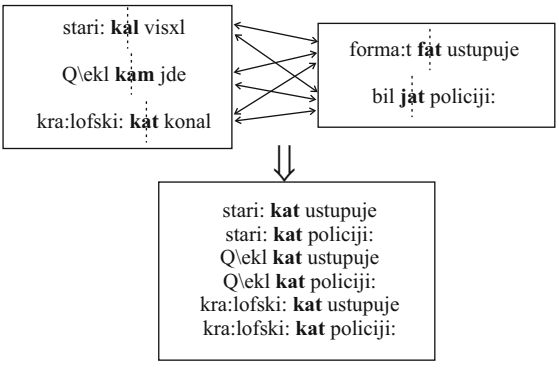

Fig. 1. Construction of the set of sentences containing only one concatenation point (SAMPA notation). The sentences were synthesized combining initial and final halves of the recorded sentences so that they contain one of the five design words in the middle (`kat` in this case).

taken into consideration. The middle mono–syllabic words contain all possible combinations of initial and final consonants. The sentences were uttered by both female and male speaker. This database is planned to be used for future concatenation costs evaluation and design, but first we need to find a reasonable way to exploit this database for that purpose.

We have divided all the sentences in the middle, i.e. at mid–vowels of the central mono–syllabic words. Then, the left and right halves of the sentences were concatenated in order to obtain sentences which are completely natural except one concatenation point in the middle (see Fig. 1). Note that no smoothing techniques were applied, and the concatenation was performed at the middle pitch mark position; this set is henceforth referred to as HS-ALL. The objective of synthesizing the HS-ALL set in this way was to present listeners with stimuli long enough for consistent assessment of the audible discontinuities while containing only one concatenation point at the same time.

In addition to the HS-ALL set, we have also included into the test stimuli for the preliminary listening test a set of sentences (henceforth referred to as DI-ALL), in which the middle words were synthesized using diphones taken from the diphone inventory created from the whole set of the recorded sentences. Thus, the synthesized sentences were natural except three concatenation points in the middle word. The inclusion of these sentences was motivated by questioning whether the range of possible discontinuities is not too narrowed by taking only the middle words into account. However, these sentences were handled with special care in the evaluation procedures because there were two other concatenation points present in the surrounding of the middle concatenation point, which was of interest. To limit the effect of surrounding joins on listeners' scores, the spectrograms and also waveforms at the surrounding concatenation points were checked, and the sentences containing some visible discontinuities at these points were removed.

Since the number of the sentences in both HS-ALL and DI-ALL sets was very large, there was a need to make a limited selection for the preliminary listening test. For that purpose, two approaches were used. First, a discontinuity metric proposed by Bellegarda [7] was implemented, using three extraction window lengths ($K = 3, 4, 5$ pitch

periods). Second, we measured the discontinuities at the concatenation points using the Euclidean distance between MFCC vectors, three window lengths were used for the feature extraction (10, 20 and 30 ms), motivated by the results presented in [10], where the authors suggest that the performance of discontinuity measures may depend heavily on the feature extraction window length.

With each of these approaches, 15 best and 15 worst concatenations were found resulting in the set of 180 sentences (6 x 2 x 15). In addition, we included 15 sentences chosen randomly and 5 completely natural sentences, 10 sentences were included twice. The objective of the inclusion of the natural and doubled sentences was to have a tool for checking the listeners' consistency and their ability to perceive discontinuities. The total number of sentences presented to the listeners in the preliminary listening test was 210, including 90 sentences from the HS-ALL set (HS-SEL), 90 sentences from the DI-ALL set (DI-SEL), 15 sentences chosen randomly (mixture of both DI and HS), 5 natural sentences and 10 sentences included twice.

The task of the listeners was to assess the concatenations in the middle vowel of the central word of each sentence on both the 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), as well as the binary scale (*perceived join* or *not perceived join*). To make the task even easier for listeners, the natural versions of the middle words were also played to them prior to the whole sentences. The overall number of participants in this preliminary listening test was 20.

3 Evaluation of the Preliminary Listening Test

3.1 Checking the Listeners' Consistency and Reliability

After collecting all the answers, the evaluation of the listeners was performed. Firstly, the ability of listeners to identify the natural sentences was estimated. The listeners, who assessed the natural sentences as containing audible joins, were given one minus point for each such decision. Based on this penalization, one listener was excluded. Secondly, the consistency of the listeners comparing their answers for the sentences included twice in the listening test was evaluated. Again, for each inconsistent decision on the binary scale the listener was given one minus point. All the listeners were found to be consistent using this measure as one mistake was allowed.

We also assessed the consistency of the listeners' answers on the five-point scale. See the penalization scheme in Tab. 3.1. If a listener, for instance, assessed the same sentence using *unnatural but not disturbing* and *highly disturbing join*, the penalization was -0.1 point. Based on this consistency measure, two listeners were excluded obtaining a score of -0.201 and -0.140, respectively. The average score of the remaining listeners was -0.008.

The answers of the 17 remaining listeners were used to calculate the Mean Opinion Score (MOS). We found the correlation between listeners' scores and MOS, and two listeners were excluded at this point as their answers correlated poorly with MOS (0.24, 0.51 respectively). The average correlation between the scores of the remaining 15 listeners and MOS was 0.76, resulting in a reliable set of listeners' scores.

Table 1. Penalization scheme based on the listeners' answers on the five–point scale. "Diff" stands for the difference in a listener's scores given to the same sentence.

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

The next step of our evaluation procedure was the formulation of "facts". By "fact" we mean a sentence which was assessed by 80% listeners in the same way on the binary scale, either as containing an audible join or as being completely natural. In the HS-SEL set we found 39 "facts" (14 continuous and 25 discontinuous), which is about 42% of HS-SEL sentences. In the DI-SEL, 9 continuous "facts" were found. Before any discontinuous "facts" can be formulated, we need to be sure that any of the perceived discontinuities were not due to poor quality of the surrounding joins.

3.2 Checking the Surrounding Joins in DI-SEL Set

As mentioned above, the sentences in the DI-SEL set were evaluated in a special way, despite being checked visually before inclusion into the test stimuli, as the listeners' scores might have been affected by the presence of some discontinuities at surrounding joins. In order to analyze these unwanted joins, which are in fact introducing some noise into the results our listening test, we have calculated the energy and pitch differences at all concatenation points.

Since human loudness perception does not scale linearly with the intensity of the signal, the energy mismatches at the concatenation points were measured on the logarithmic scale (dB). The same applies to the measuring of differences in F_0, for which Mel scale was used. We found the maximum acceptable differences in both energy and pitch at all concatenation points in sentences which were found to be continuous "facts". These values were then used as thresholds for excluding sentences from the DI-SEL set based on checking the surrounding joins. Having these joins checked in this way, the discontinuous "facts" can also be found in the DI-SEL set. In our case, 17 discontinuous "facts" in DI-SEL set were found. The total number of the "facts" obtained from our listening test results was 65 (23 continuous and 42 discontinuous).

3.3 Energy and Pitch Differences at Concatenation Points

Since the objective of our work is to measure spectral discontinuities, we need to analyze the differences in pitch and also energy at these points as these are unquestionably the sources of audible discontinuities. Note that in related works such sources of discontinuities are very often eliminated by $F0$ and energy smoothing. Nevertheless, in our opinion, any signal modification can introduce some audible artefacts, which may also affect the listeners' discontinuity perception and that is why the unmodified concatenations are preferable for the design of the concatenation cost functions, especially building the test stimuli for the listening tests.

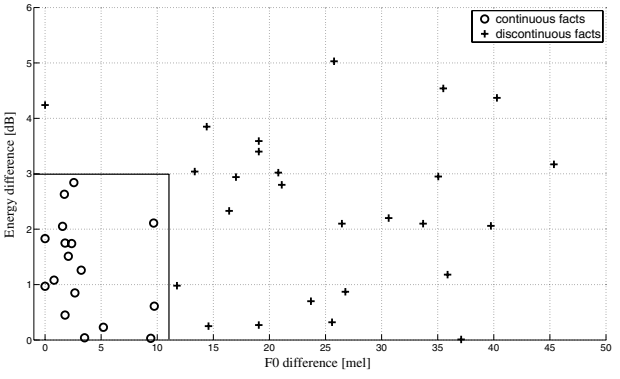

Fig. 2. Distribution of the continuous and discontinuous "facts" in the energy difference vs. $F0$ difference plain. The continuous "facts" can easily be separated from the discontinuous "facts".

In Fig. 2 the result of the analysis of the $F0$ and energy differences at concatenation points present in the "fact" sentences is shown. Surprisingly, there were no sentences among the discontinuous "facts" where the discontinuity in spectrum could be considered to be the only source of the perceived discontinuity. In fact, it was possible to classify all the sentences into continuous vs. discontinuous classes using only pitch and energy differences at the concatenation points as predictors. Obviously, there were some cases where some considerable spectral mismatches at concatenation points were observed. However, in all such cases this spectral mismatch was accompanied by either pitch or energy mismatch. Thus, there would be no need to measure spectral discontinuity to concatenate well at these points as the difference in $F0$ and energy would be sufficient components of the concatenation cost function.

4 Suggestions for Listening Test Stimuli Design

In this section, we summarize our observations related to the building of the listening test stimuli for the evaluation and design of concatenation cost functions.

One of the first questions we wanted to address was whether the concatenation points present in the synthesized sentences created from the halves of sentences as described in Sec. 2 contain enough discontinuities. Based on the obtained results, we could assume that this approach is a reasonable way as the number of perceived discontinuities follows the same trend as for the sentences in which diphones were used. The advantage of the synthesizing by halves is that we have stimuli long enough for reliable assessment by listeners, and containing only one concatenation point at the same time.

We also assume that the listeners were able to give consistent scores as the average correlation between the listeners' scores and MOS was 0.71. On the other hand, 42% "facts" do not seem to be enough, and this number needs to be taken into consideration for the design of the future listening test to obtain enough "facts" for the evaluation of the concatenation cost functions.

For the design of the larger listening test containing all the Czech vowels and utterances spoken by both of the speakers, the differences in pitch and energy at concatenation points need to be considered before the sentences are included into the listening test stimuli in order to find sentences where the spectral mismatch could be found as the only source of perceived discontinuity, which is crucial for the design of the spectral component of the concatenation costs.

It is also worth performing an analysis of listeners' scores as described in Sec. 3.1 in order to measure their ability to distinguish between continuous and discontinuous concatenation points, as well as the consistency of their answers.

5 Conclusion and Future Work

In this paper, we have addressed one of the key issues of evaluating the concatenation cost functions for the unit selection speech synthesis, the building of the listening test stimuli. To answer some questions related to this field, we have performed a preliminary listening test, the objective of which was to show that synthesizing sentences by halves could be one way of approaching the problem of "length vs. concatenation" of the test stimuli building. By the "length vs. concatenation" problem we mean that if the listening test stimuli are too short, it is difficult for the listeners to reliably score the concatenations, and having longer test stimuli requires some additional concatenation points, which might also affect the listeners' judgements.

We found that it is beneficial to check the listeners' for their ability to distinguish between smooth and discontinuous joins, and for the consistency of their answers. For this purpose, we proposed in this paper a simple method based on the inclusion of some natural and some doubled sentences into the test stimuli.

Special attention needs to be paid to the pitch and energy differences as the sources of perceived discontinuities. In our opinion, smoothing methods are not a reliable way of dealing with this problem as they may introduce some audible artefacts into the concatenation area, which might affect the listeners' scores.

In our preliminary listening test we have not observed any sentences where the spectral mismatch could be concluded to be the only source of perceived discontinuity as all discontinuous sentences contained either pitch or energy difference at the concatenation points. This particular issue is planned to be addressed in future listening tests containing all the short Czech vowels. For the evaluation and design of the spectral component of the concatenation cost functions, we need to find a set of sentences which are scored by the listeners as discontinuous and, at the same time, do not contain considerable pitch and energy discontinuities at the concatenation points.

References

1. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: ICASSP 1996, vol. 1, pp. 373–376 (1996)
2. Pantazis, Y., Stylianou, Y.: On the detection of discontinuities in concatenative speech synthesis. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (eds.) COST 277. LNCS, vol. 4391, pp. 89–100. Springer, Heidelberg (2007)

3. Vepa, J., King, S.: Join cost for unit selection speech synthesis. In: Alwan, A., Narayanan, S. (eds.) Speech Synthesis. Prentice Hall, Englewood Cliffs (2004)
4. Kawai, H., Tsuzaki, M.: Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis. In: ICSLP 2002, pp. 2621–2624 (2002)
5. Chen, J., Campbell, N.: Objective distance measures for assessing concatenative speech synthesis. In: EUROSPEECH 1999, pp. 611–614 (1999)
6. Vepa, J.: Join cost for unit selection speech synthesis. PhD Thesis, University of Edinburgh (2004)
7. Bellegarda, J.R.: A novel discontinuity metric for unit selection text–to–speech synthesis. In: EUROSPEECH 1999, pp. 611–614 (1999)
8. Tsuzaki, M.: Feature extraction by auditory modelling for unit selection in concatenative speech synthesis. In: EUROSPEECH 2001, pp. 2223–2226 (2001)
9. Klabbers, E., Veldhuis, R.: Reducing audible spectral discontinuities. IEEE Transactions on Speech and Audio Processing 9, 39–51 (2001)
10. Kirkpatrick, B., O'Brien, D., Scaife, R.: Feature extraction for spectral continuity measures in concatenative speech synthesis. In: INTERSPEECH 2006 (2006)