Pitch Contours as Predictors of Audible Concatenation Artifacts

Milan Legát and Jindřich Matoušek

Abstract—**This paper deals with the traditional problem of the occurrence of audible discontinuities at concatenation points at diphone boundaries in the concatenative speech synthesis. While most of the related studies put stress on the spectral component, we focused on the pitch contours and their role as predictors of the discontinuities. To measure the amount of information contained in the pitch contours, we trained SVM classifiers using perceptual data collected in listening tests. The results have shown that the fine grained pitch contours extracted from a vicinity of the concatenation points carry enough information for classifying continuous and discontinuous joins with a high accuracy.**

Index Terms—**speech synthesis, unit selection, concatenation cost, pitch contours.**

I. Introduction

DESPITE the increasing popularity of HMM based speech synthesis methods, the unit selection concatenative systems still represent the mainstream in many practical applications, especially in limited domains where synthesized chunks are combined with pre-recorded prompts. In such applications, the ability of the unit selection to deliver highly natural and to the recordings well fitting output are the key factors. Not surprisingly, the unit selection also remains the first choice for eBook reading applications, which have been acquiring a lot of interest over recent years.

Among the unit selection related issues that continue to be non-resolved, the audible discontinuities appearing at concatenation points play an important role. According to the original idea [1], the amount of discontinuity introduced by concatenating successive units should be reflected by a *concatenation (join) cost function*. Since phase, pitch and spectral envelope mismatches are believed to be the main sources of the discontinuities [2], ideal concatenation cost function should cover all these aspects.

In our previous work [3] dealing with vowels, and also in an informal analysis of concatenation artifacts present in the outputs of our TTS system [4], it was found out that a large number of audible discontinuities tend to appear at joins where units having originally incoherent $F0$ contours in the area of the prospective concatenation points are put together. Other possible sources of discontinuities were also identified but not in such an extend.

Many studies have been published over last one and a half decades focusing on the spectral mismatches in the first place while eliminating the other sources of discontinuities [5], [6], [7], to name but a few. Despite the considerable amount of efforts, none of them unfortunately succeeded to provide a clear answer on how to measure the discontinuities at concatenation points. The presented results have even sometimes been in contradiction. Another interesting study [8] showed that the discontinuity detection rates hardly reach 50% (at 5% false alarm rate) when using spectrum oriented methods.

In line with the observations mentioned above, we decided to extract pitch contours from the vicinity of concatenation points and use them as predictors in the discontinuity detection task performed by SVM classifiers. Four different sets of $F0$ based predictors, described in Sec. III-B, were used to answer the question of how much information is contained in concatenated $F0$ contours (their slopes, shapes, static differences, etc.) with respect to the discontinuities perceived by listeners.

A hypothesis under question was that the incoherent concatenated $F0$ contours lead to perceived discontinuities, which should be learned by the classifier, whereas coherent $F0$ contours are not sufficient condition for perceptually smooth concatenations, which should be decreasing the classifiers' sensitivity.

The perceptual data used for training and evaluation of the classifiers were collected in listening tests described in Sec. II. The classification experiment set up and the SVM models are described in Sec. III, the results are then summarized in Sec. III-D. Finally, we discuss some of our observations in Sec. IV, and draw conclusions and outline our future work in Sec. V.

II. Perceptual data collection

In order to collect data that can be used for the evaluation and design of the concatenation cost functions, we had conducted two listening tests in the past—one to collect male voice data, one for female voice data. In the following subsections, the content and the evaluation procedure of these listening tests are briefly described. More details may be found in [9], [10].

A. Test Material

The recordings covering five Czech short vowels in all consonantal contexts were made in an anechoic room by two professional speakers—male and female. The recorded scripts were composed of three word sentences containing CVC word in the middle each, e.g. `/kra:lofski:` **`kat`** `konal/` (Czech SAMPA notation). Recorded data were re-synthesized using the "half sentence" method [3]. This method consists in cutting the sentences in the middle of the vowels in the central words and combining

Manuscript received July 3, 2011; revised August 12, 2011. This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989 and by the Technology Agency of the Czech Republic, project No. TA01030476. The work has also been supported by the grant of the University of West Bohemia, project No. SGS-2010-054.

The authors are with the Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic e-mail: `legatm@kky.zcu.cz`, `jmatouse@kky.zcu.cz`

the left and right parts, which results in a large set of sentences containing only one concatenation point in the middle of the central CVC word each and covering the vowels in all possible consonantal contexts. Note that the concatenations were done pitch synchronously to avoid phase mismatches, but no smoothing algorithm was applied.

Since the whole set of synthesized sentences is too large to be entirely used as listening test stimuli, and we did not want to make a random selection, different concatenation cost functions were applied to collect a limited set of sentences, which were then included to the listening tests stimuli.

The motivation for using different concatenation cost functions was to gain control, albeit limited (due to unreliability of the the traditional concatenation cost functions), over the listening test results without having any a priori knowledge about the distribution of audible discontinuities in the synthesized data. The selection was done with the expectation to obtain from listeners slightly larger number of discontinuous ratings.

The total number of sentences presented to the listeners in each listening test was 1310, including some natural and revision sentences.

B. Subjects

The subjects were university students, all native speakers of Czech. A few 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 two tests. All subjects were paid upon completion of the tests.

C. Procedure

The task of the listeners was to assess the concatenations on both the five-point scale (*no join at all*, *unnatural but not disturbing*, *slightly perceived join*, *highly perceived join*, and *highly disturbing join*), and the 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 the synthesized sentences. Note that in the classification experiment presented in this paper only the binary scale ratings were used.

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 the silent environment and using headphones. 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 [9]. To help the listeners calibrate for the more fine grained scale, a preparation phase was included containing various examples of audible discontinuities. It was allowed to listen to the calibration sentences at any time during the listening test. There were no restrictions on how many times the listeners played each sentence before assessing it.

D. Listening Test Evaluation and Results

In order to identify listeners who did not show good agreement with the majority, a rigorous analysis of the listeners' ratings has been performed [9]. We ranked the participants according to the scores obtained by the analysis, and 9 and 6 participants were excluded from the male and female voice listening tests, respectively. The ratings of these listeners were not used to create a set of "facts" and to calculate agreement scores as described below.

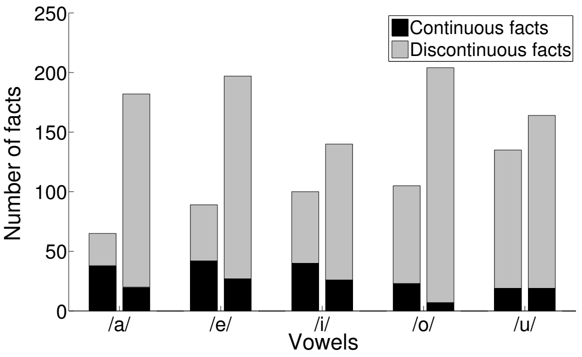

Fig. 1. The "facts" collected in the listening tests sorted by vowels—the left bar in each pair represents the male voice results.

TABLE I
AGREEMENTS SCORES (1) OF THE THREE LEAST AGREEING LISTENERS PARTICIPATING IN THE LISTENING TESTS.

	Male	Female
List1	0.84	0.82
List2	0.87	0.83
List3	0.88	0.84

As the next step, we have collected two sets of "facts", i.e. sentences that were assessed by more than 80% of the listeners in the same way on the binary scale, either as containing an audible join or being natural. The set of "facts" can be formally described as:

$$\mathrm{sent}_i \in \mathrm{FACTS} \quad \Leftrightarrow \quad \frac{\mathrm{N}_i^+}{\mathrm{N}_i} \geq 0.8 \ \vee \ \frac{\mathrm{N}_i^-}{\mathrm{N}_i} \geq 0.8,$$

where sent_i is the i-th sentence of the test stimuli, FACTS stands for the set of "facts", N_i^+, N_i^- are the numbers of continuous (i.e. *not perceived join*) and discontinuous (i.e. *perceived join*) ratings given to the i-th sentence, respectively, and N_i is a total number of ratings given to the i-th sentence.

The total numbers of the collected "facts" were 494 for the male voice and 887 for the female voice. Fig. 1 shows the distributions of the "facts" for each vowel and both speakers.

The next step was to calculate an agreement score of each listener using the following formula:

$$\mathrm{AGR\_SCORE}_i = \frac{\mathrm{NUM\_AGR}_i}{\mathrm{FACT\_COUNT}}, \tag{1}$$

where $\mathrm{AGR\_SCORE}_i$ is the agreement score of the i-th listener, $\mathrm{NUM\_AGR}_i$ is a number of ratings of the i-th listener in agreement with the "fact" rating and FACT_COUNT is the number of the collected "facts".

The agreement scores (1) of the three least agreeing listeners for each voice, which may serve as a reference for the evaluation of the classifiers' performance, are summarized in Tab. I. The score of the least agreeing listener in each test is also depicted in Fig. 3.

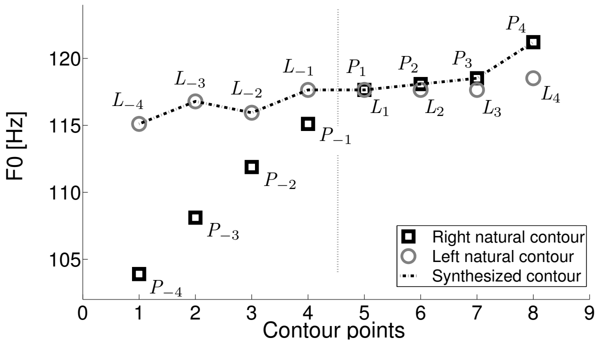

Fig. 2. Annotation scheme used for labeling the $F0$ contours. As an example, let $[L_{-4} \dots L_4]$ be the $F0$ contour extracted from the central part of the vowel /a/ in the word /t_Sak/ and $[P_{-4} \dots P_4]$ the contour of /a/ in the word /mas/. Then, the sequence $[L_{-4} \dots L_{-1}, P_1 \dots P_4]$ represents the central part of the concatenated $F0$ contour of the word created as /t_Sa-as/ (in Czech SAMPA notation).

III. Classification experiment

A. Motivation

As already mentioned in Sec. I, the experiment presented in this paper was aimed at answering the question of how much information is contained in pitch contours with respect to the discontinuities perceived by the listeners. The task was formulated as a binary classification problem using the pitch contours extracted from the vicinity of concatenation points and/or their parametrization as predictors.

The SVMs were chosen as the classification model due to their proven feasibility for different classification tasks, and the availability of the training framework.

B. Collection of Predictors

1) Sets of Predictors: Since the continuous "facts" collected in the listening tests were in some sets rather underrepresented compared to the discontinuous "facts" (see Fig. 1), we decided to include some natural sentences in order to make the experimental data better balanced.

As a preparation for collecting the sets of predictors, the recorded sentences were pitch marked using the robust multi-phase pitch marking algorithm [11]. Since no pitch smoothing method was applied during synthesis, the pitch marks remained preserved in the synthesized sentences. The $F0$ contours were then calculated, and the following sets of predictors were created:

- `Reg` : $[M_{-1}, M_1, K_L, K_R]$
- `SReg` : $[\hat{K}_L, \hat{Q}_L, \hat{K}_R, \hat{Q}_R]$,
- `Syn` : $[L_{-4} \dots L_{-1}, P_1 \dots P_4]$
- `Nat` : $[L_{-4} \dots L_{-1}, L_1 \dots L_4, P_{-4} \dots P_{-1}, P_1 \dots P_4]$,

where L_i and P_i represent $i-$th point of natural $F0$ contours pitch synchronously extracted from the vicinity of a prospective concatenation points from the vowels that were concatenated (see Fig. 2), the values M_{-1} and M_1 were calculated as:

$$M_{-1} = (L_{-2} + L_{-1})/2$$

$$M_1 = (P_1 + P_2)/2$$

The values K_L and K_R are the slopes of linear regression lines fitted to the left and right natural $F0$ contours, respectively, and the pairs $\hat{K}_L$, $\hat{Q}_L$ ($\hat{K}_R$, $\hat{Q}_R$) were obtained as parameters of linear regression lines fitted to the sequences $[L_{-2} \dots L_2]$ ($[P_{-2} \dots P_2]$), which were first smoothed by a median filter.

2) Rationale: The `Reg` set was included to address the assumption that static differences in pitch at the concatenation points together with slopes of the concatenated $F0$ contours represent the key predictors of the audible $F0$ discontinuities.

Since the estimated slopes of the $F0$ contours may be significantly affected by gross pitch marking errors, the `SReg` set was included. Considering the results of the evaluation of the accuracy of the pitch marking algorithm [11], no big differences were expected when comparing the performance of the classifiers trained on the `Reg` and the `SReg` sets of predictors.

The `Syn` set only contained synthesized $F0$ contours. These contours do not contain any information about the elements of the $F0$ sequences following the left part, or preceding the right part of a synthesized vowel in the natural data. Since no pitch smoothing was applied during concatenating halves of the recorded sentences, there might have been considerable $F0$ jumps at the concatenation points. At the same time, the synthetic $F0$ contours may also appear to be very smooth, even in cases where the original natural contours have rather different slopes as shows the example depicted in Fig. 2.

To get a full description of the concatenated $F0$ contours, the `Nat` set composed of both natural concatenated $F0$ contours extracted from the vicinity of the prospective concatenation points was added.

C. Training the Models

As suggested in [12], we decided to first try the linear kernel, which may serve as a baseline, and then compare the results with a non-linear kernel—the Gaussian (RBF) in our case.

To find the best SVM hyperparameters, we conducted a grid search using the grid points distributed on a logarithmic scale. In the first step, we used a coarse grid to find a promising region, and then we further searched for better hyperparameters' values using a finer grid. The K-fold cross-validation technique, with K set to the value 5, was used to estimate the classifiers' performance in each point on the grid.

Note that the cross-validation should help to prevent the overfitting problem.

D. Classification Results

1) Linear Kernel Models: We turn first to the results of the classification using the linear kernel SVMs. The classifiers' performance rates in terms of accuracy (ACC), sensitivity (recall rate, SENS) and specificity (SPEC) averaged across all vowels are presented in Tab. II. The sensitivity and specificity were in our case defined as follows:

$$\text{SENS} = \frac{\text{TRUE\_CONT}}{\text{TRUE\_CONT} + \text{FALSE\_DISCONT}} \qquad (2)$$

$$\text{SPEC} = \frac{\text{TRUE\_DISCONT}}{\text{TRUE\_DISCONT} + \text{FALSE\_CONT}}, \qquad (3)$$

TABLE II
CLASSIFICATION RESULTS—LINEAR KERNEL SVMS (AVERAGE ACROSS ALL VOWELS)

	Male			Female		
Predictors	ACC	SENS	SPEC	ACC	SENS	SPEC
`Syn`	0.74	0.73	0.71	0.62	0.35	0.84
`Nat`	**0.79**	**0.87**	**0.68**	**0.72**	**0.76**	**0.69**
`Reg`	0.73	0.78	0.65	0.65	0.56	0.72
`SReg`	0.76	0.83	0.68	0.66	0.47	0.82

TABLE III
CLASSIFICATION RESULTS—GAUSSIAN KERNEL SVMS (AVERAGE ACROSS ALL VOWELS)

	Male			Female		
Predictors	ACC	SENS	SPEC	ACC	SENS	SPEC
`Syn`	0.89	0.92	0.86	0.90	0.87	0.91
`Nat`	**0.93**	**0.96**	**0.91**	**0.92**	**0.95**	**0.91**
`Reg`	0.91	0.93	0.88	0.92	0.91	0.93
`SReg`	0.90	0.90	0.90	0.92	0.93	0.90

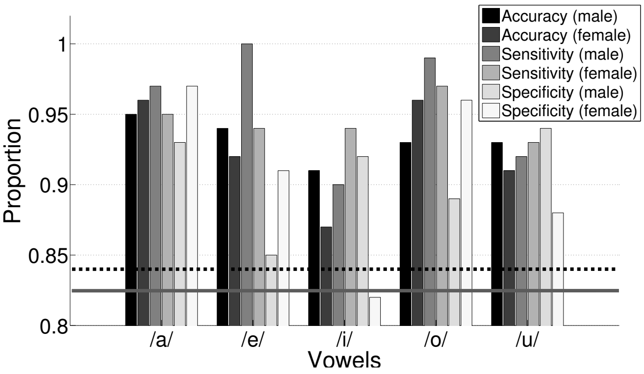

Fig. 3. Comparison of the classification results across all vowels and both speakers—`Nat` set of predictors. The dotted black and solid gray lines show the lowest agreement scores (1) obtained from the listeners participating in the listening tests (see Sec. II) for the male and female voice, respectively.

TABLE IV
HYPERPAREMETER VALUES OF THE SVM MODELS.

		Linear	RBF	
Speaker/Vowel		**C**	**C**	γ
Male	/a/	0.0385	5.6569	0.0412
	/e/	1.0718	1.7411	0.2679
	/i/	1.6245	0.5743	0.1340
	/o/	8.0000	10.556	0.1649
	/u/	2.6390	2.8284	0.3536
Female	/a/	8.0000	6.0629	1.3195
	/e/	8.5742	6.9644	0.7579
	/i/	2.4623	64.000	0.0292
	/o/	24.252	3.2490	1.7411
	/u/	1.4142	6.4980	0.4061

where TRUE_CONT is a number of continuous "facts" classified as such, FALSE_DISCONT is a number of discontinuous "facts" classified as continuous, TRUE_DISCONT is a number of correctly classified discontinuous "facts", and FALSE_CONT is a number of continuous "facts" classified as discontinuous.

These measures were calculated in order to get more insight into the performance of the classifiers as well as to address the hypothesis formulated in Sec. I. In Sec. IV, we will further discuss the obtained results.

It can be seen that the accuracy of the SVMs using linear kernel is not very high. It is, however, a promising result, taking into account the difficulty of the classification task. The classifiers performed significantly worse on the female voice data than on the male voice data. Regarding the different sets of predictors, the `Nat` set seems to be giving the best results. This observation may be attributed to the fact that using the whole $F0$ contours increases the variance in the data, which may help the linear kernel SVMs to find better separation between the two classes.

2) Gaussian Kernel Models: Having obtained the results by the linear kernel SVMs, the question was how much we can improve by introducing the non-linear kernel. The values presented in Tab. III show that all sets of predictors lead to comparatively higher performance rates.

By contrast, no significant difference was found between the averaged results for the male and the female voice data. The `Nat` predictors lead to the best classification results, and the `Syn` set shows, comparatively to the linear kernel SVMs, the worst results. This suggests that the knowledge of the whole concatenated $F0$ contours is beneficial.

If we look at the variance of the classifiers' performance across different vowels (see Fig. 3), we can see that the accuracy of the classification was significantly lower for the vowel /i/, especially for the female speaker.

3) Summary of Models' Hyperparameters: For completeness' sake, we present in Tab. IV the values of the SVM models' hyperparameters obtained during training on the `Nat` set of predictors. In can be seen that the hyperparameter values of most of the models are relatively small suggesting that the models should be capable of generalization. The exception is the Gaussian kernel SVM model for the female voice vowel /i/ where it tends to overfit, which also explains the lower accuracy estimate obtained by the cross-validation.

IV. DISCUSSION

Based on the assumption that concatenating coherent $F0$ contours is necessary but not sufficient condition of perceptually smooth concatenations (not applying any smoothing), and that concatenating incoherent $F0$ contours leads in most cases to perceptually discontinuous joins, the sensitivity was expected to be comparatively lower than specificity.

As can be seen from Tab. III, our expectation was rather not supported by the actual measurements showing that different sets of predictors lead to different results. The `Nat` set, for which we achieved the highest classification accuracy, shows the opposite of what we were originally expecting. If we look more closely at Fig. 3, we can see that the sensitivity and specificity rates may vary from vowel to vowel, and even inconsistently when comparing the two speakers.

This observation does not necessarily disconfirm the assumption that coherent $F0$ contours are the necessary condition for the perceptually smooth concatenations. Since the models were trained with respect to their accuracy, of which we believe we obtained quite robust cross-validation estimates, and the specificity and sensitivity rates may to some extend vary depending on the randomization of the training data, it may suggest that there are some clusters of $F0$ contours, which are not well separable. It is than the matter of training, into which class these clusters are ranked, which results in the variance of sensitivity and specificity measures. These clusters must however be rather non-dominant in our data since the models' accuracy remain high.

As a matter of fact, the construction of confidence intervals around cross-validation estimates is considered to be a difficult problem. Nevertheless, if we look at the models' classification accuracy, assuming that the bias of its estimates is rather towards a poorer fit (which is believed to be true for cross-validation estimates), and make the comparison with the agreement scores listed in Tab. I, which are slightly biased in the direction of higher values (due to the participation of each listener in the creation of the "facts"), we can see that the SVM classifiers perform very well, and the high obtained accuracy is clearly exceeding our expectations.

It is, however, important to mention at this point that the presented results are not meant to question the role of the spectral envelope and/or phase mismatches in the perception of the concatenation discontinuities. They should rather suggest that the discontinuities can be detected with a high accuracy using the $F0$ contours as predictors, and this knowledge is beneficial for improving the concatenative speech synthesis.

Putting more stress on the $F0$ contours during unit selection and improving their modeling may be a promising way to improve the output of our TTS system [4], which currently uses the combination of Mel-frequency cepstral coefficients (MFCCs), static $F0$ and energy differences at the concatenation points as the components of the concatenation cost function.

V. Conclusions and future work

This paper presented the results of audible discontinuity detection task performed by the SVM classifiers trained on the $F0$ contours extracted from the vicinity of concatenation points and/or their parametrization. The results suggest that the information contained in the contours is sufficient to detect audible concatenation discontinuities with a high accuracy falling into the range around 90%, which is unquestionably a very good result.

The Gaussian kernel SVMs were found to be giving better classification results than the linear kernel SVMs. The best classification accuracy was achieved using all points of the $F0$ contours extracted pitch synchronously from the vicinity of prospective concatenation points. Nevertheless, the parametrization of the contours by linear regression (no matter if the contours are pre-smoothed or not) does not significantly decrease the models' accuracy. Using only the synthesized $F0$ contours seems to be slightly inferior.

The results have also shown that the specificity and sensitivity measures may vary across vowels from speaker to speaker, and also for different set of predictors, which does not support, neither disconfirm, the assumption that concatenating coherent $F0$ contours is necessary but not sufficient condition of perceptually smooth concatenations.

Future work will focus on the incorporation of the learned knowledge into our TTS system. The challenge will be to find out to what extend the models trained in the limited prosodic context may be generalized to cover large scope of prosodic environments without the need to perform additional costly listening tests, as well as to what extend they may be used for different speakers in the same contexts. We will also more closely inspect the sentences which lead to classification errors, in order to find clusters of the $F0$ contours, which are difficult to classify.

References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP '96*, vol. 1, Atlanta, Georgia, May 1996, pp. 373–376.

[2] T. Dutoit, "Corpus-based speech synthesis," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer Berlin, Heidelberg, 2008, ch. 21, pp. 437–455.

[3] M. Legát and J. Matoušek, "Design of the test stimuli for the evaluation of concatenation cost functions," in *Proc. of the 12th International Conference TSD 2009, Lecture Notes in Artificial Intelligence*, vol. 5729. Springer Berlin / Heidelberg, 2009, pp. 339–346.

[4] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text–to–speech system ARTIC," in *Proc. of the 9th International Conference TSD 2006, Lecture Notes in Artificial Intellingence*, vol. 4188. Springer Berlin / Heidelberg, 2006, pp. 439–446.

[5] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, January 2001.

[6] J. R. Bellegarda, "A novel discontinuity metric for unit selection text–to–speech synthesis," in *SSW5 '04*, Pittsburgh, PA, June 2004, pp. 133–138.

[7] J. Vepa, S. King, and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," in *ICSLP '02*, Denver, Colorado, USA, September 2002, pp. 2605–2608.

[8] Y. Pantazis and Y. Stylianou, "On the detection of discontinuities in concatenative speech synthesis," in *Progress in Nonlinear Speech Processing*. Springer Berlin / Heidelberg, 2007, vol. 4391, ch. 6, pp. 89–100.

[9] M. Legát and J. Matoušek, "Collection and analysis of data for evaluation of concatenation cost functions," in *Proc. of the 13th International Conference TSD 2010, Lecture Notes in Artificial Intelligence*, vol. 6231. Germany: Springer Berlin / Heidelberg, 2010, pp. 345–352.

[10] ——, "Analysis of data collected in listening tests for the purpose of evaluation of concatenation cost functions," in *Proc. of the 14th International Conference TSD 2011, Lecture Notes in Artificial Intelligence*. Germany: Springer Berlin / Heidelberg, 2011, p. (accepted).

[11] M. Legát, J. Matoušek, and D. Tihelka, "On the detection of pitch marks using a robust multi-phase algorithm," *Speech Communication*, vol. 53, no. 4, pp. 552–566, April 2011.

[12] A. Ben-Hur and J. Weston, *A User's Guide to Support Vector Machines*. Springer Berlin / Heidelberg, 2010, ch. 13, pp. 223–239.