Removing Preglottalization from Unit-Selection Synthesis: Towards the Linguistic Naturalness of Synthetic Czech Speech

Jindřich Matoušek, Radek Skarnitzl, Daniel Tihelka, and Pavel Machač

Abstract—**This paper represents another step towards the linguistic naturalness of synthetic Czech. Its main goal is to eliminate undesirable occurrences of the so-called parasitic speechsounds, specifically preglottalization, from synthesized speech. After explaining the nature of parasitic speechsounds in Czech, we present procedures for both automatic detection and segmentation of these sounds in the source speech recordings. The main contribution of this paper consists in the proposal and implementation of two ways of synthesizing speech without preglottalization—cutting the parasitic sound from the signal and penalizing preglottalization during unit selection. Both these ways succeeded in suppressing the intrusiveness of preglottalization, with the latter method being evaluated as superior.**

Index Terms—**parasitic speech sound, preglottalization, linguistic naturalness, speech synthesis, unit selection.**

I. Introduction

CONCATENATIVE speech synthesis based on a unit-selection framework still appears to be the most popular approach to synthesize speech, at least as far as industrial applications are concerned [1], despite the fact that statistical parametric synthesis, especially based on the HMM framework [2], [3], is becoming increasingly more popular in the research domain. Contemporary unit-selection synthesis techniques employ very large speech corpora (for a comprehensive overview see e.g. [1] or [4]). The principle of unit-selection-based speech synthesis is to select the largest suitable segment of natural speech of the source speaker according to various phonetic, prosodic and positional criteria [5], [6], [7], [8], in order to prevent potential discontinuities in the synthesized speech; this aspect is what we may call the *technical naturalness* of concatenative synthesis. The quality of the synthesized outcome is therefore strongly dependent, apart from speech segmentation of the corpus, on the source speaker: his or her speaking style, as well as idiosyncratic habits, including potential non-standard phenomena, will be copied into the synthetic speech, and thus impair what we may call the *linguistic naturalness* of the outcome.

In any natural human activity, including speaking, we may encounter different kinds of imperfections. In speech, some “imperfections” may be perceived as neutral or even natural—this may concern for instance a certain degree of coarticulation [9] or of incomplete synchronization between glottal and articulatory activities [10]. Some imperfections may not be perceived at all, while others may not only be audible, but may have an intrusive influence on the listener.

In our preliminary study [11], we have identified in the recordings of the source speakers what we have called *parasitic sounds*, i.e., linguistically non-systematic sounds “attached” to a given speechsound and modifying its canonical realization in some way. Parasitic sounds arise in this sense as a result of a non-standard and phonetically unjustified coordination of glottal and articulatory gestures. It must be emphasized that these sounds occur very rarely in ordinary neutral Czech speech. When they do occur in normal conversation, they typically signal the speaker’s strong affective state. Paradoxically, though, these parasitic phenomena have become widespread in the speech of Czech TV and radio broadcasters who—as professionals—tend to be used as source speakers in corpus-oriented speech synthesis systems. Let us clarify once more that these phenomena in Czech have nothing in common with paralinguistic phenomena like fillers, wrappers, backchannelling, hesitation sounds, dysfluencies, filled pauses, etc. Those phenomena, abundantly manifested in spontaneous conversational speech [12], [13] and researched in the context of expressive or spontaneous speech synthesis (see e.g. [14], [15], [16]), are desirable in synthetic speech so as to increase its naturalness. This is not the case with our parasitic sounds—*preglottalization*, *postglottalization*, and *epenthetic schwa*: they cannot be considered to form a natural part of the Czech phonological system and they are, on the contrary, highly unnatural.

In [17], the form of these parasitic phenomena is thoroughly described and their classification is proposed. Briefly, *preglottalization* may be regarded mostly—though not exclusively—as a post-pausal phenomenon which involves non-standard fortification of the given consonant in the form of a glottal stop or, in the broader sense, glottalization [18] (e.g., [?dobri:] *good*), possibly also accompanied by a schwa-like vocalic element (e.g., [?@dobri:])[1]. *Postglottalization*, on the other hand, tends to occur mainly in the final, pre-pausal position; while glottalization is very frequent in

Manuscript received January 22, 2012. This research was supported by the Grant Agency of Czech Republic, project No. GAČR 102/09/0989. The access to the MetaCentrum computing facilities provided under the programme “Projects of Large Infrastructure for Research, Development, and Innovations” LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic is highly appreciated.

J. Matoušek and D. Tihelka are with the Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic, e-mail: {jmatouse, dtihelka}@kky.zcu.cz.

R. Skarnitzl and P. Machač are with the Institute of Phonetics, Faculty of Arts, Charles University in Prague, Czech Republic, e-mail: {radek.skarnitzl, pavel.machac}@ff.cuni.cz

[1] Canonically, the glottal stop may occur only before word- or morpheme-initial *vowels* in Czech.

(Advance online publication: 27 February 2012)

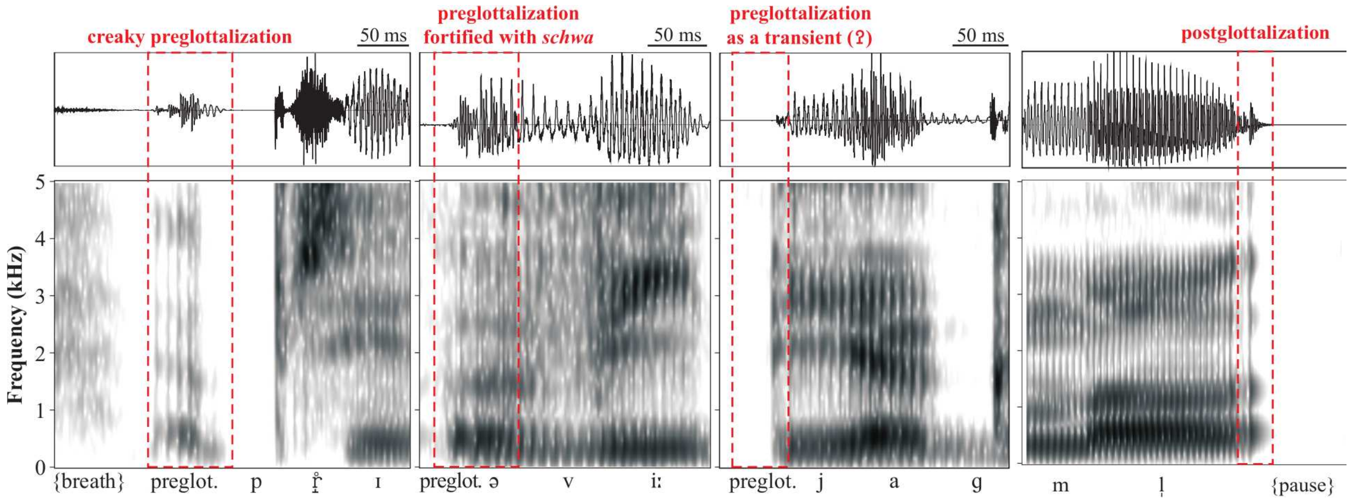

Fig. 1. Examples of parasitic sounds: preglottalization and postglottalization.

utterance-final positions [19], typically in the form of creaky phonation, additional postglottalization (i.e., after articulation has ceased) may sound unnatural in these positions when the vocal folds are firmly and abruptly pressed together at the end of a vowel, or when a vowel ends in a vocalic, schwa-like element. *Epenthetic schwa*, then, involves—for our purposes—the insertion of an unnaturally long vocalic element between consonants, sometimes even leading to the perception of an additional syllable (e.g., [pjet@ kluku:] *five boys*). Examples of parasitic sounds are given in Figure 1.

Subsequent perceptually oriented studies investigated parasitic phenomena from the perspective of their *perceptibility* [10] (i.e., whether listeners are, in fact, able to detect them in an AXB listening test paradigm) and the *degree of intrusiveness* [20]. Preglottalization turned out to manifest the greatest degree of intrusive effect. The presence of preglottalization in synthesized speech therefore is likely to create an impression of affectedness and to disturb the natural character of speech, especially when they are cumulated. For an example of preglottalization see Figure 1 or Figure 6a in Section V. The appearance of epenthetic schwa was most disturbing between voiceless consonants. Given the fact that postglottalization is comparatively rarer in our speech corpora (cf. Section II), its intrusiveness was not investigated in our studies.

It is obvious that, due to the enormous size of speech corpora employed in contemporary unit-selection-based speech synthesis (usually more than 10 hours of speech), manual annotation of parasitic sounds is almost impossible. Therefore, the parasitic sounds are hidden in the corpora and, following the principle of unit selection, they may unintentionally find their way into the synthesized speech. This may lead to two kinds of problems. First, as already mentioned, the presence of parasitic sounds (especially preglottalization) may disturb the natural character and fluency of speech. Second, when such parasitic sounds are not detected in the source recordings, speech contexts in which the parasitic sounds could appear are to be synthesized with no a priori information about the presence of such a sound. As a result, it is possible that speech contexts both with and without the described phenomena are concatenated, which would most likely be perceived as discontinuity in the synthesized speech.

It follows that it would be beneficial to have at one's disposal information about the actual presence or absence of a parasitic sound in a given context. With this information, one can try to avoid using such speech contexts in unit-selection synthesis—if the position of the parasitic sound is known, it may be cut out of the speech signal, or the particular speech unit containing the parasitic sound may be penalized during the unit selection mechanism. In some limited applications, for example when synthesizing highly affective speech, a unit containing a parasitic sound may even be used intentionally so as to increase the naturalness in that particular situation.

Procedures for both the automatic detection of the presence of parasitic sounds and the automatic determination of their boundaries in speech signals were designed in [11] and [21], respectively, and they are briefly recalled in Section II and III. In Section IV, we present the next step in our attempt to synthesize linguistically more natural speech as we describe two approaches to speech synthesis without the intrusive preglottalization sounds. These attempts are evaluated in Section V, and conclusions are drawn in Section VI.

II. Automatic Detection of Parasitic Phenomena

For the purpose of this study, we utilized randomly selected recordings of the source male speaker used as the primary voice in the Czech TTS system ARTIC [6], corresponding in total to approximately 14 minutes of read speech. The recordings were manually searched for instances of parasitic sounds. We identified 123 instances of preglottalization, 71 instances of epenthetic schwa, and 45 instances of postglottalization (in our other source speaker, however, postglottalization featured only four times). The boundaries of all these phenomena were determined in source speech signals (see [11], as well as [22] for a more detailed description of segmentation criteria). The next aim was to detect parasitic sounds automatically in the speech signals. Two different kinds of classifiers were used to this end: an HMM-based classifier and a BVM classifier. Both types of classifiers were trained on the same training data set and evaluated on the test data set as described in [11].

The HMM-based classifier follows the well-established techniques known from the field of automatic speech recognition (ASR) and automatic phonetic segmentation (APS). As this classifier was also utilized for the automatic segmentation of preglottalization and postglottalization, it is described here in more detail. In this framework, each phone or speechsound is modelled by a hidden Markov model (HMM)—firstly, the parameters of each HMM are estimated; then, *forced alignment* based on Viterbi decoding is performed to find the best alignment between the HMMs and the corresponding speech data.

In our experiments, a set of single-speaker three-state left-to-right context-independent multiple-mixture HMMs was employed, corresponding to all standard Czech phones plus the parasitic sounds. For the estimation of model parameters, we employed isolated-unit training utilizing Baum-Welch algorithm with model boundaries fixed to those labelled manually. For each utterance from the test data (described by feature vectors of mel frequency cepstral coefficients, MFCCs, extracted each 4 ms), the trained HMMs of all phones and of the parasitic sounds were concatenated according to the phonetic transcription of the utterance and aligned with the speech signal by means of Viterbi decoding. In this way, the best alignment between HMMs and the corresponding speech data is found, producing a set of boundaries which delimit the speechsounds belonging to each HMM. Thus, the position of each phone-like unit—including the parasitic phenomena—is identified in the stream of speech signal. Within this process, the automatic detection of the presence of each parasitic sound is carried out by creating multiple phonetic transcripts per utterance with all combinations of the presence/absence of the individual parasitic sounds in the defined target contexts. Consequently, the transcript which "best matches" the data is chosen as the maximum likelihood estimation (MLE) of the utterance. Using this procedure, it was possible to detect parasitic sounds in the given contexts (see Figure 2 for a schematic view of the detection process).

Ball Vector Machines (BVM) is a simplified version of Core Vector Machines (CVM) classification method from the family of kernel methods. Unlike the computationally demanding SVM, a CVM finds an approximative solution by applying methods of computational geometry. The training phase is formulated as finding an approximation of the *minimum enclosing ball* (MEB), or specifically, its so called $(1+\varepsilon)$-approximation. A BVM further simplifies the problem by finding a $(1+\varepsilon)$-approximation of *enclosing ball* (EB) with a fixed radius instead of MEB (see [23] for more detail). The reason why we have chosen a kernel-based classifier is that it often outperforms the other types of classifiers [24]. We used the RBF (radial basis function) kernel in our BVM classifier.

In order to obtain the input features for the classifier, we employed the TRAPS parametrization technique in our experiments. Such a technique enables the classifier to take long-term temporal trajectories into account. We used the setup similar to [25]. To ensure better granularity, the parametrization was modified to obtain the feature vectors each 4 ms. Using the same manually labelled time-aligned data as for the HMM-based classifier, we identified positive and negative examples for the BVM classifier. Eight feature vectors closest to the centre of the given parasitic sound were used as the positive examples, while as the negative examples we used those eight feature vectors which lay closest to the boundary where the given sound may occur but actually did not. The parameters of the BVM classifier were determined using the grid-search algorithm with 10-fold cross-validation.

The evaluation of the automatic classification was performed in a "standard" way, i.e. using true positive rate (TPR, i.e. hit rate), false positive rate (FPR, i.e. false alarm rate) and detection accuracy

$$ACC = \frac{P \cdot TPR + N \cdot (1 - FPR)}{P + N}, \qquad (1)$$

where P is the number of "positive examples" in the test data (i.e. how many times the parasitic sound really occurred in the given context) and N is the number of "negative examples" in the test data (i.e. how many times the parasitic sound could occur in the given context but actually did not occur (N)). In order to take into account also the classification "accuracy" which occurred by chance, Cohen's kappa κ is also indicated (generally, $\kappa \geq 0.70$ is considered satisfactory). The results of the detection are summarized in Tables I-III and discussed in [11] in more detail. The slightly different numbers N of negative examples are caused by different pre-processing of the data for the two classifiers.

TABLE I
Results of the automatic detection of preglottalization.

Detection rates	HMM	BVM
P	50	50
N	56	59
TPR	0.92	0.92
FPR	0.11	0.02
ACC	0.91	0.95
chance level	0.50	0.51
κ	0.81	0.91

TABLE II
Results of the automatic detection of postglottalization.

Detection rates	HMM	BVM
P	26	26
N	106	132
TPR	0.77	0.96
FPR	0.02	0.00
ACC	0.94	0.99
chance level	0.70	0.73
κ	0.70	0.98

TABLE III
Results of the automatic detection of epenthetic schwa.

Detection rates	HMM
P	17
N	36
TPR	0.29
FPR	0.11
ACC	0.70
chance level	0.62
κ	0.21

While glottalization is relatively well defined in terms of the context of occurrence (pre-pausal and post-pausal), epenthetic schwa may occur in various contexts whose

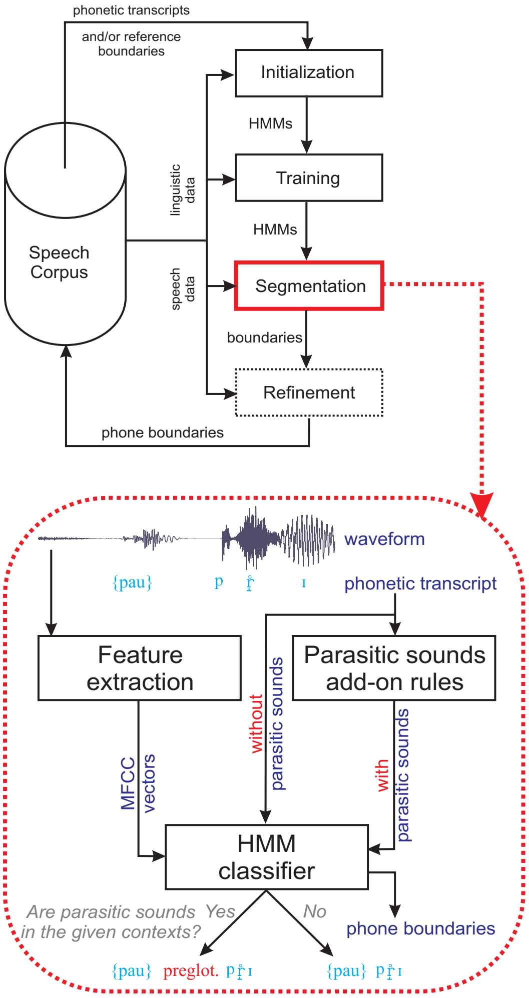

Fig. 2. Simplified scheme of HMM-based automatic phonetic segmentation including the detection of parasitic sounds.

delimitation is less straightforward. Our analyses indicate that epenthetic schwa occurs most frequently in a) prepausal sonorant consonants, b) between two speechsounds of the same place of articulation, and c) after a non-syllabic preposition; only these contexts were then taken into account in our initial experiments with the automatic detection of schwa. Since the BVM classifier could suffer from the training set being heavily biased towards negative examples, we only employed the HMM-based classifier for automatic detection at this stage.

As the results presented in the tables indicate, automatic detection of glottalization phenomena was quite successful (cf. especially the values of Cohen's kappa). The detection of epenthetic schwa, on the other hand, was poorer. This may have been caused partly by the fact that the context in which epenthetic schwa may occur is considerably more varied. Moreover, the acoustic contrast between schwa and some of the possible co-occurring speechsounds is very low, especially in the case of sonorant consonants.

For this reason, we will focus only on the glottalization phenomena in the following sections. Since it appears that it is especially preglottalization which may have an intrusive effect on listeners [20], we ran a separate automatic detection on our entire speech corpus which includes the total of 12,065 source recordings. Automatic detection of preglottalization yielded 9,075 instances of preglottalization in 6,819 recordings. This means, in other words, that more than one half of the source recordings we have been using lately for synthesizing Czech speech appear to include preglottalization.

TABLE IV
Results of the automatic segmentation of preglottalization and postglottalization sounds.

	MAE (ms)	RMSE (ms)	Tol10 (%)	Tol20 (%)
PRG−⋆	7.50	10.56	83.33	90.48
⋆−POS	8.33	11.16	65.38	92.31
GST−⋆	7.00	9.39	73.08	96.15
⋆−⋆	6.45	11.66	82.40	94.65

III. Automatic Segmentation of Glottalization

Essentially, there were two options regarding the segmentation of glottalization: it may be carried out within the HMM-based detection process (respecting multiple phonetic transcriptions which distinguish between the presence/absence of a glottalization sound as described in Section II), or after the HMM- or BVM-based detection only on those speech contexts in which preglottalization had already been detected. In our experiments, the segmentation was performed within the HMM-based detection process [21]. Though the accuracy of the detection of the HMM-based classifier was slightly worse when compared to the BVM classifier (see Tables I and II), one of the advantages of the HMM-based classifier is that, as boundaries between HMMs are produced during the alignment, the position of each modelled sound in the utterance could be located. A simplified scheme of the automatic phonetic segmentation utilizing the HMM-based classifier is shown in Figure 2. Optionally, it is possible to subsequently refine the boundaries obtained by the HMM-based classifier, as described e.g. in [26], [27].

The results of the automatic segmentation of preglottalization (PRG) and postglottalization (POG) in terms of mean absolute error (MAE), root mean square error (RMSE) and percentage of boundaries deviating less than the 10ms (Tol10) or 20ms (Tol20) tolerance region are shown in Table IV. Notice that only the ending boundaries of preglottalization (PRG−⋆) and the starting boundaries of postglottalization (⋆−POG) are specified. The other types of boundaries (⋆−PRG and POG−⋆) are located in pauses (see Figure 1), and, due to the smooth concatenation of speech signals in silence, the precise identification of these boundaries is of lower importance. So as to allow for comparison, the segmentation accuracy of a phonetically similar unit, the pre-vocalically occurring glottal stop (GST−⋆) is also indicated in Table IV, as well as the average segmentation accuracy of all other Czech phonetic units (⋆−⋆) . The comparison of the segmentation accuracy of all boundary types is shown in Figure 3.

Looking at the results of the automatic segmentation in Table IV and in Figure 3, it can be shown that:

- For both speech corpora, the segmentation accuracy of preglottalization (PRG−⋆) is comparable to the segmentation accuracy of glottal stop (GST−⋆), a phonetic unit similar to preglottalization, which has already been used in synthesis of Czech speech (according to the

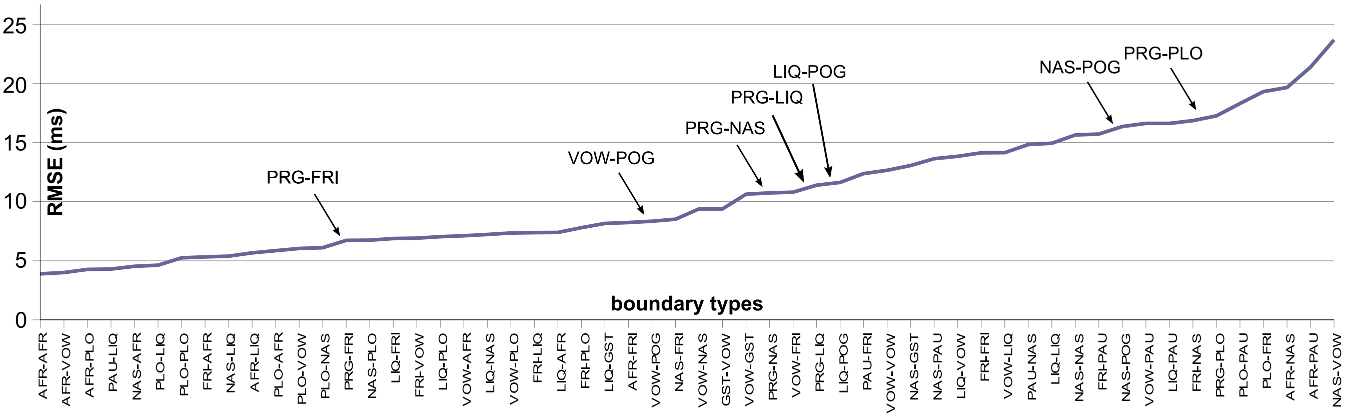

Fig. 3. Comparison of the automatic segmentation accuracy of different boundaries types in terms of RMSE (PRG = preglottalization, POG = postglottalization, VOW = vowels, FRI = fricatives, PLO = plosives, AFR = affricates, NAS = nasals, LIQ = liquids)

unpaired *t*–test the difference in MAE is not statistically significant, two-tailed P-value = 0.8095).

- Comparing the segmentation accuracy of (`PRG-*`) to the average segmentation accuracy of all other phonetic boundaries (`*-*`), preglottalization tends to be worse in terms of MAE but, on the other hand, it tends to be better in terms of RMSE (with the difference in MAE being not statistically significant, unpaired *t*–test, two-tailed P-value = 0.4876).
- The segmentation of postglottalization (`*-POG`) is less accurate than the segmentation of preglottalization (the difference is statistically not significant, unpaired *t*–test, two-tailed P-value = 0.6607).
- Segmentation results in Figure 3 confirm that the segmentation of both preglottalization and postglottalization sounds does not deviate from the segmentation of all other phone sounds.

Moreover, the average segmentation accuracy of the automatic phonetic segmentation (APS) system with the parasitic sounds included (MAE = 6.45 ms, RMSE = 11.66 ms) is better than the average segmentation accuracy of the standard APS system with no parasitic sounds included (MAE = 6.71 ms, RMSE = 16.21 ms), which means that explicit modelling of parasitic sounds does increase the accuracy of the segmentation of other phones (although the difference is statistically not significant, unpaired *t*–test, two-tailed P-value = 0.5879).

To sum up, the results indicate that, based on the automatic segmentation, it should be possible to remove preglottalization from the speech signals and thus to prevent this parasitic sound from being transferred into synthesized speech.

IV. Speech synthesis without preglottalization

In previous sections, we described procedures designed to detect preglottalization phenomena in our speech corpus and to find their boundaries in the signal. With this information at our disposal, it was possible to consider ways in which preglottalization could be eliminated from synthesized speech, thus yielding a linguistically more natural outcome. We proposed two scenarios, both based on the unit-selection framework [28]. The first scenario employs the standard unit-selection mechanism, with the resulting speech signal subsequently being post-processed—preglottalization sounds are physically removed, or cut out (see Section IV-A). The second scenario employs a modified unit-selection mechanism in which the items containing preglottalization are penalized during the selection process, so that they are less likely to appear in the resulting speech signal (see Section IV-B).

A. *Cutting out preglottalization*

In this first experiment, speech was synthesized with standard settings of our unit-selection TTS system, as described e.g. in [29], [30]. Figure 6a shows a specific example illustrating the fact that undesirable parasitic preglottalization phenomena may find their way into synthesized speech. In order to obtain synthetic speech free of preglottalization, it was necessary to employ post-processing consisting in cutting the signal corresponding to preglottalization out of the speech signal. To do that, an overlap-add-like procedure was applied as illustrated in Figure 4.

From what we have said, it is obvious that this approach requires not only the knowledge concerning the presence of preglottalization in the synthesized contexts (see Section II), but also concerning its precise location in the source speech units (and, by extension, also in the synthesized speech signal—see Section III). It should be pointed out at this stage that the primary objective of this experiment was to determine the true potential of the cutting out procedure. Therefore, in order to avoid synthesis errors caused by imperfect automatic segmentation of preglottalization, we made use of manually performed segmentation from expert phoneticians throughout this experiment.

The successfulness of this method in removing preglottalization from the resulting synthetic speech is evaluated further in Section V. The advantage of this method is that a standard, well-tuned unit selection mechanism can be utilized. On the other hand, the need for (very precise) automatic segmentation of the parasitic glottalization sounds can be viewed as a clear disadvantage. It is also worth mentioning that the decision to cut preglottalization on-line was deliberately preferred to off-line cutting out, since the latter option would require storing another large, modified speech unit database.

B. *Penalization of preglottalization*

The idea behind this second experiment is the notion of producing linguistically natural synthetic speech by gener-

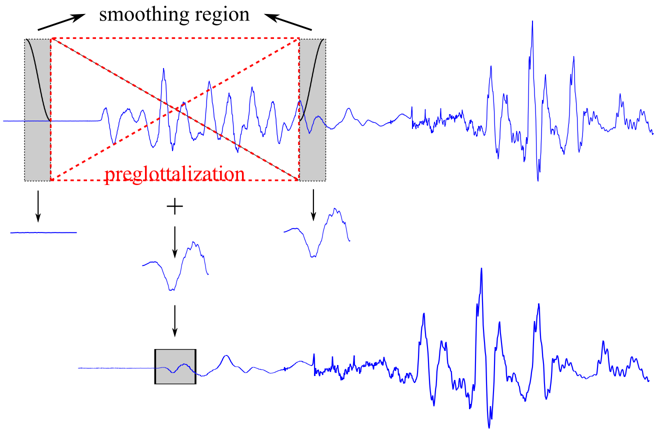

Fig. 4. An illustration of the cutting-out algorithm. The undesirable preglottalization sound is cut out of the synthetic signal (the upper part of figure), and the remaining parts of the signal are smoothly concatenated within a smoothing region. The resulting signal is shown at the bottom of the figure.

ating it from clear, unmarked, preglottalization-free speech segments. This technique would thus not be affected by the somewhat cumbersome and—due to automatic segmentation possibly even imprecise—cutting-out process described in the previous section. To this end, we proposed a modified unit selection scheme which consisted in the addition of another criterion into the unit-selection algorithm—the knowledge of the presence/absence of preglottalization in each diphone candidate (see Figure 5). In order to minimize the chance of a diphone with preglottalization being selected, the unit-selection algorithm was tuned to prefer a diphone candidate free of preglottalization phenomena. Penalization of this criterion was set as very high in comparison with other criteria (phonetic and prosodic contexts). Therefore, diphone candidates with the undesirable preglottalization should be selected only when other criteria fail (actually, such a case never occurred for our test utterances—see Section V).

Note that, unlike in the previous experiment, it is not necessary to have information about the boundaries of preglottalization in the source recordings, nor in the synthetic speech. The only additional information which is required as compared to the standard speech synthesis system is the knowledge of the presence/absence of preglottalization in the source diphone candidates; as described in Section II, that is something that can be obtained automatically. Similarly—as in Section IV-A—manual detection of preglottalization was utilized throughout this experiment, so as to avoid errors caused by automatic processes.

The comparison between both approaches to speech synthesis is provided further in Section V. The advantage of the approach described in this subsection is clear—the location of the parasitic preglottalization phenomena is not needed, and their automatic segmentation thus need not be provided. On the other hand, it is necessary to modify the well-established unit-selection algorithm (possibly with some amount of experiments needed to fine-tune the modified algorithm).

V. Evaluation & discussion

To emphasize the need for special handling of the parasitic preglottalization in Czech speech synthesis—in other words,

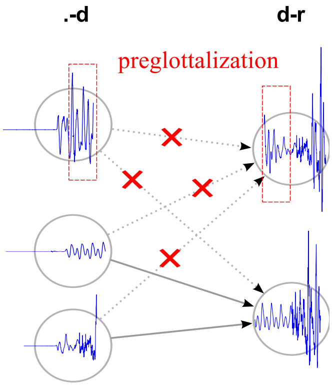

Fig. 5. An illustration of the modified unit-selection algorithm. In order to obtain synthetic speech free of undesirable preglottalization, segments containing preglottalization are penalized during unit selection.

to illustrate the point that we are not talking about a marginal, negligible phenomenon in the speech corpus—approximately 965k unique sentences were synthesized using the original version of our TTS system. Subsequently, statistics about the usage of each diphone from the speech corpus were recorded. In this way, 335k sentences were identified to contain at least one half of any of the preglottalization items from the speech corpus. This means that every third sentence synthesized by our TTS system contained preglottalization.

To evaluate the impact of preglottalization on the quality of the resulting synthetic speech, we selected 18 representative sentences for further analysis. Each of these sentences was then synthesized with the three versions of our speech synthesis system—the original system (ORG), in which preglottalization was not handled, and two versions in which preglottalization was cut out (CUT, see Section IV-A) or penalized during unit selection (PEN, see Section IV-B). An example of these three synthetic versions of one sentence is given in Figure 6. The resulting synthetic sentences were analyzed by the two phoneticians in the author team, and the perceptual effect stemming from the potential presence of preglottalization was marked as *not intrusive*, *slightly intrusive* or as *very intrusive*.

Table V shows that 11 of the 18 sentences contained intrusive preglottalization when synthesized with the original, unmodified speech synthesis system (ORG), and that five of these eleven instances were perceived as very intrusive. It can also be seen that both the proposed methods succeeded in suppressing the intrusiveness of the parasitic preglottalization—only two sentences still contained audible preglottalization after preglottalization sounds had been removed from the corresponding synthetic speech signal using the cutting out procedure (CUT), and no preglottalization at all was audible when source units containing preglottalization

TABLE V
Comparison of synthetic speech of 18 test sentences with respect to the intrusiveness of preglottalization.

	Intrusiveness		
Synthesis scenarios	None	Slightly	Very
ORG	7	6	5
CUT	16	1	1
PEN	18	0	0

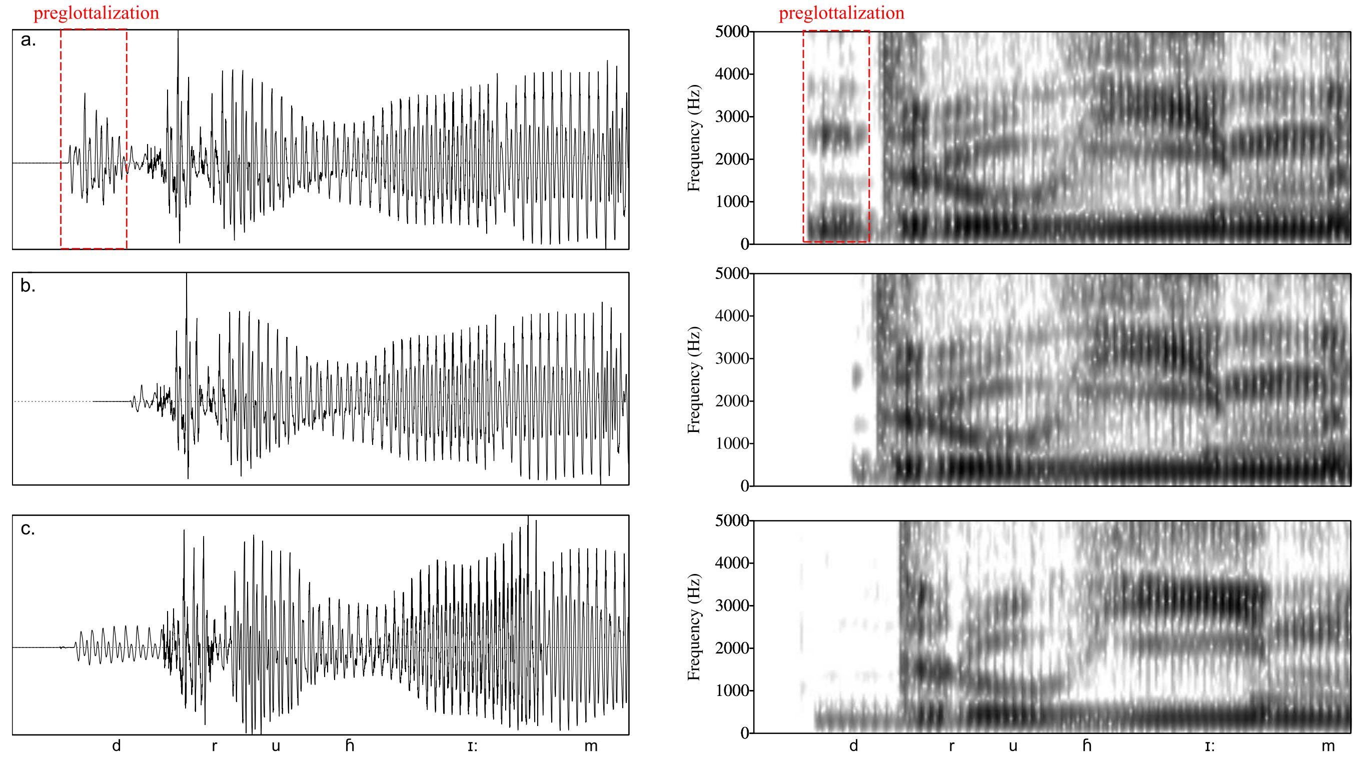

Fig. 6. Examples of synthetic speech: a. using the original system with a preglottalization sound included (ORG); b. with the preglottalization sound removed (CUT); c. with preglottalization diphones penalized during unit selection (PEN).

had been penalized during unit selection (PEN). The persisting presence of preglottalization in the two units—even after their removal using the cutting out procedure—may have two causes. First, it is possible that, although segmentation of preglottalization phenomena had been performed manually by phoneticians in this experiment, the segmentation may have been imperfect. Second, it is simply possible that some sort of audible parasitic phenomenon arises in the concatenation process although the signal corresponding to preglottalization has been removed.

It is obvious—and it can also be seen in Figure 6—that penalizing preglottalization in the unit-selection process leads to the selection of different diphone segments. In other words, the resulting synthesized versions, after the application of the CUT and PEN procedures, sound differently. Although PEN outperforms CUT in the suppression of the intrusive effect of preglottalization, the forced usage of different—and from the standard viewpoint less ideal—diphone candidates may degrade the *overall quality* of the resulting speech. We regarded it therefore as necessary to conduct another informal listening test, so as to compare the overall quality of synthetic speech produced by both the PEN and CUT synthesis scenarios. The results, shown in Figure 7, indicate that PEN outperforms CUT also with respect to the overall quality: no PEN version has been assessed as being worse in overall quality than CUT, while a PEN version was preferred to the respective version in five cases. In the remaining 13 cases, the two synthetic versions were judged as equivalent in their overall quality.

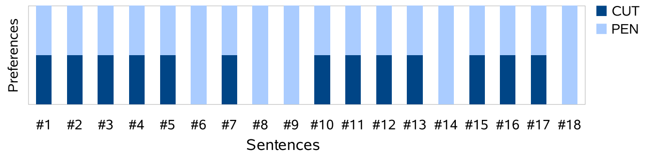

Fig. 7. Results of preference listening test comparing the overall quality of synthesized speech produced by the PEN and CUT speech synthesis scenarios.

It is worth pointing out that the procedure consisting in penalizing units with preglottalization sounds is likely to have even more of an edge over the cut-out algorithm in real-life applications. Since we used manually determined boundaries in our experiments—and their precise knowledge is necessary only for the cut-out procedure—the performance of the cut-out procedure is likely to drop further once automatic segmentation of parasitic sounds is applied.

VI. Conclusion

This paper is part of a larger endeavour focused on improving the linguistic naturalness of synthetic Czech. Previous investigations [10], [20] have verified the perceptual intrusiveness of what we have described as *parasitic phenomena*, especially preglottalization and epenthetic schwa. While essentially non-existent in neutral, unmarked speech of ordinary speakers, these parasitic sounds have been shown to be quite frequent in the speech of professionals from the media who, in turn, are often recruited to record large corpora for speech synthesis. It was therefore desirable to remove the parasitic sounds from the synthetic speech. This paper focused specifically on preglottalization; one reason for not including epenthetic schwa was the fact that its automatic detection was comparatively lower (see [11]), the second reason is the higher degree of perceived intrusive effect of preglottalization. Postglottalization was much less frequent in our corpora, and our informal observations indicated a lower degree of intrusive effect.

The first step in our analysis consisted in the automatic detection and segmentation of all three parasitic phenomena. Given the lower detection rate of epenthetic schwa, we subsequently focused on the segmentation of only the two forms of glottalization. In the next stage, two speech synthesis scenarios were proposed and employed to synthesize speech without preglottalization, the most intrusive of all the three parasitic speechsounds. The first scenario was based on actually removing the signal corresponding to preglottalization from the synthesized speech, the second exploited penalizing speech units containing preglottalization during selection. The results of our experiments with sample sentences were encouraging in that the synthesized speech was evaluated as linguistically more natural than the speech produced by the original system. Comparing the two synthesis scenarios, penalization of preglottalization during unit selection appears to provide better results, at least for two reasons. First of all, penalizing the target units outperformed cutting preglottalization out of the synthetic speech, and this concerns the ability to suppress the intrusive effect, as well as the overall quality of the resulting synthetic speech. Second, the procedure penalizing preglottalization does not require the knowledge of the precise boundary of the parasitic preglottalization sounds in the signal. On the other hand, some fine-tuning of the unit-selection algorithm may be necessary to find an optimal trade-off between the ability to suppress preglottalization and the overall quality of synthetic speech.

In our future work, we will also conduct experiments with other source speakers used in our TTS system and, most importantly, utilize the knowledge acquired in this research in a real Czech TTS system. We will also investigate the effect of automatic procedures for both the detection and segmentation of preglottalization on the quality of the resulting synthetic speech.

REFERENCES

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

[2] K. Tokuda, H. Zen, and A. Black, "An HMM-based approach to multilingual speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds. New Jersey: Prentice Hall, 2004, pp. 135–153.

[3] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[5] A. Hunt and A. Black, "Unit selection in concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, USA, 1996, pp. 373–376.

[6] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text-to-speech system ARTIC," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, vol. 4188, pp. 439–446.

[7] J. Romportl, J. Matoušek, and D. Tihelka, "Advanced prosody modelling," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, vol. 3206, pp. 441–447.

[8] J. Romportl, "Prosodic phrases and semantic accents in speech corpus for Czech TTS synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, vol. 5246, pp. 493–500.

[9] B. Kühnert and F. Nolan, "The origin of coarticulation." Cambridge University Press, 1999, pp. 7–30.

[10] R. Skarnitzl and P. Machač, "Domain-initial coordination of phonation and articulation in czech radio speech," *AUC Philologica*, vol. 12, no. 1/2009, pp. 21–35, 2010.

[11] J. Matoušek, R. Skarnitzl, P. Machač, and J. Trmal, "Identification and automatic detection of parasitic speech sounds," in *Proc. INTERSPEECH*, Brighton, Great Britain, 2009, pp. 876–879.

[12] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *J. Int. Phon. Ass.*, vol. 31, no. 1, pp. 153–169, 2001.

[13] J. E. Arnold, M. Fagnano, and M. K. Tanenhaus, "Disfluencies signal theee, um, new information," *J. Psycholinguist. Res.*, vol. 32, no. 1, pp. 25–36, January 2003.

[14] N. Campbell, "Towards synthesising expressive speech; designing and collecting expressive speech data," in *Proc. INTERSPEECH*, Geneve, Switzerland, 2003, pp. 1637–1640.

[15] R. Carlson, K. Gustafson, and E. Strangert, "Cues for hesitation in speech synthesis," in *Proc. INTERSPEECH*, Pittsburgh, USA, 2006, pp. 1300–1303.

[16] J. Adell, A. Bonafonte, and D. Escudero, "Synthesis of filled pauses based on a disfluent speech model," in *Proc. ICASSP*, Dallas, USA, 2010, pp. 4810–4813.

[17] P. Machač and R. Skarnitzl, "Phonetic analysis of parasitic speech sounds," in *Proc. Czech-German Workshop Speech Process.*, Prague, Czech Rep., 2009, pp. 61–68.

[18] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407–429, Oct. 2001.

[19] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, no. 4, pp. 423–444, 1996.

[20] R. Skarnitzl and P. Machač, "Míra rušivosti parazitních zvuků v řeči mediálních mluvčích," *Naše Řeč*, 2012, (in Czech; in print).

[21] J. Matoušek, "Automatic segmentation of parasitic sounds in speech corpora for TTS synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, vol. 6231, pp. 369–376.

[22] P. Machač and R. Skarnitzl, *Principles of Phonetic Segmentation*. Prague: Epocha, 2009.

[23] I. W. Tsang, A. Kocsor, and J. T. Kwok, "Simpler core vector machines with enclosing balls," in *Proc. ICML*, Corvallis, Oregon, USA, 2007, pp. 911–918.

[24] J. Trmal, J. Zelinka, J. Psutka, and L. Müller, "Comparison between GMM and decision graphs based silence/speech detection method," in *Proc. SPECOM*, St. Petersburg, Russia, 2006, pp. 376–379.

[25] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, vol. 3206, pp. 465–472.

[26] S. S. Park and N. S. Kim, "On using multiple models for automatic speech segmentation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2202–2212, 2007.

[27] J. Matoušek and J. Romportl, "Automatic pitch-synchronous phonetic segmentation," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008.

[28] J. Matoušek, R. Skarnitzl, D. Tihelka, and P. Machač, "Towards linguistic naturalness of synthetic speech," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2011, WCECS 2011*, San Francisco, USA, 2011, pp. 561–566.

[29] D. Tihelka and J. Matoušek, "Unit selection and its relation to symbolic prosody: a new approach," in *Proc. INTERSPEECH*, Pittsburgh, USA, 2006, pp. 2042–2045.

[30] D. Tihelka, J. Kala, and J. Matoušek, "Enhancements of Viterbi search for fast unit selection synthesis," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 174–177.