Towards Linguistic Naturalness of Synthetic Speech

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

***Abstract*—This paper presents another step towards linguistic naturalness of synthetic Czech. The main goal of this study is to avoid unintended occurrences of parasitic speech sounds (namely preglottalization) in synthesised speech. Firstly, we explain what we mean by the term parasitic speech sound. Secondly, procedures for both automatic detection and segmentation of these sounds in source speech recordings are presented. Then, two speech synthesis scenarios are proposed and employed to synthesise speech without preglottalization. Both scenarios succeeded in suppressing intrusiveness of preglottalization, with the scenario of penalisation of preglottalization during unit selection being evaluated as the better one.**

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

I. Introduction

CONTEMPORARY concatenative speech synthesis techniques based on a unit-selection framework employ very large speech corpora (for a comprehensive overview see e.g. [1] or [2]). As 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 [3], [4], [5], in order to prevent potential discontinuities (i.e., what we may call the *technical naturalness* of concatenative synthesis), the synthesised outcome is strongly dependent on the source speaker: his or her speaking style and idiosyncractic habits, including potential non-standard phenomena, are 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, some may not be perceived at all, while others may have an intrusive influence on the listener. In [6], 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 it 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 emphasised that these sounds occur very rarely in ordinary neutral speech; when they do occur in normal conversation, they signal the speaker's strong affective state. Paradoxically, though, these parasitic phenomena are widespread in the speech of Czech TV and radio broadcasters who—as professionals—tend to be used as source speakers in speech synthesis systems. Let us clarify that these phenomena in Czech are really a consequence of imperfect speaking and have nothing in common with paralinguistic phenomena like fillers, wrappers, backchannelling, hesitation sounds, disfluencies, filled pauses, etc. present in spontaneous conversational speech and researched in the context of expressive or spontaneous speech synthesis (see e.g. [7], [8], [9]).

The parasitic phenomena—*preglottalization*, *postglottalization*, and *epenthetic schwa*—are thoroughly described and classified in [10]. It should be emphasised once more that these speechsound modifications cannot be considered to form a natural part of the Czech phonological system: on the contrary, they are highly unnatural. In perceptually oriented studies, we investigated their perceptibility [11] and the degree of intrusiveness [12]. Preglottalization—non-standard fortification of the following consonant in the form of a glottal stop, possibly also accompanied by a schwa-like vocalic element—turned out to manifest the greatest degree of intrusive effect. (The glottal stop may occur only before word- or morpheme-initial *vowels* in Czech.) The presence of preglottalization in synthesised speech therefore creates an impression of affectedness and disturbs the natural character of speech, especially when they are cumulated. For an example of preglottalization see Figure 6a in Section V.

It is obvious that, due to the enormous size of present speech corpora employed in 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 can unintentionally get into the synthesised speech. In fact, two kinds of problems can then arise. Firstly, as already mentioned, the presence of parasitic sounds (namely preglottalization) can disturb the natural character and fluency of speech. Secondly, when such parasitic sounds are not detected in the source recordings, speech contexts in which the parasitic sounds could appear are to be synthesised with no a priori information about the presence of such a sound. As a result, the speech contexts both with and without the described phenomena could be concatenated, which would be most likely perceived as a discontinuity in synthetic speech.

Having information about the presence/absence of a parasitic sound in a given context, one can avoid using such speech contexts in unit-selection synthesis—if the position of the parasitic sound is known, it could be cut out of the speech signal, or the particular speech unit containing the parasitic sound could be penalised during the unit selection

Manuscript received June 24, 2011; revised August 11, 2011. 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

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

mechanism, or, even, such a unit could be intentionally used in speech synthesis in order to increase the naturalness of synthetic speech in some limited applications.

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 [6] and [13], respectively, and they are briefly recalled in Section II and III. In Section IV, we present a next step in our attempt to synthesise linguistically 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 preglottalization

For the purpose of this study, randomly selected recordings of a source male speaker used as the "main" voice in the Czech TTS system ARTIC [3], in total approx. 14 minutes of read speech were utilised. The recordings were analysed with the aim to identify preglottalization. Consequently, 123 instances of preglottalization were found, and their boundaries in source speech signals were determined (see [6] for a more detailed description).

The aim of the automatic detection of parasitic preglottalization sounds was to detect, or identify the presence of these parasitic sounds in speech signals. Two different kinds of classifiers were used: 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 [6].

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 utilised for the automatic segmentation of preglottalization, it is described here in more detail. In this framework each phone or sound 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 corresponding to all Czech phones and preglottalization was employed. For the estimation of model parameters, we employed isolated-unit training utilising Baum-Welch algorithm with model boundaries fixed to the hand-labelled ones. 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 preglottalization were concatenated according to the phonetic transcription of the utterance and aligned with a 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 speech sounds belonging to each HMM. Thus, the position of each phone-like unit and of preglottalization is identified in the stream of speech signal. Within this process, the automatic detection of the presence of each preglottalization sound is carried out by creating multiple phonetic transcripts per utterance with all combinations of the presence/absence of preglottalization in the defined contexts. Consequently, the transcript which "best matches" the data is chosen as the maximum likelihood estimation (MLE) of the utterance. In this way, preglottalization in given contexts could be detected.

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, 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. BVM further simplifies the problem by finding a $(1+\varepsilon)$-approximation of *enclosing ball* (EB) with a fixed radius instead of MEB. For greater details, see [14]. The reason why we have chosen a kernel based classifier is that it often outperforms the other types of classifiers [15]. We used RBF (radial basis function) kernel in the BVM classifier.

In our experiments, the TRAPS parametrisation technique was employed to obtain the input features for the classifier. Such a technique enables the classifier to take the long-term temporal trajectories into account. We used the setup similar to [16]. To ensure better granularity, the parametrisation was modified to obtain the feature vectors each 4 ms. Using the same hand-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. As the negative examples, eight feature vectors closest to the boundary where the given sound is possible to occur but actually did not were used. The parameters of BVM classifier were determined using 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 also the classification "accuracy" occurred by chance into account, Cohen's kappa κ is also indicated (generally, $\kappa \geq 0.70$ is considered satisfactory).

TABLE I
Results of the automatic detection of preglottalization (slightly different numbers N of negative examples are caused by different pre-processing of the data for a particular classifier).

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

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

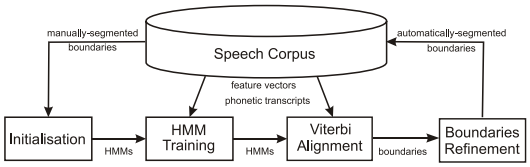

Fig. 1. Simplified scheme of HMM-based automatic phonetic segmentation.

The results of the detection are summarised in Table I and discussed in [6] in more detail.

Having applied the automatic detection of preglottalization on all 12,065 source recordings in our speech corpus, 9,075 instances of preglottalization were found in 6,819 recordings. It means that more than one half of the source recordings used lately for speech synthesis include preglottalization.

III. Automatic Segmentation of Preglottalization

The segmentation of preglottalization could 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 HMM- or BVM-based detection only on speech contexts with the detected preglottalization phenomenon. In our experiments, the segmentation was performed within the HMM-based detection process [13]. A simplified scheme of the automatic phonetic segmentation utilising the HMM-based classifier is shown in Figure 1. Optionally, the boundaries segmented by the HMM-based classifier can be refined as described e.g. in [17], [18].

The results suggest that the automatic segmentation of preglottalization (and also postglottalization) sounds is comparable to the segmentation of other phones (see Figure 2). It indicates that, based on the automatic segmentation, it should be possible to remove preglottalization from the speech signals and thus to prevent them from getting into synthesised speech. More detailed results and their discussion can be found in [13].

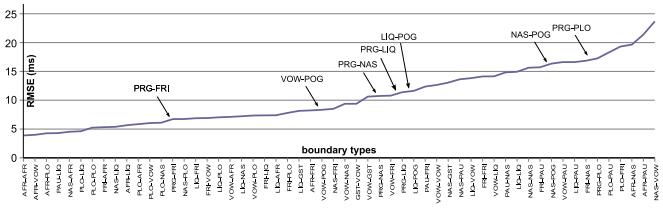

Fig. 2. Comparison of the automatic segmentation accuracy of different boundaries types in terms of RMSE (PRG = preglottalization, POG = postglottalization, VOW = vowels, FRI = fricatives, PLO = plosives, AFR = affricates, NAS = nasals, LIQ = liquids)

IV. Speech Synthesis Without Preglottalization

Two speech synthesis scenarios, both based on the unit-selection framework, were proposed to synthesise linguistically more natural speech without parasitic preglottalization phenomena. The first scenario employs standard unit-selection mechanism, and resulting speech signal is then post-processed—preglottalization sounds are cut out (see Section IV-A). The second scenario employs a modified unit-selection mechanism in which the items with preglottalization are penalised during the selection process and so they are likely not to be present in the resulting speech signal (see Section IV-B).

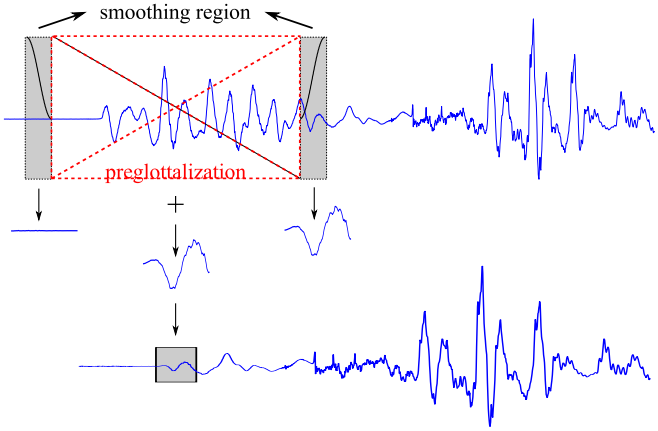

Fig. 3. An illustration of the cutting-out algorithm. The undesirable preglottalization sound is cut out of the synthetic signal (the upper part of figure), and the remaining parts of the signal are smoothly concatenated within a smoothing region. The resulting signal is shown at the bottom of the figure.

A. Cutting out preglottalization

In this experiment, speech was synthesised with a standard setting of our unit-selection TTS system as described e.g. in [19], [20]. As shown in Figure 6a., synthetic speech signal containing the undesirable parasitic preglottalization phenomena could be produced. In addition, to have synthetic speech free of preglottalization, a post-processing of resulting speech signal has to be done. The post-processing consisted in cutting the signal corresponding to preglottalization out of the synthetic speech. To do that, an overlap-add-like procedure was employed as illustrated in Figure 3.

It is obvious that this approach requires not only to know about the presence of preglottalization in the synthesised contexts (see Section II), but also its precise location in source speech units (and, thus also in the synthesised speech signal—see Section III). Let us note that, in order to avoid synthesis errors caused by imperfect automatic segmentation of preglottalization and to find the true potential of this method, manual segmentations from expert phoneticians were used throughout this experiment.

The successfulness of this method in removing preglottalization from synthetic speech is evaluated further in Section V. The advantage of this method is that a standard, well-tuned unit selection mechanism can be utilised. On the other hand, the need for (very precise) automatic segmentation of glottalization sounds can be viewed as a clear disadvantage.

B. Penalisation of preglottalization

The idea behind this experiment is to produce linguistically natural synthetic speech by making it up of linguistically clear, preglottalization-free speech segments not affected by the cutting-out process described in the previous section. Hence, a modified unit selection scheme was proposed. The modification 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 4). In order to minimise a chance that

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

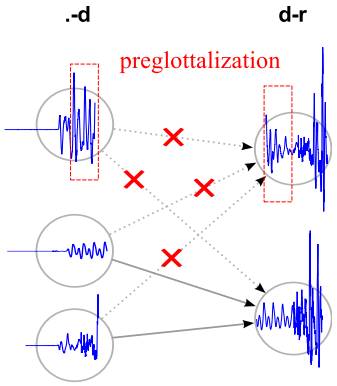

Fig. 4. An illustration of the modified unit-selection algorithm. In order to have synthetic speech free of undesirable preglottalization, segments containing preglottalization are penalised during unit selection.

a diphone including preglottalization is selected, the unit-selection algorithm was tuned to prefer a diphone candidate free of preglottalization phenomena. Penalisation 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 there is no need to know the boundaries of preglottalization in the source recordings or in the synthetic speech. The only extra information when compared to the standard speech synthesis system is the knowledge of the presence/absence of preglottalization in the source diphone candidates, which can be automatically obtained as described in Section II. Similarly as in Section IV-A, in order to avoid errors caused by an imperfect automatic detection of preglottalization, manual detection from expert phoneticians was used throughout this experiment.

The comparison between both approaches to speech synthesis is given further in Section V. The advantage of the approach described in this subsection is clear—location of preglottalization is not needed; thus, automatic segmentation of these sounds 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 emphasise the need for special handling of preglottalization in Czech speech synthesis, approximately 965k unique sentences were synthesised using the original version of our TTS system, and 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. It means that every third sentence synthesised by our TTS system contains preglottalization.

For the evaluation of the impact of preglottalization on synthetic speech quality, 18 representative sentences, in which different distinct preglottalization sounds occurred, were chosen for further analysis. Each of these sentences was then synthesised with the three versions of our speech synthesis system—the original system (ORG), in which preglottalization was not handled, and two versions in which

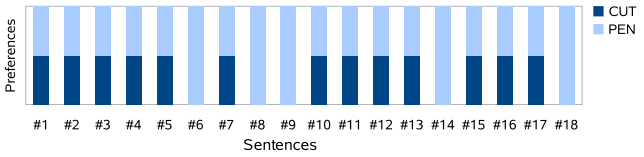

Fig. 5. Results of preference listening test for the comparison of the quality of synthetic speech produced by PEN and CUT speech synthesis scenarios

preglottalization was cut out (CUT, see Section IV-A) or penalised during unit selection (PEN, see Section IV-B). The example of different synthetic speech of one sentence is given in Figure 6. The resulting synthetic sentences were analysed by the two phoneticians in the author team, and the intrusiveness stemming from the potential presence of preglottalization was marked either as *slightly intrusive* or as *very intrusive*.

As can be seen in Table II, 11 of 18 sentences contained intrusive preglottalization (5 of them were perceived as very intrusive) when synthesised with the original speech synthesis system (ORG). It can be also seen that both proposed methods succeeded in the suppression of intrusiveness of preglottalization phenomenon—only 2 sentences still contained audible preglottalization after preglottalization sounds had been cut out of the corresponding synthetic speech signal (CUT), and no preglottalization at all was audible after preglottalization had been penalised during unit selection (PEN). The persisting presence of preglottalization after cutting them out could be caused by an imperfect segmentation of these sounds in source utterances (though it was performed by experts in this experiment), or preglottalization may be also present in the context of the cut-out sounds.

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

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

As can be seen in Figure 6, the penalisation of preglottalization causes different diphone segments are selected; thus, CUT and PEN versions sound differently. Although PEN outperforms CUT in the suppression of intrusiveness of preglottalization, the forced usage of different diphone candidates could change the *overall quality* of resulted speech. Therefore, another informal listening test was carried out to compare the overall quality of synthetic speech produced by both PEN and CUT synthesis scenarios. The results in Figure 5 show that PEN outperforms CUT also with respect to the overall quality, as it was never assessed worse than CUT.

VI. Conclusion

The next step towards linguistic naturalness of synthetic Czech was presented in this paper. First, the automatic detection and segmentation of preglottalization, the most intrusive parasitic phenomenon which degrades the linguistic naturalness of Czech speech, were briefly presented. Then,

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

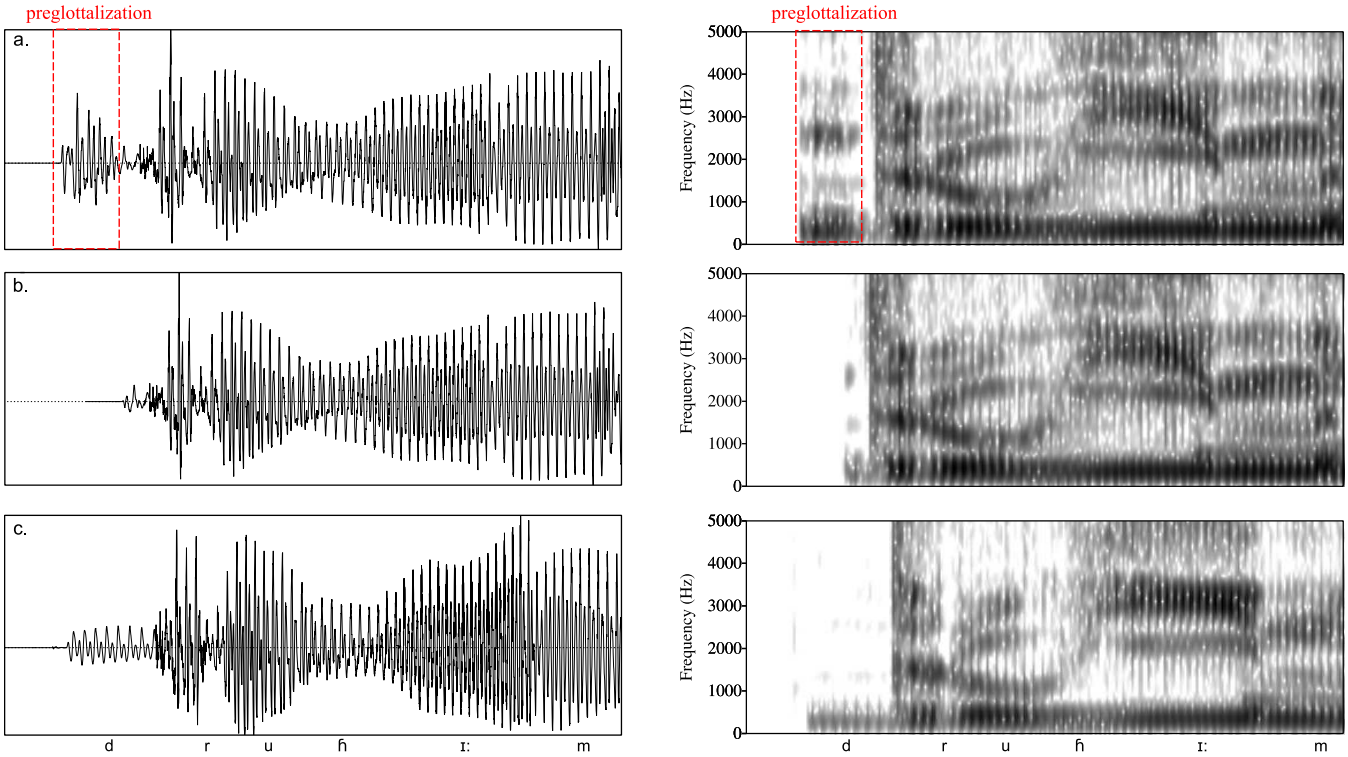

Fig. 6. Examples of synthetic speech: a. from the original system with a preglottalization sound included (ORG); b. with the preglottalization sound cut out (CUT); c. with preglottalization penalised during unit selection (PEN)

two speech synthesis scenarios were proposed and employed to synthesise speech without preglottalization. Indeed, the resulting speech was evaluated as more natural (from the linguistic point of view) than speech produced by the original system. Comparing the two synthesis scenarios, penalisation of preglottalization during unit selection seems to be a better choice, at least for two reasons—firstly, it outperformed cutting preglottalization out of synthetic speech both in the ability to suppress the intrusiveness and in the overall quality of the resulting synthetic speech; secondly, no delimitation of preglottalization sounds is needed. On the other hand, some fine-tuning of the unit-selection algorithm may be needed to find an optimal trade-off between the ability to suppress preglottalization and the overall quality of synthetic speech.

Our future work will be directed both towards experiments with other speakers and towards a utilisation of the knowledge acquired in this research in a real Czech TTS system. An influence of the automatic procedures for both the detection and segmentation of preglottalization on the quality of synthetic speech will be also investigated.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] ——, "Míra rušivosti parazitních zvuků v řeči mediálních mluvčích," *Naše Řeč*, 2011, (in Czech; in print).

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] D. Tihelka and J. Matoušek, "Unit selection and its relation to sym-

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

bolic prosody: a new approach," in *Proc. INTERSPEECH*, Pittsburgh, USA, 2006, pp. 2042–2045.

[20] 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.

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)