INTERSPEECH 2009 BRIGHTON

Identification and Automatic Detection of Parasitic Speech Sounds

Jindřich Matoušek[1], *Radek Skarnitzl*[2], *Pavel Machač*[2], *Jan Trmal*[1]

[1]Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Rep.
[2]Institute of Phonetics, Faculty of Arts & Philosophy, Charles University in Prague, Czech Rep.
{jmatouse, jtrmal}@kky.zcu.cz, {radek.skarnitzl, pavel.machac}@ff.cuni.cz

10.21437/Interspeech.2009-265

Abstract

This paper presents initial experiments with the identification and automatic detection of parasitic sounds in speech signals. The main goal of this study is to identify such sounds in the source recordings for unit-selection-based speech synthesis systems and thus to avoid their unintended usage in synthesised speech. The first part of the paper describes the phonetic analysis and identification of parasitic phenomena in recordings of two Czech speakers. In the second part, experiments with the automatic detection of parasitic sounds using HMM-based and BVM classifiers are presented. The results are encouraging, especially those for glottalization phenomena.

Index Terms: parasitic speech sound, linguistic naturalness, speech synthesis, unit selection, HMM, BVM

1. Introduction

Contemporary standard of speech synthesis of Czech has reached the stage when the synthesised speech is comfortably intelligible. The area where work still remains to be done is the naturalness of the resulting speech. Our present system applies corpus-based concatenative synthesis, using *unit selection*. In this area of speech synthesis, we can distinguish two aspects of naturalness.

On the one hand, we may consider *technical naturalness*, which is related to the way the selected units are concatenated. In this study, however, we are looking at the other aspect, which we may call *linguistic naturalness*. Unit selection synthesis increases the naturalness of speech from the technical point of view, since it attempts to select the largest suitable segment of natural speech and thus to introduce the lowest possible number of potential discontinuities into the connected speech signal. On the other hand, it allows the speaker's possible idiosyncratic habits to interfere with the linguistic naturalness of speech — more so than is the case in classical diphone synthesis. Although much has been done in the research of "technical" aspects of unit selection speech synthesis paradigm (e.g., tuning of both target and join costs, speech unit inventory creation, automatic phonetic segmentation, etc. — see e.g. [1] for more detail), linguistic naturalness of speech synthesised by a unit selection system has not been studied extensively.

The objective of this study is to take a step towards the linguistic naturalness of synthetic Czech by identifying and, subsequently, automatically detecting in the speech of the source speakers linguistically *non-systematic*, or *parasitic phenomena*. In neutral, unmarked synthetic speech — which is our ultimate goal at this stage — the presence of such phenomena may lead to an intrusive effect on listeners, for example in the form of hypercorrectness, carelessness, affectedness etc. Even worse, 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 will be most likely perceived as a discontinuity in synthetic speech. Having information about the presence/absence of a parasitic sound in a given context, we can avoid mixing such speech contexts in unit selection speech synthesis — the parasitic sound could be cut out of the speech signal or the particular speech unit containing the sound could be penalised during the unit selection mechanism, or, even, such a unit could be intentionally used in speech synthesis in order to increase the naturalness of synthetic speech.

This paper is primarily focused on the phonetic analysis and identification of parasitic sounds. The technical aspects of the automatic detection of these sounds are not detailed. The main goal was to find out whether the utilisation of automatic classifiers for this task is possible at all.

This research was supported by the grants GAČR 102/09/0989, MŠMT LC536 and VZ MSM 0021620825. The access to the META-Centrum clusters provided under the research intent MSM6383917201 is highly appreciated.

Table 1: *Description of the speech material in reference and test data (used in the context of the automatic detection techniques explained in Section 3): number of utterances, amount of data in minutes, length in phones and occurrences of the most frequent parasitic phenomena.*

	male			female		
	all	ref.	test	all	ref.	test
utterances	119	70	49	88	58	30
amount	13.75	8.84	4.91	15.04	11.34	3.70
phone len.	9,850	6,298	3,552	9,979	8,010	1,969
preglot.	123	73	50	74	53	21
postglot.	45	16	29	4	0	4
ep. schwa	71	39	32	173	151	22

2. Identification of parasitic speech sounds

The first step consisted in the analysis of the source speakers' recordings, aiming to identify fine phonetic detail whose presence may negatively affect the naturalness and unmarked character of speech. We searched for any detail in the speech signal which cannot be regarded as part of the canonical sound patterns of Czech.

The material for analysis comprised randomly selected recordings of the two source speakers (one female, one male) used in Czech speech synthesis system ARTIC [2], in total approx. 28 minutes of read speech (see Table 1 for more detailed description). Both speakers have background in radio broadcast-

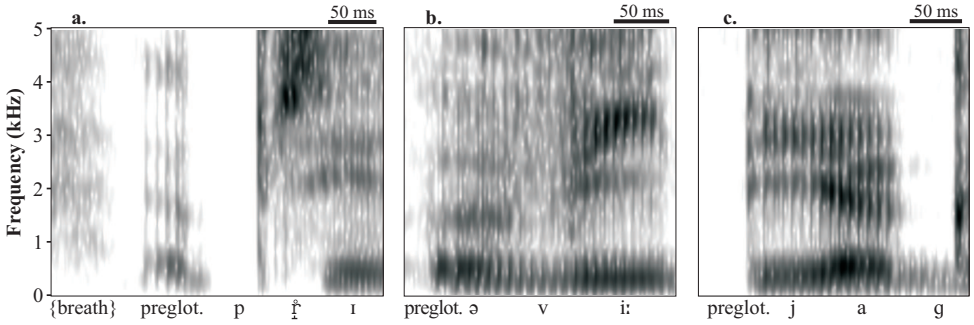

Figure 1: *Examples of preglottalization: a. creaky phonation before a voiceless plosive; b. reinforcement by schwa; c. transient.*

ing. The recordings were analysed and annotated in the phonetic analysis software Praat [3] by two human experts.

The results of the phonetic analyses are summarised in the lower part of Table 1. The presence of glottalization (preglottalization and postglottalization) and the insertion of epenthetic schwa turned out to be the most frequent non-standard phenomena in the analysed sample. It should be pointed out that, while frequent in our two speakers, these phenomena appear to be some sort of "media mannerism" and not part of everyday speech. Other, considerably less frequent features include excessive aspiration, nasalization, and breathiness.

2.1. Glottalization

For the purposes of this study, we regard glottalization essentially as a short, aperiodic noise produced by the vocal folds. As such, it occurs as one of the phonetic realisations of the phonological category glottal stop [4]. In Czech, the glottal stop is used naturally only before a post-pausal vowel. The orthoepic norm requires the usage of the glottal stop after a non-syllabic preposition (e.g., *k oknu* [k ʔoknu], to the window) and recommends it in some other prevocalic positions. The occurrence of glottal stops in preconsonantal positions, or *preglottalization*, is not usual in Czech, and it may be viewed as marked, and potentially intrusive.

As shown in Table 1, preglottalization is very frequent in both our source speakers. It can be realised in various ways (see Figure 1 for examples): as a brief transient (plosion) which can be both voiced and voiceless, as several irregular periods of creaky phonation, in very few cases also with nasalization. Moreover, both the transient and creaky phonation can be accompanied, or reinforced by a schwa-like vocalic element.

In our sample, preglottalization was found almost exclusively in the position after a pause. In only four cases, preglottalization occurred after a prosodic boundary but with no pause. The segmental context of glottalization occurrences varied to a surprisingly great extent. We expected to find it predominantly before voiced speechsounds (both sonorants and voiced obstruents). However, 13 % of preglottalization items in the male speaker's recording and 30 % in the female speaker's recording were produced before a voiceless obstruent. The two speakers also differ in the extent to which preglottalization is reinforced by the following schwa sound: while the male speaker pronounced around 10 % with schwa, it was nearly 70 % in the female speaker.

Table 1 also mentions 49 instances of *postglottalization*. This label describes aperiodic activity of the vocal folds before a pause. In 37 cases, postglottalization occurred at the end of a pre-pausal vowel. In eight items, it marked the end of a sonorant consonant. In the remaining four items, it occurred in the form of a glottal squeak [5], a sudden shift to relatively high F0, at the end of pre-pausal voiceless obstruents.

2.2. Epenthetic schwa

The Czech vocalic system does not have the vowel *schwa*. Due to articulatory reduction, however, some vowels (mostly /ɛ/ and /o/) may be realised as [ə] in everyday speech. Moreover, a schwa-like vocalic element appears regularly in certain consonantal combinations, for instance between two voiced plosives (e.g., *odkdy* [ʔodᵊgᵊdɪ], since when). For the purpose of this study, we were interested especially in positions where the occurrence of [ə] is not usual in Czech and where it may be considered hypercorrect, and potentially intrusive (as in mannerist pronunciation). The introduction of epenthetic schwa will most probably have an intrusive effect on listeners if it leads to the perceptual impression of an additional syllable.

Epenthetic schwa was considerably more frequent in the female speaker. Indeed, this can be considered as a mannerist, idiosyncratic feature of her pronunciation. It should be emphasised, however, that not all the items enumerated in Table 1 will have an intrusive effect on listeners. Let us therefore look at the cases with epenthetic [ə] in more detail.

One context when epenthetic schwa is relatively frequent has already been mentioned: it serves as a vocalic element in words like *kdy, kde, když* (when, where, if), but also *dva* [dᵊva], *dvacet* [dᵊva͡tsɛt] (two, twenty).

Second, [ə] can function as some sort of fortification, or reinforcement of a pre-pausal speechsound. This happens in 23 items of the male speaker recordings and in 16 items of the female speaker recordings, with the sonorant consonants [m, n, l] being the most likely candidates for this fortification.

Finally, epenthetic schwa appears rather frequently on the word boundary, reinforcing the final speechsound of the first of the two words. As long as this does not lead to the impression of an additional syllable, it will most likely not have an intrusive effect when the two speechsounds have identical place of articulation, e.g., *tam máš* [tamᵊ ˈmaːʃ] (there you have) or *tam půjdeš* [tamᵊ ˈpuːjdɛʃ] (there you will go). In this context, it essentially serves as an alternative to no audible release ([tamˈˈmaːʃ]). Similarly, it appears with non-syllabic prepositions in Czech, e.g., *z vody* [ˈzᵊvodɪ] (from the water).

These, let us say standard contexts can be found in the male speaker's recordings. However, one can find many instances of word-boundary epenthetic schwa in the female speaker's recordings in other contexts where it can be intrusive: *účast v delegaci* [ˈuːt͡ʃastə ˈvdɛlɛga͡tsɪ] (participation in the delegation).

The intrusive effect may become even stronger when the introduction of epenthetic [ə] induces wrong assimilation of voicing. Essentially, Czech has regressive assimilation of voicing

(see, e.g., [6] for more detail), so that isolated *zákaz* (ban) is [ˈzaːkas], while *zákaz zastavení* (no stopping) should be [ˈzaːkaz ˈzastavɛɲɪː]. As the sounds on the word boundary are identical, one could say [ˈzaːkazᵊ ˈzastavɛɲɪː]. The possibility [ˈzaːkasə ˈzastavɛɲɪː], as pronounced by the female speaker, is not allowed.

3. Automatic detection of parasitic sounds

For our experiments with the automatic detection of parasitic sounds the most frequent ones were chosen, i.e. glottalization and epenthetic schwa. The speech material was divided into reference (training) and test data as shown in Table 1. Two different kinds of classifiers were used: an HMM-based classifier and BVM classifier. The evaluation of the automatic classification was performed in the "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 $ACC = [P \cdot TPR + N \cdot (1 - FPR)]/(P+N)$, 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. In order to take also the classification "accuracy" occurred by chance into account, Cohen's kappa κ will be also indicated (in our case, $\kappa = 1$ means perfect performance of a classifier, $\kappa \leq 0$ indicates worse performance than that obtained by random classification — generally, $\kappa \geq 0.70$ is considered satisfactory).

3.1. HMM-based classifier

The most successful approaches to automatic phonetic segmentation (APS) are based on *hidden Markov models* (HMMs). In this approach, each phone or sound is modelled by an HMM: firstly the parameters of each HMM are estimated and then *force-alignment* based on Viterbi decoding is performed to find the best alignment between the HMMs and the corresponding speech data. For more details, see e.g. [2, 7].

In our experiments, a set of single-speaker three-state left-to-right context-independent multiple-mixture HMMs corresponding to all Czech phones and the parasitic sounds was employed. For models parameters estimation we employed isolated-unit training utilising Baum-Welch algorithm with model boundaries fixed to the hand-labelled ones (the reference data). The detection of a parasitic sound was then performed in a similar way as in the APS task — for each utterance from the test data (described by feature vectors of mel frequency cepstral coefficients extracted each 4 ms), the trained HMMs of all phones and parasitic sounds were concatenated according to the phonetic transcripts of the utterance (in our case, there are multiple phonetic transcripts per utterance with all the combinations of the presence/absence of the given parasitic sound in the defined contexts) and the transcript which "best matches" the data is chosen as the maximum likelihood estimation (MLE) of the utterance. In this way, the parasitic sounds in given contexts could be detected.

It is obvious that, using this type of a classifier, the accuracy of the detection of parasitic sounds crucially depends also on other sounds, and in fact on the performance of each HMM. On the other hand, as boundaries between HMMs are produced during the alignment, the position of each parasitic sound in the utterance could be located.

3.2. BVM classifier

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 [8]. The reason why we have chosen a kernel based classifier is that it often outperforms the other types of classifiers [9]. We used RBF (radial basis function) kernel in the BVM classifier.

In our experiments, the TRAPS parameterization 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 [10]. To ensure better granularity, the parameterization 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.

Table 2: *Detection of parasitic sounds.*

		male		female	
		HMM	BVM	HMM	BVM
Preglottalization	P	50	50	21	21
	N	56	59	28	29
	TPR	0.92	0.92	0.81	0.52
	FPR	0.11	0.02	0.07	0.00
	ACC	0.91	0.95	0.88	0.80
	chance level	0.50	0.51	0.52	0.54
	κ	0.81	0.91	0.75	0.56
Postglottalization		male		female	
		HMM	BVM	HMM	BVM
	P	26	26	4	4
	N	106	132	60	64
	TPR	0.77	0.96	0.0	0.75
	FPR	0.02	0.00	0.00	0.00
	ACC	0.94	0.99	0.94	0.98
	chance level	0.70	0.73	0.94	0.90
	κ	0.70	0.98	0.00	0.85
Glottalization		male		female	
		HMM	BVM	HMM	BVM
	P	73	76	25	25
	N	168	121	91	96
	TPR	0.95	0.95	0.76	0.52
	FPR	0.07	0.00	0.04	0.00
	ACC	0.93	0.98	0.91	0.90
	chance level	0.56	0.53	0.67	0.73
	κ	0.85	0.96	0.74	0.63
Epenthetic schwa		male		female	
		HMM	BVM	HMM	BVM
	P	17	–	9	–
	N	36	–	21	–
	TPR	0.29	–	0.89	–
	FPR	0.11	–	0.24	–
	ACC	0.70	–	0.80	–
	chance level	0.62	–	0.53	–
	κ	0.21	–	0.58	–

3.3. Experiments

As preglottalization occurs mostly before post-pausal consonants, only these contexts were taken into account. In the first experiment, classifiers were trained on the contexts where preglottalization really occurred (P) and where it was possible to occur but did not (N). The second experiment was targeted at postglottalization phenomena. Taking relatively small number of occurrences of postglottalization in the source data and the similarity of both post- and preglottalization sound into account, the classifiers were trained both on postglottalization and preglottalization examples. Using such a joint sound "model", only postglottalization phenomena were then detected in the appropriate contexts (i.e. in pre-pausal contexts). In the third experiment, preglottalization and postglottalization phenomena were merged into glottalization phenomena. Hence, they were trained and detected as one phenomenon. The advantage of merging is that there is more data available for training of the classifiers. The kind of glottalization can be distinguished after the classification according to the context glottalization appears in. Results of all experiments are shown in Table 2 (slightly different numbers of examples for the same data are caused by different preprocessing of the data for a particular classifier).

Epenthetic schwa can occur in various more vaguely defined contexts, but most frequently occurs in a) pre-pausal sonorant consonants, b) with two speechsounds of the same place of articulation, and c) after a non-syllabic preposition. That is why only these contexts were taken into account in our initial experiments with the detection of schwa. As HMM-based classifier is less sensitive to the context definition (BVM classifier could suffer from the training set being heavily biased towards the negative examples), only the HMM-based classifier has been employed at this moment. The results are shown in Table 2.

4. Discussion

The results of the automatic detection of glottalization phenomena in Table 2 are very good, especially for the male speaker. This is confirmed by high values of Cohen's kappa coefficient, too. The detection of glottalization has been poorer for the female speaker than for the male speaker. This may have been caused by the different distribution of "canonical", aperiodic items of glottalization and items which were reinforced with a periodic element. While the male speaker, who produced more glottalization items, pronounced a majority of them with "pure" glottalization, the female speaker reinforced most of the time. Therefore, given the limited amount and greater variability of data, the glottalization model for the female speaker is more difficult to be trained.

The detection of epenthetic schwa has been poorer than that of the glottalization phenomena. This is probably caused by the fact that, while glottalization was detected only in the adjacency of pauses, the context in which schwa occurs is significantly more variable. Moreover, the acoustic contrast between schwa and some of the possible co-occurring sounds is very low, especially in the case of sonorants.

Comparing both classifiers, BVM seem to outperform HMMs when enough representative data is available. Unlike HMMs, the training set for BVM classifiers is biased significantly towards the negative samples. This is the reason for $FPR \rightarrow 0$ in all experiments. Even though the parameters determined by the cross-validation signalised near 100% performance, the performance of the classifier dropped considerably when run on the test data set. This can have two reasons: first, the classification methodology may have changed throughout the process of labelling. The second reason is related to the large TRAPS vector size (330 features per vector), that captures parasitic sounds in their contexts. Since the training and testing sets were selected randomly, it is possible that some of the ill-classified parasitic sounds occur in contexts that were not present in the training data. On the other hand, as HMM-based classifiers are trained only on the positive examples and HMMs of the surrounding speechsounds also exist, they seem to be less sensitive to the amount of the training data and its context richness.

5. Conclusions

The first steps towards the linguistic naturalness of synthetic Czech were presented in this paper. Firstly, linguistically non-systematic phenomena, which occurred in source recordings used for synthesis of Czech speech, were identified. Secondly, two classifiers were designed and employed to automatically detect parasitic sounds corresponding to these phenomena in speech signals. Having such sounds detected and located, we can avoid to use them unintentionally during speech synthesis. Results of the automatic detection are encouraging, especially for glottalization phenomena.

In our future research, we will try to refine the criteria for classifying nonstandard speech phenomena with the aim to determine, with the help of listening tests, which of them really have an intrusive effect. Our work will be also directed to ensuring more representative data and to the optimisation of classifiers' parameters (e.g., number of frames used and their placement around the context in which the parasitic sound could occur, other kernel functions, or other classifiers).

6. 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] 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 Artificial Intelligence. Berlin, Heidelberg: Springer, 2006, vol. 4188, pp. 439–446.

[3] P. Boersma and D. Weenink, "Praat - doing phonetics by computer," www.praat.org, 2008.

[4] R. Skarnitzl, "Acoustic categories of nonmodal phonation in the context of the Czech conjunction "a"," in *AUC Philologica 1/2004, Phonetica Pragensia X*, Z. Palková and J. Veroňková, Eds. Prague: Karolinum, 2008.

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

[6] J. Volín and R. Skarnitzl, "Fonologická výjimečnost české znělé labiodentály," in *Kapitoly z fonetiky a fonologie slovanských jazyků*, Z. Palková and J. Janoušková, Eds. Prague: FF UK, 2006.

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

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

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

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