Automatic Segmentation of Parasitic Sounds in Speech Corpora for TTS Synthesis*

Jindřich Matoušek

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz

Abstract. In this paper, automatic segmentation of parasitic speech sounds in speech corpora for text-to-speech (TTS) synthesis is presented. The automatic segmentation is, beside the automatic detection of the presence of such sounds in speech corpora, an important step in the precise localisation of parasitic sounds in speech corpora. The main goal of this study is to find out whether the segmentation of these sounds is accurate enough to enable cutting the sounds out of synthetic speech or explicit modelling of these sounds during synthesis. HMM-based classifier was employed to detect the parasitic sounds and to find the boundaries between these sounds and the surrounding phones simultaneously. The results show that the automatic segmentation of parasitic sounds is comparable to the segmentation of other phones, which indicates that the cutting out or the explicit usage of parasitic sounds should be possible.

Keywords: parasitic speech sound, speech synthesis, unit selection, HMM, automatic phonetic segmentation.

1 Introduction

Contemporary *concatenative speech synthesis* techniques based on a *unit-selection framework* employ very large speech corpora. As the principle of unit-selection-based speech synthesis is to select the largest suitable segment of natural speech in order to prevent the potential discontinuities in the connected speech signal [1], attributes like the voice identity, style of speaking, speaking habits, the quality of speaking, etc. are copied to synthetic speech. In order to produce speech as natural as possible, source utterances in the speech corpora have to be spoken naturally, i.e., among others, with a natural intonation, natural speech rhythm or common pronunciation (and, more over, depending on the resulting application possibly also in various expressive or affective states, emotions, etc.). As a result, due to affectedness, carelessness or possibly also hypercorrectness related to the natural way of speaking, the recorded utterances can include so called *parasitic speech sounds* (parasitic from the point of view of both Czech canonical pronunciation and the fluency and overall acceptability of synthetic speech), linguistically non-systematic phenomena. Such sounds, if used in text-to-speech (TTS) synthesis too often or, even, unintentionally, can negatively affect the acceptability,

* This research has been supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 369–376, 2010.

fluency (or, in other words, naturalness) of synthetic speech and can have an intrusive effect on listeners, especially when neutral, unmarked synthetic speech (which is still required in a majority of TTS applications) is about to be produced.

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 annotations of parasitic sounds are almost impossible. Thus, the parasitic sounds are hidden in the corpora and, following the principle of concatenation-based unit selection speech synthesis, they can unintentionally get into synthesised speech. 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, one can avoid mixing such speech contexts in unit selection speech 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 mechanism, or, even, such a unit could be intentionally used in speech synthesis in order to increase the naturalness of synthetic speech in some applications.

In [2], the phonetic analysis and identification of parasitic speech sounds were carried out, and a procedure for the identification and *automatic detection* of the presence of parasitic sounds in speech signals was designed. In this paper the next step in the process of the precise localisation of parasitic sounds is presented. The objective is to propose an algorithm for the *automatic segmentation* of parasitic sounds in speech signals. The paper is organised as follows. Parasitic speech sounds are briefly introduced in Section 2. In Section 3, the results of the automatic detection of the parasitic sounds are shown. Experiments with the automatic segmentation of parasitic sounds, results and their discussion are provided in Sections 4 and 5. Finally, conclusions are drawn in Section 6.

2 Parasitic Speech Sounds in Czech

For the purpose of our study, randomly selected recordings of two source speakers (one female, one male) used in the Czech TTS system ARTIC [3], in total approx. 28 minutes of read speech (see Table 1 for more detailed description) were utilised. The recordings were analysed with the aim to identify parasitic sounds — sounds whose fine phonetic detail cannot be regarded as part of the canonical sounds pattern in Czech and whose presence in synthetic speech may negatively affect the perceptibility of synthetic speech.

The results of the phonetic analyses are summarised in the lower part of Table 1. The presence of *glottalization* (preglottalization and postglottalization) turned out to be the most frequent non-standard phenomena in the analysed sample. Glottalization can be defined as a short aperiodic noise produced by the vocal folds. In Czech, one of the phonetic realisations of glottalization is *glottal stop*, which naturally occurs only before a post-pausal vowel [4]. In this context, glottalization is perceived as a natural part of pronunciation. Thus, glottal stop is a well-established unit in the phonetic

Table 1. Description of the speech material in reference and test data sets (used in the context of the automatic detection and segmentation techniques explained in Sections 3 and 4): 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 [min.]	13.75	8.84	4.91	15.04	11.34	3.70
phone length	9,850	6,298	3,552	9,979	8,010	1,969
preglottalization	123	73	50	74	53	21
postglottalization	45	16	29	4	0	4

system of Czech and is used in synthesis of Czech speech [3]. On the other hand, the occurrence of glottal stops in preconsonantal positions (almost exclusively after a pause), or *preglottalization*, is not usual in Czech, and it may be viewed as marked, and potentially intrusive. Similarly, *postglottalization*, aperiodic activity of the vocal folds before a pause (either after a consonant or a vowel), is perceived as non-standard and intrusive. For more details see [2] and [5] where phonetic analysis is described more thoroughly, and other parasitic phenomena like epenthetic schwa are discussed as well. Although, as revealed by the phonetic analysis, these sounds do occur in Czech speech, they are not included in the standard Czech phonetic inventory and they have not been coped with in the synthesis of Czech speech yet.

3 Automatic Detection of Parasitic Sounds

The aim of the automatic detection of parasitic sounds was to detect, or identify the presence of the parasitic sounds (preglottalization and postglottalization in our case) in speech signals. Two different kinds of classifiers were used: an HMM-based classifier and BVM classifier. Both types of classifiers were trained on the reference (training) data set and evaluated on the test data set specified in Table 1.

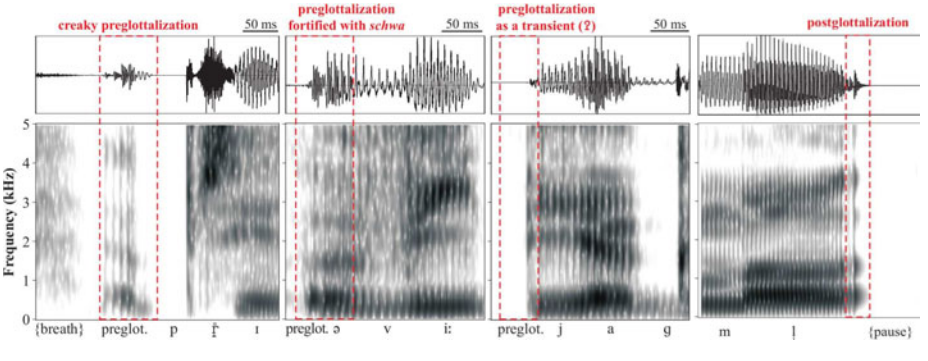

Fig. 1. Examples of parasitic sounds: preglottalization and postglottalization

Table 2. Results of the automatic detection of parasitic sounds [2]

Detection measures	Preglottalization				Postglottalization			
	male		female		male		female	
	HMM	BVM	HMM	BVM	HMM	BVM	HMM	BVM
P	50	50	21	21	26	26	4	4
N	56	59	28	29	106	132	60	64
TPR	0.92	0.92	0.81	0.52	0.77	0.96	0.0	0.75
FPR	0.11	0.02	0.07	0.00	0.02	0.00	0.00	0.00
ACC	0.91	0.95	0.88	0.80	0.94	0.99	0.94	0.98
chance level	0.50	0.51	0.52	0.54	0.70	0.73	0.94	0.90
κ	0.81	0.91	0.75	0.56	0.70	0.98	0.00	0.85

The HMM-based classifier follows the well-established techniques known from the field of automatic speech recognition (ASR) and automatic phonetic segmentation (APS), see e.g. [6,7], or for Czech [8,9]. In this framework each phone or sound is modelled by an hidden Markov model (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. As this classifier was utilised also for the automatic segmentation of parasitic sounds, it is described further in Section 4 in more detail.

Ball Vector Machines (BVM) classifier, one from the family of kernel methods, was used with RBF (radial basis function) kernel. The TRAPS parametrisation technique with the setup similar to [10] was employed to obtain the input features for the classifier. The parameters of the BVM classifier were determined using grid-search algorithm with 10-fold cross-validation. More details and reasons why this classifier was preferred over the similar ones, CVM or SVM, could be found in [2].

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 is 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). Results of the detection are summarised in Table 2.

4 Automatic Segmentation of Parasitic Sounds

Though the accuracy of the detection of the HMM-based classifier is slightly worse when compared to the BVM classifier (see Table 2), 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. Therefore, the HMM-based classifier was used for the automatic segmentation of parasitic sounds.

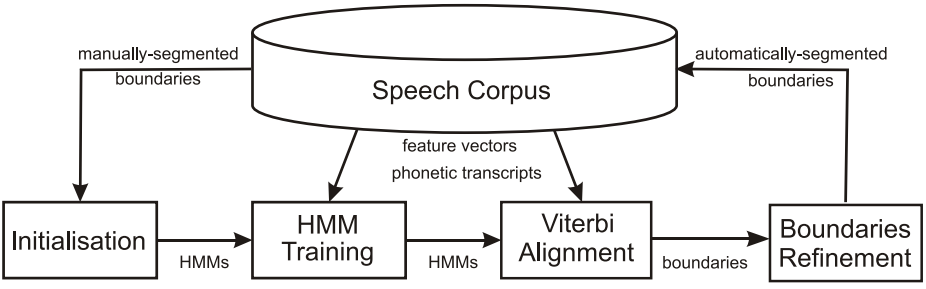

Fig. 2. Simplified scheme of HMM-based automatic phonetic segmentation

In our experiments, a set of single-speaker three-state left-to-right context-independent multiple-mixture HMMs corresponding to all Czech phones and parasitic sounds was employed in a similar way as in the automatic phonetic segmentation task in [9]. 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). 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 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 parasitic sound is identified in the stream of speech signal. Within this process, the automatic detection of the presence of each parasitic sound mentioned in Section 3 is carried out by creating multiple phonetic transcripts per utterance with all combinations of the presence/absence of the given parasitic sound 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, the parasitic sounds in given contexts could be detected. A simplified scheme of the automatic phonetic segmentation utilising the HMM-based classifier is shown in Figure 2.

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 tolerance region 10 ms (Tol10) or 20 ms (Tol20) are shown in Table 3. The results for postglottalization in the female speech corpus are not shown due to the small number of occurrences of postglottalization in female speech signals (see Table 1). 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 in pauses (see Figure 1), and, due to the smooth concatenation of speech signals in silence, the precise location of these boundaries is not so important. For comparison, the segmentation accuracy of a similar unit, glottal stop (GST-*), and the average segmentation accuracy of all other Czech phonetic units (*-*) are also shown in Table 3. The comparison of the segmentation accuracy of all boundary types for the male speech corpus is shown in Figure 3. Similar results were obtained also for the female corpus.

Table 3. Results of the automatic segmentation of preglottalization and postglottalization sounds in the male and female corpus

	male				female			
Boundary	MAE (ms)	RMSE (ms)	Tol10 (%)	Tol20 (%)	MAE (ms)	RMSE (ms)	Tol10 (%)	Tol20 (%)
PRG-*	7.50	10.56	83.33	90.48	7.13	10.20	66.67	100.00
*-POS	8.33	11.16	65.38	92.31	—			
GST-*	7.00	9.39	73.08	96.15	6.87	12.59	75.00	90.00
-	6.45	11.66	82.40	94.65	6.73	12.58	80.56	94.20

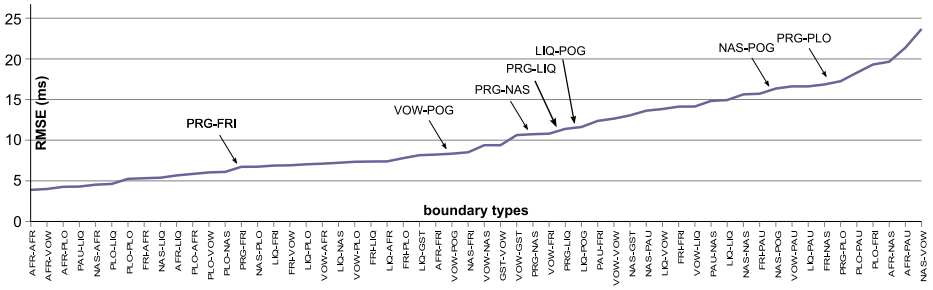

Fig. 3. Comparison of the automatic segmentation accuracy of different boundaries types (VOW = vowels, FRI = fricatives, PLO = plosives, AFR = affricates, NAS = nasals, LIQ = liquids) in terms of RMSE

5 Discussion

Despite the slightly worse results of the automatic detection of the presence of parasitic sounds, the HMM-based classifier was preferred over the BVM classifier in our experiments. It provides us with an "all-in-one" solution — during a single, well-established procedure of the automatic phonetic segmentation, the detection of the presence of parasitic sounds is carried out simultaneously. As a result, the segmentation of all phones and parasitic sounds (if detected in the appropriate contexts) is obtained. Based on these segmentations (boundaries between phones in speech signals), speech unit inventories for unit-selection-based speech synthesis can be built.

Looking at the results of the automatic segmentation in Table 3 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 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 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 (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 (statistically not significant, unpaired *t*–test, two-tailed *P*-value = 0.5879).

The results indicate that, based on the automatic segmentation, it should be possible to cut the parasitic sounds out of the speech signals and thus to prevent them from getting into synthesised speech. Or, more specifically, parasitic sounds could be used, within reasonable measure, as regular units in speech synthesis in order to increase the naturalness of synthetic speech in some applications.

6 Conclusion

In this paper, the introduction of preglottalization and postglottalization, parasitic speech sounds from the point of view of the fluency and overall acceptability of synthetic speech, was presented. Beside the automatic detection of the presence of the parasitic sounds in speech signals, the research was focused on the automatic segmentation of these sounds in speech. The main goal was to find out whether the segmentation was accurate enough for the parasitic sounds to be cut out of synthetic speech. Alternatively, the precise localisation of their positions in source speech corpora would enable explicit usage of the parasitic sounds in speech synthesis. HMM-based classifier was employed to detect the parasitic sounds and to find the boundaries between these sounds and the surrounding phones simultaneously. The results show that the automatic segmentation of parasitic sounds is comparable to the segmentation of other phones. It indicates that the cutting out or the explicit usage of parasitic sounds should be possible.

In our future work, the utilisation of both proposed classifiers, the BVM one for the detection of parasitic sounds and the HMM-based one for the segmentation of parasitic sounds in the contexts detected by the BVM classifier, will be researched. Furthermore, speech synthesis with the parasitic sounds either excluded or intentionally included will be investigated. The quality of synthetic speech will be compared to the quality of synthetic speech produced by the standard TTS system.

References

1. Tihelka, D., Romportl, J.: Exploring Automatic Similarity Measures for Unit Selection Tuning. In: Proceedings of Interspeech, Brighton, Great Britain, pp. 736–739 (2009)
2. Matoušek, J., Skarnitzl, R., Machač, P., Trmal, J.: Identification and Automatic Detection of Parasitic Speech Sounds. In: Proceedings of Interspeech, Brighton, Great Beritain, pp. 876–879 (2009)
3. Matoušek, J., Tihelka, D., Romportl, J.: Current State of Czech Text-to-Speech System ARTIC. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 439–446. Springer, Heidelberg (2006)
4. Skarnitzl, R.: Acoustic Categories of Nonmodal Phonation in the Context of the Czech Conjunction "a". In: Palková, Z., Veroňková, J. (eds.) AUC Philologica 1/2004, Phonetica Pragensia X, Karolinum, Prague (2008)
5. Machač, P., Skarnitzl, R.: Phonetic Analysis of Parasitic Speech Sounds. In: Proceedings of the 19th Czech-German Workshop on Speech Processing, Prague, Czech Rep., pp. 61–68 (2009)
6. Byrne, W., Doerman, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.: Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. IEEE Transactions on Speech and Audio Processing 4, 420–435 (2004)
7. Toledano, D., Gómez, L., Grande, L.: Automatic Phonetic Segmentation. IEEE Transactions on Speech and Audio Processing 11(6), 617–625 (2003)
8. Vaněk, J., Psutka, J.V., Zelinka, J., Pražák, A., Psutka, J.: Discriminative Training of Gender-Dependent Acoustic Models. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS (LNAI), vol. 5729, pp. 331–338. Springer, Heidelberg (2009)
9. Matoušek, J.: Automatic Pitch-Synchronous Phonetic Segmentation with Context-Independent HMMs. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS (LNAI), vol. 5729, pp. 178–185. Springer, Heidelberg (2009)
10. Schwarz, P., Matějka, P., Černocký, J.: Towards Lower Error Rates In Phoneme Recognition. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 465–472. Springer, Heidelberg (2004)