On Modelling Glottal Stop in Czech Text-to-Speech Synthesis*

Jindřich Matoušek and Jiří Kala

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz, jkala@students.zcu.cz

Abstract. This paper deals with the modelling of glottal stop for the purposes of Czech text-to-speech synthesis. Phonetic features of glottal stop are discussed here and a phonetic transcription rule for inserting glottal stop into the sequences of Czech phones is proposed. Two approaches to glottal stop modelling are introduced in the paper. The first one uses glottal stop as a stand-alone phone. The second one models glottal stop as an allophone of a vowel. Both approaches are evaluated from the point of view of both the automatic segmentation of speech and the quality of the resulting synthetic speech. Better results are obtained when glottal stop is modelled as a stand-alone phone.

1 Introduction

Nowadays, concatenative synthesis is the most widely used approach to speech synthesis. This approach employs an acoustic unit inventory (AUI) which should comprise all relevant sounds of a language to be synthesized. Although diphones or triphones are the most often used units in the current speech synthesis systems, a system designer should always start with a phonetic inventory (i.e. with either phonemes or phones) of the given language. Such a phonetic inventory then constitutes a ground for more specific diphone or triphone inventories.

In our previous work we have designed ARTIC, a modern Czech concatenative text-to-speech (TTS) system (see e.g. [1]). It employs a carefully designed triphone AUI [1, 2]. In the first versions of our system we used a *phonemic inventory* of the Czech language as a ground for the triphone inventory. Later, the fundamental phonemic inventory was extended with the significant allophonic variants of some phonemes and thus it was replaced by a *phonetic inventory*, that incorporated 47 Czech phones (see SAMPA [3] for their notations).

Although the synthetic speech produced by our system sounded highly intelligibly, certain distortions were observed in some speech contexts, especially in words starting with a vowel. In natural speech, such contexts are characterized by the presence of so-called *glottal stop*. Since glottal stop is not usually considered as a phoneme in the Czech language (and its occurrence is often not mandatory), it was not included in our baseline phonetic inventory. However, in

* Support for this work was provided by GA ASCR, project No. 1ET101470416.

order to eliminate the distortions in the synthetic speech, an explicit glottal stop modelling and synthesis seems to be necessary.

Two approaches to modelling glottal stop for the purposes of Czech TTS synthesis are introduced in this paper. The first one uses glottal stop as a *stand-alone phonetic unit*. The second one models glottal stop as an *allophone of a following vowel*, separating vowels with glottal stop from the vowels without glottal stop. In this paper, both approaches are evaluated from the point of view of both the automatic segmentation of speech and the quality of the resulting synthetic speech.

The paper is organized as follows. Section 2 briefly describes the phonetic features of glottal stop in the Czech language and discusses the automatic phonetic transcription of glottal stop. The Section 3 deals with the modelling of glottal stop in Czech synthetic speech. In Section 4 the results are presented. Finally, Section 5 concludes the paper by summarizing the main findings and outlines our future work.

2 Phonetic Features of Glottal Stop

From the point of view of Czech phoneticians [4], glottal stop (denoted as [?] in SAMPA [3]) is a glottal plosive that originates in vocal cords when a vowel or diphthong is articulated at the beginning of a word (especially after a silence, e.g. [?ahoj]) or inside a word at a morphological juncture (e.g. [na?opak]). When the articulation of a vowel in these contexts starts, vocals cords clamp tightly. Then, more audible separation of the vowel (or diphthong) from a preceding syllable is perceived. However, the point of view of Czech phoneticians is a little bit vague. Although glottal stop could distinguish the meaning of Czech words (e.g. *s uchem* [s?uxem] vs. *suchem* [suxem]), it is not common to consider it as a phoneme of the Czech language. Nevertheless, it could be considered as a stand-alone phone. On the other hand, as glottal stop could be viewed just as a beginning of a phonation, it can be defined as an allophone of the following vowel. In this paper, both conceptions are adopted and further described in Section 3.

As for the acoustic waveforms of glottal stop, it differs from context to context. In post-pausal contexts the acoustic waveform resembles the waveform of plosives. On the other hand, in intervocalic contexts it rather looks like a very transitional segment of speech. So, the contexts of glottal stop should be taken into account when modelling and synthesizing speech.

2.1 Phonetic Transcription of Glottal Stop

Obviously, there is a need of the automatic processing of the input text in text-to-speech synthesis tasks. The automatic conversion of the input text to its pronunciation (i.e. phonetic) form (so-called phonetic transcription) forms an important part of the text processing. As for the phonetic transcription of the Czech language, the phonetic transcription rules in the form of

$$A \rightarrow B \,/\, C\_D \tag{1}$$

(where letter sequence A with both left context C and right context D is transcribed as phone sequence B) were proposed to transcribe Czech texts in a fully automatic way [5].

If glottal stop is to be modelled in Czech text-to-speech synthesis, the set of phonetic transcription rules introduced in [5] should be extended with the rules describing the pronunciation of glottal stop. After a series of experiments (with respect to the observations of Czech phoneticians, e.g. in [4]) we proposed a phonetic transcription rule for inserting glottal stop into the sequence of Czech phones

$$\mathrm{VOW} \rightarrow \mathrm{?VOW} \;/\; \langle |, \mathrm{PREF}- \rangle \;\_\,, \tag{2}$$

where "VOW" stands for a vowel (or diphthong), "?" is a symbol for glottal stop, "PREF" is a prefix or a part of a compound word, and the symbol "−" marks the morphological juncture. The symbol "|" marks the word boundary. The text before the symbol "→" describes the input text to be transcribed, the phones after "→" express the result of the transcription. The symbol "_" separates the left and right context of the input text. If various contexts are allowed (denoted by "<" and ">"), individual components are separated by a comma.

3 Modelling Glottal Stop

To be able to synthesize glottal stop in concatenative speech synthesis, glottal-stop-based units must be included in an acoustic unit inventory. In our previous work, we proposed a technique for the automatic construction of the acoustic unit inventories [1, 2]. Based on a carefully designed speech corpus [6], *statistical approach* (using three-state left-to-right single-density model-clustered crossword-triphone hidden Markov models, HMMs) was employed to create AUI of the Czech language in a fully automatic way. As a part of this approach, decision-tree-based clustering of similar triphone HMMs was utilized to define the set of basic speech units (i.e. clustered triphones) used later in speech synthesis. As a result, all the speech available in the corpus was segmented into these triphones. Then, the most suitable instance of all candidates of each triphone was selected off-line and used as a representative of the unit during synthesis. In this paper the process of AUI construction is extended with the modelling and segmentation of context-dependent glottal stop units.

Two approaches to modelling glottal stop have been proposed. Both approaches describe the glottal stop sounds in the context of the surroundings units, i.e. as the triphones. Hence, there is no need to explicitly differentiate between various acoustic waveforms of glottal stop as mentioned in Section 2, because triphones implicitly catch the context of the surrounding units. In the first approach the glottal stop is considered to be an independent phone of the Czech language (see Section 3.1 for more details). In the second approach the glottal stop is modelled as an allophone of a vowel (see Section 3.2).

The same corpus as described in [6] was used for the experiments with glottal stop modelling. The corpus was designed very carefully to contain phonetically balanced sentences. It comprises 5,000 sentences (about 13 hours of speech).

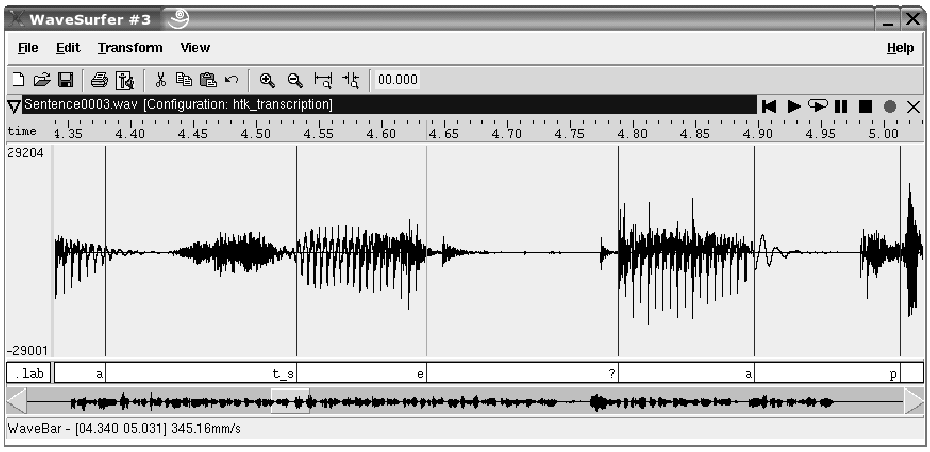

Fig. 1. An example of the delimitation of glottal stop represented as a stand-alone phone [?] in Czech synthetic speech.

Each sentence is described by linguistic and signal representations of speech. As for linguistics, both orthographic and phonetic transcriptions of each sentence are used. Speech signals are represented by their waveforms and their spectral properties are described by vectors of Mel Frequency Cepstral Coefficients (MFCCs) calculated using 20 ms windowed speech signal with 4 ms shift. In the current system 12 MFCCs plus normalized energy together with corresponding first, second and third differential coefficients (52 coefficients in total) are used.

3.1 Approach 1: Modelling Glottal Stop as a Stand-Alone Phone

In this approach (let's denote it APP1) the most straightforward modelling of glottal stop is performed. The phonetic inventory of Czech phones is extended with a single phone [?] that describes the glottal stop. The phonetic transcription rule (2) is employed to estimate the occurrences of glottal stop in Czech synthetic speech. The example of the phonetic transcription and the delimitation of glottal stop in the synthetic speech is shown in Fig. 1. The modelling and synthesis of glottal stop is the same as the modelling and synthesis of the other units in the statistical approach described above. The impact of glottal stop modelling both on the accuracy of segmentation of speech and on the quality of the resulting synthetic speech is described in Section 4.

3.2 Approach 2: Modelling Glottal Stop as an Allophone of a Vowel

Since it is sometimes hard to delimit glottal stop in the stream of continuous speech (especially in the intervocalic contexts), the alternative approach (APP2) to glottal stop modelling was proposed. In this approach glottal stop is not considered as a single phone but as an allophone of a corresponding vowel or

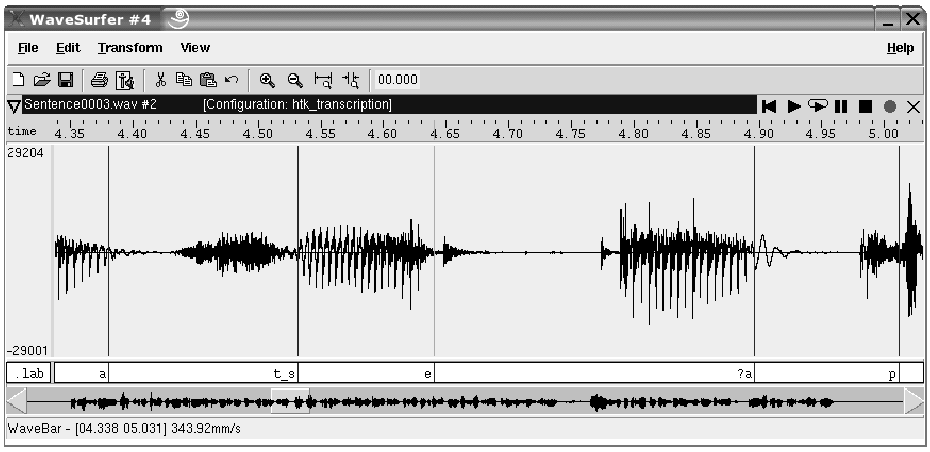

Fig. 2. An example of the delimitation of glottal stop represented as an allophone of a vowel ([a] in this case) in Czech synthetic speech.

diphthong. In fact, each vowel and diphthong is then represented by two different "phones" – vowel with glottal stop (e.g. [?a]) and vowel without glottal stop (e.g. [a]). Since there are 10 vowels and 3 diphthongs in Czech, there is a need to extend the phonetic inventory with 13 new "phones". If we use the phonetic transcription rule (2) to obtain glottal stop, a post-transcription is needed to convert glottal stops to corresponding vowel (or diphthong) units (see Fig. 2). Again, glottal stop units are then modelled and synthesized in the same way as the other units in the system. The impact of this approach on glottal stop modelling and synthesis is analyzed in Section 4.

4 Results

In this section the results of both approaches to glottal stop modelling are discussed. Since the quality of the resulting synthetic speech to a large extent depends on the quality of acoustic unit inventory and the quality of AUI is influenced by the accuracy the units are segmented with, the accuracy of the automatic segmentation of speech is evaluated in Section 4.1. The quality of the resulting synthetic speech is assessed in Section 4.2.

4.1 Segmentation Accuracy

To evaluate the segmentation accuracy of both approaches to glottal stop modelling described in Section 3, statistics of the deviation between the automatic and reference manual segmentation were computed (see Table 1). The reference segmentation consists of 50 sentences segmented by hand with no a priori information about the phone boundaries.

Table 1. The comparison of the segmentation accuracy. |MD| denotes the absolute mean deviation between the automatic and reference manual segmentation, |SD| is its standard deviation. Both values are given in ms. Acc10 and Acc20 express the segmentation accuracy in tolerance regions 10 and 20 ms.

Approach	Phone	|MD| [ms]	|SD| [ms]	Acc10 [%]	Acc20 [%]
APP1	Glottal stop	8.63	8.32	69.12	95.71
	Vowels	6.73	9.73	80.40	96.14
	All phones	6.60	9.16	82.29	95.71
APP2	Glottal stop	7.40	7.80	82.29	92.62
	Vowels	6.89	9.95	80.09	95.97
	All phones	6.70	9.28	82.20	95.37

The segmentation accuracy is also often expressed as a percentage of automatically detected boundaries which lie within a tolerance region around the human labelled boundary. The tolerance region used to be chosen somewhat arbitrarily. We chose smaller (10 ms) and bigger (20 ms) regions.

Table 1 shows the evaluation of the automatic segmentation of speech. The segmentation accuracy of both approaches to glottal stop modelling is very similar. When comparing the segmentation of glottal stop (the stand-alone phone [?] in APP1 and the allophones of vowels in APP2), APP2 demonstrates better results. On the other hand, vowels were slightly better segmented in APP1. The total segmentation accuracy (when all phones were counted in) was slightly better in APP1 as well.

4.2 Listening Tests

To evaluate the quality of the resulting synthetic speech generated using acoustic unit inventories based on both approaches to glottal stop modelling, two informal listening tests were carried out. Since glottal stop affects mainly the intelligibility of speech, the tested sentences were synthesized with neutral prosodic characteristics. 18 listeners participated in the tests.

The first test (TEST1) consisted of 18 specially designed sentences or collocations. Some contexts in these sentences could be pronounced with or without glottal stop. The presence or absence of glottal stop affects the meaning of sentences (e.g. "*Vypil asi dvě piva.*" [vipil ?asi dvje piva] and "*Vypila si dvě piva.*" [vipila si dvje piva] or "*Při boji za mír upadl*" [p\Qi boji za mi:r ?upadl=] and "*Při boji za Míru padl.*" [p\Qi boji za mi:ru padl=]). Of course, 9 sentences were synthesized using APP1 and 9 sentences were synthesized using APP2. The sentence order was chosen randomly. Some sentences were synthesized with glottal stop present and the rest of sentences were synthesized with no glottal stop present. The listeners were given a single synthetic waveform for each sentence

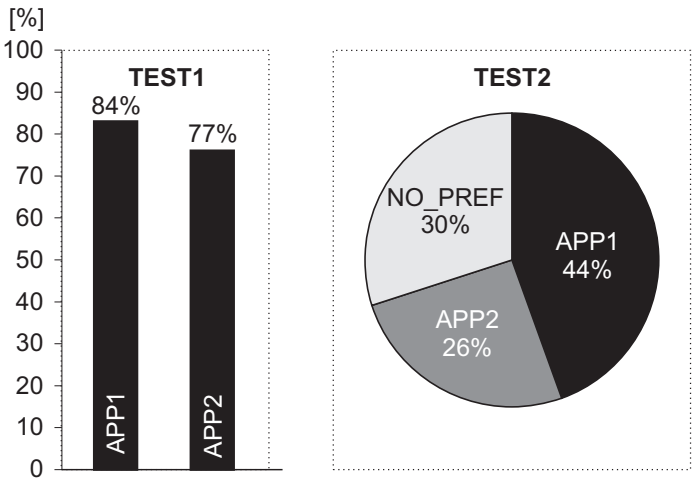

Fig. 3. The results of the listening tests.

and two textual transcriptions which differed just by the presence/absence of glottal stop. The task of listeners was to choose such a transcription which best matches the synthetic waveform.

The aim of the second test (TEST2) was to compare two synthetic waveforms (generated using both approaches to glottal stop modelling) of the same sentence directly. The same 18 sentences as in TEST1 were utilized. Now, the task of listeners was to choose such a variant which sounded more intelligibly and more "fluently" for them. Again, to ensure the independence of the test, in 9 cases the first waveform in the pair of the waveforms was synthesized using APP1 and in 9 cases the first waveform was synthesized using APP2.

The results of both tests are shown in Fig. 3. TEST1 shows the correct mappings of the written sentences to the played synthetic waveform (in percentage). In TEST2 the listeners' evaluation of both approaches to glottal stop modelling is presented (in percentage). NO_PREF denotes cases when no preference was given. It can be seen that the listeners preferred modelling glottal stop as a stand-alone phone (APP1).

To evaluate the contribution of explicit glottal stop modelling to the increase of the intelligibility of synthetic speech, another listening test was carried out. The previous version of our synthesizer with no explicit glottal stop modelling (AUI contained no glottal stop sounds) was compared to the APP1 version of glottal stop modelling. All listeners did prefer the synthetic speech with glottal stops.

5 Conclusion & Future Work

In this paper an explicit modelling of glottal stop for the purposes of Czech text-to-speech synthesis was described. Phonetic inventory of the Czech language was

extended with glottal stop units. A phonetic transcription rule for inserting glottal stop to the sequence of Czech phones was also proposed. Two approaches to glottal stop modelling were proposed and examined as well. The first approach (APP1) models glottal stop as a stand-alone phone. The second approach considers glottal stop to be an allophone of a vowel. The results presented in Section 4 showed the superiority of the first approach (mainly from the point of view of the quality of the synthetic speech assessed by the listening tests).

Moreover, when comparing both approaches to glottal stop modelling from the systemic point of view, it is more convenient to employ the first approach (APP1) in speech synthesis tasks, because the phonetic inventory of Czech phones is extended just by a single phone. Beside the worse results presented in Section 4, there are also other drawbacks of the second approach (APP2):

- There is a need of more phone-sized units in the system (13 new "phones" should be added to the phonetic inventory of the Czech language).
- The occurrence of glottal stop in some contexts (e.g. in front of [O] or [o_u]) is very rare. So, a special care should be dedicated to sentence selection when recording the speech corpus for AUI creation in order to assure that such contexts will be present in the corpus.
- Due to separated modelling of vowels and diphthongs with/without glottal stop, some rare vowels (e.g. [O]) or diphthongs (e.g. [e_u]) could not have robust models resulting in less accurate segmentation of these units and less quality representatives in AUI.

In our next work we will continuously aim at improving the quality of the synthetic speech produced by our TTS system. Beside other aspects (e.g. enhanced prosody generation or dynamic unit selection) a substantial attention will be paid to the improvements in the quality of the automatically designed acoustic unit inventories. We will focus mainly on the increase of the accuracy of the automatic segmentation of speech and on defining the optimal set of units present in the acoustic unit inventory.

References

1. Matoušek, J., Romportl, J., Tihelka, D., Tychtl, Z.: Recent Improvements on ARTIC: Czech Text-to-Speech System. Proceedings of ICSLP 2004, vol. III. Jeju Island, Korea (2004) 1933–1936.
2. Matoušek, J., Tihelka, D., Psutka, J: Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction. Proceedings of Eurospeech 2003. Geneva (2003) 301–304.
3. Czech SAMPA. http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm.
4. Palková, Z.: Phonetics and Phonology of Czech (in Czech). Karolinum, Prague (1994).
5. Psutka, J.: Communication with Computer by Speech (in Czech). Academia, Prague (1995).
6. Matoušek, J., Psutka, J., Krůta, J.: On Building Speech Corpus for Concatenation-Based Speech Synthesis. Proceedings of Eurospeech2001, vol 3. Ålborg (2001) 2047–2050.