Building of a Speech Corpus Optimised for Unit Selection TTS Synthesis

Jindřich Matoušek, Daniel Tihelka, Jan Romportl*

Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz

*Institute of Formal and Applied Linguistics
Charles University, Czech Republic
romportl@philosophia.cz

Abstract

The paper deals with the process of designing a phonetically and prosodically rich speech corpus for unit selection speech synthesis. The attention is given mainly to the recording and verification stage of the process. In order to ensure as high quality and consistency of the recordings as possible, a special recording environment consisting of a recording session management and "pluggable" chain of checking modules was designed and utilised. Other stages, namely text collection (including) both phonetically and prosodically balanced sentence selection and a careful annotation on both orthographic and phonetic level are also mentioned.

1. Introduction

It is generally known that the quality of synthetic speech produced by a corpus-based concatenative synthesis system crucially depends on the quality of its acoustic unit inventory. Several factors contribute to the quality of the acoustic unit inventory, such as speech corpus from which the units are extracted, the type of the units (i.e. phone, diphone, triphone etc.), labelling accuracy, the number of instances per each unit, prosodic richness of each unit etc.

A process of the speech corpus preparation involves several steps like text collection preprocessing, sentence selection according to specified criteria, recording by a suitable speaker and both orthographic and phonetic annotation. The present paper summarises those steps, describing the preparation of a new speech corpus for the Czech text-to-speech (TTS) system ARTIC (Matoušek et al., 2006). An extra attention is paid to the recording and verification process.

Two new Czech speech corpora have been built upon the principles described in this paper – a female corpus consisting of 5,139 utterances (approx. 10.5 hours of speech) and a male corpus consisting of 12,242 utterances (approx. 18.5 hours of speech). The new speech corpora are intended to provide enough data for robust unit selection text-to-speech synthesis as well as for prosodic-syntactic parsing and explicit prosody modelling. The special care is thus given to assuring segmental and supra-segmental balance of the recorded utterances together with the exact correspondence with their orthographic form, and also to the high-quality speech recording and verification process.

2. Text collection

There is usually an effort to cover significant linguistic events in the collected text. Traditionally, phonetic criteria are taken into account aiming to collect text sentences that follow the desired distribution of phonetic units. Our basic algorithm obeys this strategy and takes phones and diphones into account. This algorithm is further extended to be able to cope also with prosodic features of speech.

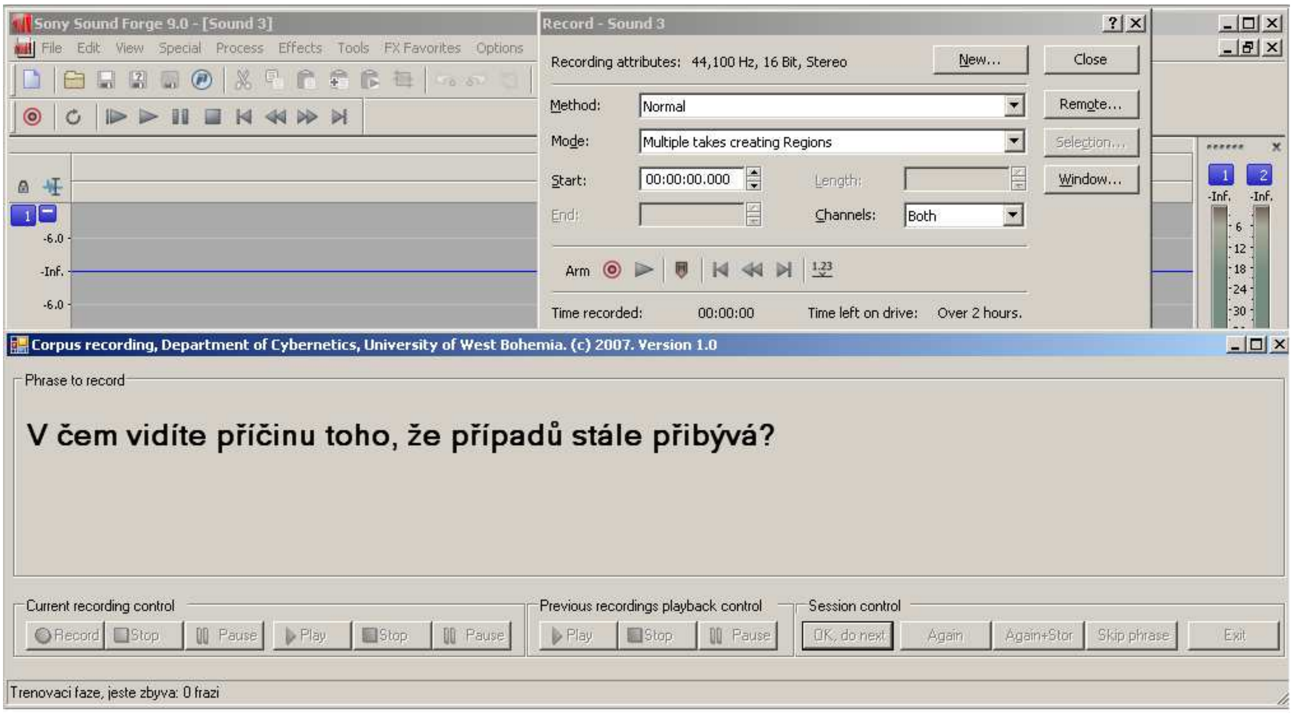
We have decided to select phonetically rich text sentences, i.e. sentences containing all phonetic events with as much uniform distribution as possible (also known as uniformly balanced sentences) – this in contrast with another possible strategy of selecting naturally balanced sentences (which contain phonetic events with respect to their frequency in natural speech). The reasons supporting our decision are discussed in (Matoušek and Romportl, 2006).

The basic sentence selection algorithm is inferred from a modified version of a greedy maximum entropy algorithm (Matoušek et al., 2001). In addition to the requirement of phonetically rich sentences, we also want all phonetic units in the list of the sentences selected so far to occur at least P-times, where P ranges from 12 to 50. Phones as the phonetic units were used for this "preselection". When this criterion is fulfilled, the algorithm continues with selection in such a way to maximise overall entropy of diphones in the selected sentences.

The extended selection algorithm incorporates prosodic features through so called *prosodemes*. Prosodemes are abstract prosodic units established in certain communication functions within the language system and in the process of sentence selection they distinguish diphones according to their involvement within these functions. For these purposes we distinguish 6 types of prosodemes: declarative, "expressive" (imperative or optative), inquiring and supplementary interrogative (all of these being terminating prosodemes), non-terminating and "null" prosodemes. Each word of a sentence belongs to a certain prosodeme according to the rules of the prosodic phrase grammar (Romportl, 2006).

This way each diphone is differentiated into 6 types according to the prosodeme it appears in. The sentence selection algorithm then works with the text data represented by this extended set of diphones. Its effort to balance the diphone occurrences thus also implicitly leads to better prosodic balancing. Although the prosodeme placement itself is carried out on the text automatically by a rule-based algorithm (obviously, it must be done before the sentences are actually uttered) and, therefore, the final utterances often prosodically differ (due to the influences of the speaker) from what was expected during the balancing process, the sentence selection algorithm using the extended diphone set still se-

This research was supported by the Ministry of Education of the Czech Republic projects No. 2C06020 and LC536, and the Grant Agency of the Czech Republic project No. 102/06/P205.



Sony Sound Forge 9.0 - [Sound 3]

Record - Sound 3

Recording attributes: 44,100 Hz, 16 Bit, Stereo

Method: Normal

Mode: Multiple takes creating Regions

Start: 00:00:00.000

Channels: Both

Time recorded: 00:00:00

Time left on drive: Over 2 hours.

Corpus recording, Department of Cybernetics, University of West Bohemia. (c) 2007. Version 1.0

Phrase to record

V čem vidíte příčinu toho, že případů stále přibývá?

Trenovaci faze, jeste zbyva: 0 frazi

Figure 1: The screenshot of the window of recording session manager (the lower half of image), with SoundForge window beneath.

lects diphone instances in as many prosodic situations (i.e. prosodemes) as possible.

3. Recording and verification process

In the next phase, the collected text material is to be recorded. For the purposes of such large corpus recording, special environment was build. It is based on SoundForge digital audio editing software (in version 9) which provides a solid platform for audio recording and post processing. Although the SoundForge in itself cannot be used for automated corpus recording (it cannot show text to record or manage the recording session), it can be extended through well defined C# interface. It allowed us to write a special recording session manager which controls the whole recording, and uses the SoundForge software for audio data recording and files handling (see screenshot in Figure 1).
The recording manager interacts with speaker through a GUI, showing the text of sentences to record, as well as buttons for individual actions, like start/stop recording/playback, record the next sentence and so on. By enabling and disabling appropriate buttons, the manager guides the speaker through recording sessions. In addition, the manager also carries out the large amount of work in supervising, to some extend the quality and consistency of data recorded, as described further.

3.1. Recording session management

It usually takes several weeks to record the whole corpus with several thousand sentences, and even when the speaker is instructed to speak in his/her the most natural and comfortable style, it is clear that non-professional speaker can unconsciously vary his/her voice or speech style, which may cause unnatural glitches when concatenating (without additional modifications) sequences of speech units varying in voice colour or tempo. To prevent this, the manager is configured to start each session by one of the following stages, in which the speaker should "tune" his/her voice:

- *warming stage* – is carried out at the very beginning of the recording when no phrases are recorded at all. The speaker is supposed to find the most natural style which is convenient for him/her. The sentences in this stage are selected randomly from the whole corpus to be recorded; we set their number to 10.
- *tune-training stage* – is carried out at the beginning of each session, except the first one. The given number of sentences (set to 6) recorded in previous sessions are played to the speaker (together with the text shown) to remind the style used in the recordings; the speaker is supposed to tune his/her mind to that style (meaning colour, pitch, tempo, loudness and other characteristics), and he/she can ask to repeat the playback of each sentence as many times as wanted. The sentences to listen are chosen randomly by the manager, to lie in interval $\langle x, y \rangle$ days from the day of recording (in our case $\langle 3, 10 \rangle$). If there were not enough phrases in the given interval, the manager would enlarge the interval, first by all $y = y + 1$ and then by $x = x - 1$, until the interval covers all recording days completed so far.
- *tune-checking stage* – always follows the *tune-training* stage in our case (although can also start a session, if required). The speaker has to record the given number of sentences (set to 8) recorded in previous sessions, and after each recording the version of sentence recorded currently is played, immediately followed by the reference version recorded earlier. The speaker has the responsibility to confirm that both versions sound

similar in the style; naturally the recording can be repeated until the speaker feels to be tuned into the style of reference recordings. The sentences for this stage are selected in the same way as in the previous *tune-training* stage, but none of the sentences used in the previous stage is chosen.

After the pre-recording stage the manager switches to the real recording stage where sentence-by-sentence are shown to the speaker, who records each separately. The manager accepts the recording or lets the sentence to record again, according to the result of checking by modules mentioned below in Section 3.2.. Moreover, to check the consistency during the current recording (if the same style is still being preserved), the manager switches randomly into *tune-checking* stage, after a given minimum number but before a given maximum number of sentences recorded (set to 80 and 120) without notifying the speaker in advance, and only after his/her recording, the recorded and reference sentence versions are played with the requirement to confirm the similarity of the style.
Unfortunately, it is currently impossible to measure the similarity of voices or speech styles automatically, therefore we must rely on the speaker, who has, in each step, the responsibility to confirm his/her conviction to the manager that the aim of the step has been met, or to ask for repetition. None of the stages or phrases can, nevertheless, be skipped or omitted.

3.2. Checking modules

The manager contains "pluggable check modules architecture" which allows to add a chain of modules checking the audio data recorded and rejecting the recording if the audio signal does not match conditions of any module in the chain. In this case, the speaker is instructed by the manager to record the sentence again, until all checks are passed. In the ideal case, also high level characteristics, like speaking style, voice colour and others, should be checked to keep the consistency of recording (it currently relies on speaker's claim, as mentioned in Section 3.1.), and the architecture is flexible enough to allow it. However, it is only little known how to measure such characteristics, and thus solely the following check modules were developed and used during our recording:

- *intensity level* – checks the absolute value of each sample and RMS (Root Mean Square) of audio in the whole channel to prevent large intensity differences in the corpus. To accept the recording, one of four decision possibilities can be chosen: either maximum sample or RMS value or their AND/OR combination lie within defined intervals. The module was used to check both speech and glottal signal, each with different setting though.
- *pauses length* – checks if each recording begins and ends with the pause of defined length. It is known that the pauses are important for HMM-based automatic segmentation, where the composed model have better chance to fit the whole phrase. The pause is measured on the interval of defined length by the same means as *intensity level* (also with the four possible decision possibilities), and the recording is rejected if the pause detected is shorter than the interval specified (set to at least 1 sec). The module was used to check pauses in speech signal only.
- *glottal signal corruption* – as glottal signal is recorded along with speech, the EGG machine used to record it is powered by batteries during the recording. However, the batteries go very low when recording session lasts long. It causes random peaks in the glottal signal which cannot then be used for pitch-mark detection, and the phrases affected would have to be recorded again (illustrated in Figure 2). The detection method is based on heuristics and observations – undamaged glottal signal does not contain sharp ascending edges, as are appearing in the damaged signal, and its energy is higher than the energy of random peaks. The positive difference of EGG signal (keeping the character of ascending edges) is multiplied by the inverted value of short-term RMS computed from the same EGG signal (emphasising energy of random peaks in parts where vocal chords are not oscillating and thus no significant energy is expected). The result is thresholded, and the overall sum of thresholded values is expected to be close to zero for undamaged signal. For more detailed description see (Grůber et al., 2007). Naturally, EGG signal was checked only.

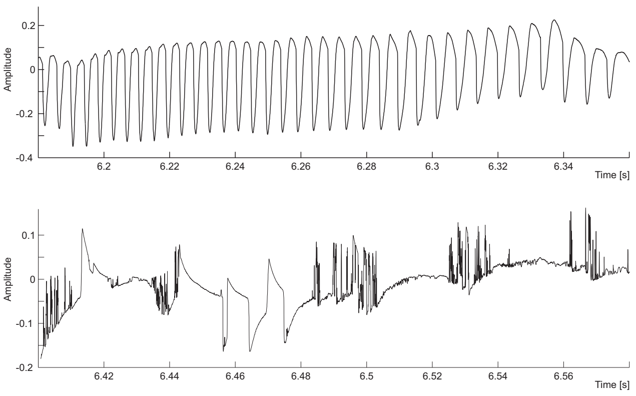


Figure 2: The illustration of correct glottal signal (upper waveform), and glottal signal recorded when the batteries of EGG machine go low.

- *low-frequency hum* – was observed in a part of our previous corpus (Grůber et al., 2007), probably caused by an improper setting of the recording system (it might be 50Hz power line hum; see example in Figure 3). To avoid it in the current recordings, the zero-crossing rate at the beginning and the end of speech signal is computed (the rate in signal without hum is much higher), and the value is compared to the average value and standard deviation obtained in advance from errorless sentences from a previous recording. Only speech signal was checked by the module.

Let us note that all the parameters and thresholds used by check modules were set heuristically by experts, and/or were based on the analysis of 50 sentences recorded by the speaker several weeks before the main recording, using the same room and equipment.

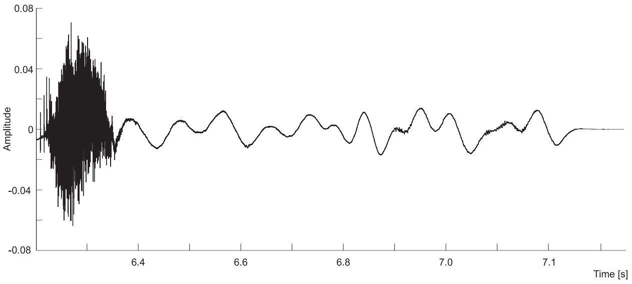


Figure 3: The illustration of low-frequency hum. The noise-like signal at the beginning is the sound of the last phone in the recorded phrase.

4. Annotation

As the very exact correspondence between the recorded speech signals and their linguistic representations is very important for corpus-based speech synthesis (where the annotation serves as a base for indexing large speech unit inventories, and any misannotation often causes glitches in the synthesised speech), a great attention was paid to the annotation process. The numbers presented in this section concerns the female speech corpus.

4.1. Orthographic annotation

The orthographic annotations were done using the annotation software Transcriber (Barras et al., 2000) by two skilled annotators in two phases. In the first phase, the first annotator (ANN1) transcribed all recordings in the way they were really pronounced following a set of annotation rules (including conventions for transcribing numbers, abbreviations, acronyms, punctuation and also rules for transcribing mispronunciations, exceptional words like non-Czech words, non-speech events like breathing or clicking, etc.). In the second phase, the initial annotations ANN1 were revised and possibly corrected by the second annotator (ANN2).

As a result, the final revised annotations comprise 62,332 running words (7.60% of them being non-speech events and 2.62% being exceptional words) in 5,139 sentences. The lexicon made from the annotations contains 17,630 different words, 0.02% of which being non-speech events and 6.11% being exceptions. When comparing both annotations, approximately 96% of all sentences and more than 99% of all words were found to be the same in both annotations. It means that second-phase annotation has corrected 237 words which would cause – if being left uncorrected – fairly noticeable problems in resulting synthesised speech because wrongly assessed segments from these words would be repetitively used in the concatenation process during unit selection. Four categories of differences between the annotations (missing the special annotation of exceptional words, different words, extra words in ANN1 and words missed in ANN1) were found and analysed. A detailed description of the orthographic annotation process can be found in (Matoušek and Romportl, 2007).

4.2. Phonetic annotation

Unlike the orthographic annotation, the phonetic annotations were done in a fully automatic way based on both the revised annotations and the acoustic signals of the recorded sentences themselves. The initial phonetic annotations were obtained using a set of approximately 155 expert phonetic transcription rules specially designed for Czech language, see (Psutka et al., 2006) for more details. The rules are in the form of $A \rightarrow \ / C\_D$, which means that a letter sequence A with both left context C and right context D is transcribed as a phone context B. Since some alternative phonetic transcriptions of the same letter contexts are allowed in Czech, the more probable transcription was preferred in this stage.

After having the initial phonetic annotations, an automatic HMM-based speech segmentation process was started, supplementing each recorded speech signal with the estimates of boundaries between phones (Matoušek et al., 2006). In this approach, corrections of the initial phonetic annotations are also made within the alternative phonetic transcriptions based on the corresponding acoustic waveforms.

5. Conclusion

In this paper, we have summarised the whole process of creation of the new Czech speech corpus for unit selection text-to-speech synthesis together with the requirements posed on it, as well as the aims this corpus has been intended with. Three main modules (text collection, speech recording and both orthographic and phonetic annotation) have been described and their role within the whole process has been discussed. The focus has been given on the speech recording and verification process.

6. References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2000. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Martin Grůber, Milan Legát, and Daniel Tihelka. 2007. Corpus recording and checking on the recorded data. In *Proc. Young Researchers Conference on Applied Sciences*, pages 174–179, Pilsen, Czech Rep.

Jindřich Matoušek and Jan Romportl. 2006. On building phonetically and prosodically rich speech corpus for text-to-speech synthesis. In *Proc. Computational Intelligence*, pages 442–447, San Francisco, USA.

Jindřich Matoušek and Jan Romportl. 2007. Recording and annotation of speech corpus for Czech unit selection speech synthesis. *Lecture Notes on Computer Science*, pages 326–333.

Jindřich Matoušek, Josef Psutka, and Jiří Krůta. 2001. Design of speech corpus for text-to-speech synthesis. In *Proc. Interspeech*, pages 2047–2050, Alborg, Denmark.

Jindřich Matoušek, Daniel Tihelka, and Jan Romportl. 2006. Current state of Czech text-to-speech system ARTIC. *Lecture Notes on Computer Science*, 4188:439–446.

Josef Psutka, Luděk Müller, Jindřich Matoušek, and Vlasta Radová. 2006. *Talking with Computer in Czech*. Academia, Prague.

Jan Romportl. 2006. Structural data-driven prosody model for TTS synthesis. In *Proc. Speech Prosody*, pages 549–552, Dresden, Germany.