Recording and Annotation of Speech Corpus for Czech Unit Selection Speech Synthesis*

Jindřich Matoušek[1] and Jan Romportl[2]

[1] University of West Bohemia, Faculty of Applied Sciences,
Department of Cybernetics, Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz
[2] SpeechTech, s.r.o., Morseova 5, 301 00 Plzeň, Czech Republic
jan.romportl@speechtech.cz

Abstract. The paper gives a brief summarisation of preparation and recording of a phonetically and prosodically rich speech corpus for Czech unit selection text-to-speech synthesis. Special attention is paid to the process of two-phase orthographic annotations of recorded sentences with regard to their coherence.

1 Introduction

Quality of synthetic speech produced by a concatenation-based 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 orthographic annotation with its revision. This paper summarises such steps in the preparation of a new speech corpus for the Czech TTS system ARTIC [1] and specially pays attention to the orthographic annotation and its revision.

The new speech corpus is intended to provide enough data (approx. five thousand sentences) for robust unit selection text-to-speech synthesis as well as for prosodic-syntactic parsing and explicit prosody modelling. Special care is thus given to assuring segmental and supra-segmental balance of recorded sentences together with exact correspondence with their orthographic form.

2 Selection of Sentences

The sentences have been selected from a large collection of Czech newspaper texts covering various domains like news, sport, culture, economy, etc. Further discussion on suitability of such kind of texts is provided in [2] together with a detailed description of sentence preprocessing.

* Support for this work was provided by the Ministry of Education of the Czech Republic, project No. 2C06020, and the EU 6th Framework Programme IST-034434.

V. Matoušek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 326–333, 2007.

From the total number of 524,472 sentences a selection algorithm has automatically chosen approximately five thousand sentences so as the resulting selection was phonetically and prosodically (or segmentally and supra-segmentally) balanced. The question has arisen, whether the sentences should be balanced *naturally* or *uniformly* – as [2] shows, we have decided for the latter, i.e. frequency of all segmental units (diphones) should be uniform (obviously, this cannot be fulfilled but at least it ensures rare units to appear as frequently as possible). In addition to this, further restrictions on the selection process were imposed: the sentences shorter than 3 and longer than 30 words were excluded; 3,500 sentences had to be declaratory, 900 interrogative ("yes/no") questions, 310 supplementary ("wh-") questions and 311 exclamatory or imperative sentences. Moreover, other sentences have been selected "by hand" (due to requirements for specific contexts) so that the final total number was 5,139.

Both segmental and supra-segmental balancing processes were carried out by a greedy algorithm maximising diphone entropy of selected sentences. Prosodic richness was introduced into this process by creating six variants of each diphone based on its position within a prosodic structure of a given sentence [3]. The selection algorithm then treated these variants as different units and thus maximised entropy also among them, which basically lead to better deployment of various prosodic contexts in the selected sentences [2].

3 Recording of the Corpus

Unlike speech recognition tasks, where some kind of noise depending on the environment the speech recogniser will run in is almost always desirable [4], high-quality noise-free recordings are required for concatenative speech synthesis. Hence, our corpus was recorded in a soundproof studio. An AKG C 3000B large-diaphragm cardioid condenser microphone with a pop filter installed to reduce the force of air puffs emerging from bilabial plosives and other strongly released stops was used. A high fidelity capture card capable of up to 96 kHz AD conversion was utilised. For our purposes, 48 kHz AD conversion has been actually performed. Glottal signal measured by an electroglottograph device was recorded along with the speech signal. The glottal signal is suitable for the detection of glottal closure instants (also called pitch-marks) which are used for accurate pitch contours estimation, pitch-synchronous speech synthesis, very precise voiced/unvoiced signal detection, or smooth concatenation of speech segments in unit selection speech synthesis [1,5].

A female voice talent possessing a pleasant voice, good voice quality and professional recording experience was chosen to record the corpus. The recording ran in a sentence-by-sentence manner. The speaker was instructed to read each sentence naturally but with no emotions and no amount of expressiveness. She was also asked to speak clearly and to keep her normal speaking rate and volume. Being aware of the importance to keep the recordings consistent both in phonetic and prosodic (within the framework of symbolic prosody description [5]) terms, an expert in acoustic phonetics and orthoepy supervised the recordings; his job was to check the consistency of recordings and also the constancy of speaker's voice quality and pronunciation. The average duration of a recording session was about 4 hours which resulted in about 13 recording sessions.

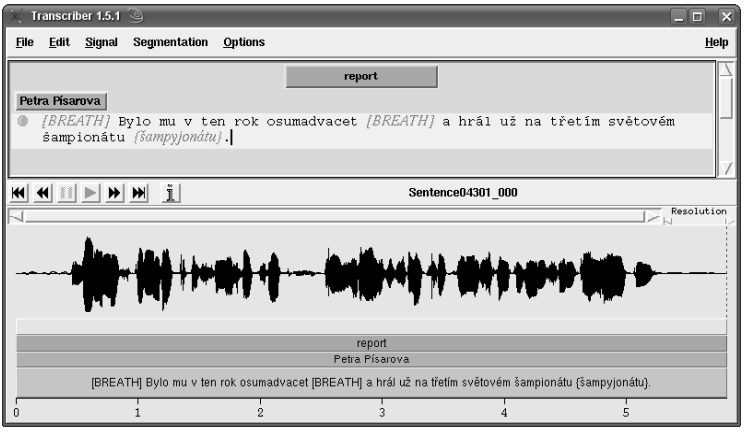

Fig. 1. A typical window of the `Transcriber`

4 Annotation of the Corpus

To know the correspondence between the speech signals and their linguistic representations on orthographic (and later on phonetic) level, the orthographic annotation of each recorded sentence is necessary. As the very precise annotation 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), the annotation process was divided into two phases. In the first phase the recordings were transcribed by a skilled annotator and the "initial" annotation (ANN1) was obtained in this way. In the second phase the annotation ANN1 was revised and possibly corrected by another annotator ("revised" annotation ANN2).

The annotations were done using the special annotation software `Transcriber`, a tool for assisting the creation of speech corpora [6] (see Figure 1). It makes it possible to manually segment, label and transcribe speech signals for later use in automatic speech processing. `Transcriber` is freely available from Linguistic Data Consortium (LDC) web site `http://www.ldc.upenn.edu/`.

4.1 Annotation Rules

During the annotation process, each sentence is transcribed in the way it was really pronounced. Unlike the "prescribed" sentences selected by the sentence selection algorithm in Section 2 (denoted as "patterns" and marked as ANN0 hereafter), the really uttered sentences can contain mispronunciation, unintelligible pronunciation, missing or extra words, various non-speech events like breathing or clicking, and very rarely also some kinds of noises (in our case mostly caused by a failure of the recording system). The rules used for the annotation of the corpus were adopted from [4,7] (where they were utilised for the purposes of speech recognition) and are listed here (if specified further in the examples, P means pattern sentences (i.e. what should have been read) and R means what was actually read):

Table 1. List of non-speech events and their brief description

Event	Description
BREATH	audible breath
CLICK	extraordinary mouth click
NO-SILENCE	no leading or trailing silence
UNINTELLIGIBLE	unintelligible pronunciation
NOISE	noise
NO-SPEECH	no speech present in the signal

1. For each sentence, the annotator is given an orthographic pattern of the sentence, i.e. what the speaker was expected to read. These are the sentences selected in Section 2.
   E.g.: `V Tatrách začíná univerziáda.`
2. Non-speech events and noises are indicated by a descriptor enclosed in square brackets. The descriptors contain only capitalised alphabetic characters and dashes. The list of the descriptors used during the annotation is given in Table 1. The descriptor is placed at the point at which the non-speech event occurred. Two non-speech events caused by the speaker are distinguished: `[BREATH]`, indicating an "audible" breath, and `[CLICK]`, marking a loud extraordinary mouth click.
   E.g.: `[BREATH] Využívání služby je jednoduché a pohodlné.`
3. Each recording should start and end with a silence. If not, a special "non-speech-event-like" mark `[NO-SILENCE]` must be put on the appropriate place (either at the beginning or the end of the sentence).
   E.g.: `Další důležitou změnou je věk [NO-SILENCE].`
4. The conventions in Czech written texts are abided by (including punctuation) – each sentence starts with a capitalised word, all other words except for proper names (e.g. Josef) and acronyms (e.g. NATO) are transcribed with low-case letters.
   E.g.: `V Kosovu jsou vojáci NATO, kteří střeží bezpečnost.`
5. Everything uttered is transcribed as words, including numbers or dates. Again, rules for correct writing of Czech numbers are followed.
   E.g. (P): `Skončil až někde na` *`163.`* `místě.`
   E.g. (R): `Skončil až někde na` *`sto šedesátém třetím`* `místě.`
6. If any word was pronounced differently from the given pattern (either as another meaningful word or a non-sense word) and the mispronunciation was clear and intelligible, the original word must be replaced with the really uttered word (the non-sense word must be enclosed by * to indicate that it is not a typo).
   E.g. (P): `V minulých dvou` *`dnech`* `si` *`zdříml`* `jen příležitostně.`
   E.g. (R): `V minulých dvou` *`letech`* `si` *`*zdříml*`* `jen příležitostně.`
7. If any word was pronounced differently from the given pattern and the mispronunciation was not intelligible (e.g. stammering, hesitation or surprise when uttering a word, a word corrupted by the recording system failure, etc.), non-speech event `[UNINTELLIGIBLE]` is placed in front of the word.
   E.g.: `V ruce žmoulá [UNINTELLIGIBLE] kapesník.`

8. The pronunciation of some words (especially the foreign ones) does differ from its written form and does not obey the Czech pronunciation rules [8]. Such words must be followed by a "commentary" notation, containing the "pronunciation form" of the word, and will be referred as exceptions henceforth.
   E.g.: `V Tatrách začíná univerziáda {unyverzyjáda}.`
9. Abbreviations are transcribed as they were spelled, also using the "commentary" notation, e.g. IBM can be transcribed as "aj bí em", "i b@ m@", or "í bé em", where "@" stands for a reduced vowel (schwa).
   E.g.: `V Kosovu jsou jednotky KFOR {káfor} z mise OSN {ó es en}.`
10. Although the recordings were made in a soundproof studio with a high-quality recording system, one must always take some possible noises into account. Indeed, sometimes (very rarely) there was a failure of the recording system causing some portions of the recordings to sound like a buzzing. Such portions of the signal (either silence, a single word, or a sequence of words) must be denoted by a special descriptor `[NOISE]` (in the case of a sequence of words, the beginning of the noise event is denoted by `[NOISE>]` and the end by `[<NOISE]`, or by descriptor `[NO-SPEECH]` if the speech signal was completely missing.
    E.g.: `To je také důvod, proč [NOISE>] píšu [<NOISE].`

Since there were relatively many recordings with `[NO-SILENCE]`, `[NO-SPEECH]` or `[NOISE]` events (around 5 %), sentences with these events were recorded once more.

4.2 The 1st Phase of Annotation

For the 1st-phase annotation, the "prescribed" sentences (ANN0) selected by the algorithm briefly described in Section 2 were used as patterns. Following the annotation rules described in Section 4.1, the annotation ANN0 was modified by the first annotator. As a result, ANN1 annotation was obtained. Approximately 72% of all transcribed sentences and 96% of all words were identical in ANN0 and ANN1 (as there were no non-speech events available in ANN0, they were not counted in during the comparison). As the exceptional words were not marked as exceptions in ANN0 (see Section 4.1, rule No. 8), most of the differences were the exceptions themselves. The results of the comparison are shown in Table 2 in section ANN0-ANN1. The results after suppressing the influence of exceptions (by supplementing ANN0 with "pronunciation forms" of the exceptions from a dictionary of exceptions – note that not all exceptions were actually present in the dictionary and that some exceptions were mistyped in ANN1 because the relative occurrences of different words increased) are shown in section ANN0-ANN1*.

4.3 The 2nd Phase of Annotation – Revision

Being aware of the importance of the precise annotation of the source speech data for corpus-based speech synthesis, all annotations were subject of a revision. The revision ANN2 was made by another annotator – she used ANN1 annotations and corrected them if needed. Approximately 96% of all sentences and more than 99% of all words

Table 2. Comparison of the pattern (ANN0), initial (ANN1) and revised (ANN2) annotations (relative occurrences in percents) and percentage of words and sentences which were equal in the comparisons. Sections with $*$ denote that ANN0 was supplemented with "pronunciation forms" of the exceptions.

Differences	ANN0-ANN1	ANN0-ANN1*	ANN0-ANN2	ANN0-ANN2*	ANN1-ANN2
Missing exceptions	2.60	0.13	2.77	0.16	0.17
Different words	0.79	0.83	0.88	0.94	0.09
Extra words	0.07	0.07	0.08	0.08	0.08
Missing words	0.05	0.05	0.05	0.05	0.03
Words OK	96.49	98.92	96.22	98.77	99.62
Sentences OK	72.32	87.13	70.83	87.15	96.24

Table 3. Differences between words and their examples as annotated in ANN1 and revised in ANN2. Percentage is shown within all differences.

Typo	Perch. [%]	ANN1	ANN2
TYPO1	47.37	blondýna	blondýnka
LAST	19.30	jak	jako
LENGTH	14.04	benzinů	benzínů
TYPO2	10.53	jak	jako
MISP	8.77	Jankulovski	Jarkulovski

were found to be the same in both annotations. The comparison of both annotations is given in Table 2 in section ANN1-ANN2 and comprises also non-speech events.

Both the words missed in ANN1 ("missing words") and deleted in ANN2 ("extra words") were mostly non-speech events (about 73%, or 82% respectively). The rest were mostly monosyllabic words. The differences between words in both annotations ("different words") are summarised in Table 3. They typically consist in:

- the last letter of a word was missing or extra (LAST);
- a vowel letter was shortened or lengthened (LENGTH);
- a word was mistyped in ANN1 as another meaningful word (TYPO1);
- a word was mistyped in ANN1 as a non-sense word (TYPO2);
- a word had been pronounced as a non-sense word but the original transcription from ANN0 was left in ANN1 (MISP).

As for the exceptional words not marked as exceptions in the initial annotations (ANN1), six types of exceptions were observed and are analysed in Table 4:

- words containing "s" were pronounced with [z] (S-Z);
- consonant [j] was inserted between [i] and a vowel (INS-J);
- "d", "t", "n" were pronounced as non-palatal consonants [d], [t], [n] when followed by "i" (typical for foreign words in Czech, DTN) [8];

Table 4. Missing exceptions and their examples. Percentage is shown within all differences. PRON stands for the "pronunciation form" of a word.

Exceptions	Perch. [%]	Word	PRON
S-Z	28.18	Klausem	Klauzem
INS-J	23.64	policie	policije
DTN	20.00	politika	polityka
LEN	12.73	Rudolfinum	Rudolfínum
OTHER	9.09	pokeru	pokru
DBL	6.36	Gross	Gros

- words containing a short vowel were pronounced with a corresponding long vowel (LEN);
- words containing a double letter were pronounced with the corresponding single consonant (DBL);
- the other words (OTHER).

The final revised annotations comprise 62,332 running words (7.60% of them being non-speech events and 2.62% being exceptions as defined by rule No. 8 in Section 4.1) 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.

5 Conclusion

We have briefly 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. The emphasis was actually mainly placed on the important step of the orthographic annotations carried out as a two-phase process, where the importance of the second annotation phase (i.e. revision) has been discussed.

The Table 2 clearly shows the improvement of the annotation coherence and its correspondence with the speech data after the annotation revision. Although the difference between the numbers of correct words in both annotation steps (ANN1 and ANN2) may first seem rather insignificant (from the point of view of ANN2 there were 99.62% words correctly annotated in ANN1), but concerning the total number of word tokens from the corpus (62,332), the 2nd-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 (since generally the whole corpus is used at once).

As can be further seen, careful classification and annotation of non-speech events and speaker's mistakes is of great importance too. The column ANN0-ANN2* from the

Table 2 can thus be regarded as a very rough "measure" of how the speaker was correct and precise in recording. Indeed this value comprises possible errors in the source text (ANN0) corrected by the speaker herself and perhaps also mistakes that have been made (or unnoticed) both in ANN1 and ANN2, but the core is definitely constituted by the differences in what the speaker was supposed to read and what has actually read.

References

1. 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)
2. Matoušek, J., Romportl, J.: On Building Phonetically and Prosodically Rich Speech Corpus for Text-to-Speech Synthesis. In: Proc. Computational Intelligence. San Francisco, U.S.A, pp. 442–447 (2006)
3. Romportl, J.: Structural Data-driven Prosody Model for TTS Synthesis. In: Proc. Speech Prosody. Dresden, Germany, pp. 549–552 (2006)
4. Radová, V., Psutka, J.: Recording and Annotation of the Czech Speech Corpus. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2000. LNCS (LNAI), vol. 1902, pp. 319–323. Springer, Heidelberg (2000)
5. Tihelka, D., Matoušek, J.: Unit Selection and its Relation to Symbolic Prosody: a New Approach. In: Proc. Interspeech. Pittsburgh, U.S.A., pp. 2042–2045 (2006)
6. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication 33, 1–2 (2000)
7. Psutka, J., Radová, V., Müller, L., Matoušek, J., Ircing, P., Graff, D.: Large Broadcast News and Read Speech Corpora of Spoken Czech. In: Proc. Eurospeech. Ålborg, Denmark, pp. 2067–2070 (2001)
8. Psutka, J., Müller, L., Matoušek, J., Radová, V.: Talking with Computer in Czech. Academia, Prague (in Czech) (2006)