Design of Speech Corpus for Text-to-Speech Synthesis

Jindřich Matoušek, Josef Psutka, Jiří Krůta

Department of Cybernetics
University of West Bohemia in Pilsen, Czech Republic
jmatouse@kky.zcu.cz, psutka@kky.zcu.cz, ihik@students.zcu.cz

Abstract

This paper deals with the design of a speech corpus for a concatenation-based text-to-speech (TTS) synthesis. Several aspects of the design process are discussed here. We propose a sentence selection algorithm to choose sentences (from a large text corpus) which will be read and stored in a speech corpus. The selected sentences should include all possible triphones in a sufficient number of occurrences. Some notes on recording the speech are also discussed to ensure a quality speech corpus. As some popular speech synthesis techniques require knowing the moments of principal excitation of vocal tract during the speech, pitch-mark detection is also a subject of our attention. Several automatic pitch-mark detection methods are discussed here and a comparison test is performed to find out the best method.

1. Introduction

In our previous work, we have designed ARTIC, a new Czech TTS system based on segment concatenation [1]. Generally, the synthetic speech quality of a concatenation-based synthesis system crucially depends on the quality of a speech unit database. Several factors contribute to the quality of a speech unit database, such as speech corpus from which the units are extracted, the type of unit (i.e. diphone, triphone etc.), labeling method (manual or automatic), number of instances (segments) per each unit, prosodic richness of each unit etc. This paper proposes a way to prepare and record a speech corpus for the use in concatenation-based TTS synthesis applications. Some speech synthesis techniques (e.g. PSOLA or some methods of harmonic/stochastic synthesis) also require knowing the moments of principal excitation of vocal tract (usually the moments of glottal closure – so called pitch-marks) during the speech. Therefore the exact detection of these time instants is very important for these techniques. So speech corpus should contain information about the proper placement of pitch-marks if such techniques are intended to be used to generate speech. Phonetic representation of speech should also be included in the speech corpus, since it represents the pronunciation form of recorded sentences. Czech phonetic transcription process is typically done by rules and is described e.g. in [2]. The scheme of a speech corpus design process is shown in Fig. 1. The paper deals with the following tasks: selection of sentences, recording of sentences and detection of pitch-marks.

The paper is organized as follows. Section 2 describes a method for selection of sentences to record from a large text corpus. In Section 3 recording conditions are proposed to record a quality speech corpus. Section 4 is dedicated to pitch-mark detection methods. Finally, Section 5 contains the conclusion and outlines our future work in this field.

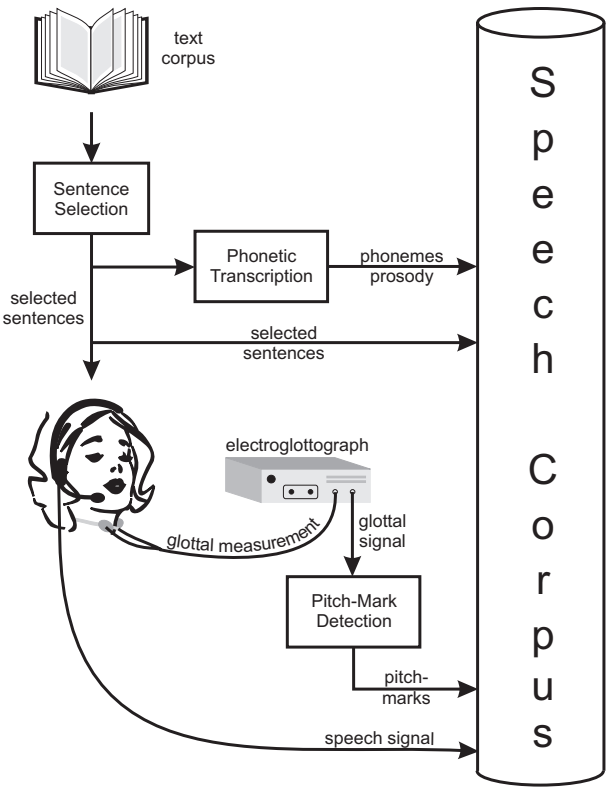

Figure 1: *The scheme of a speech corpus design process.*

2. Selection of sentences

First of all, we have to select sentences to record. In [1] no selection was performed since the speech recordings were obtained from radio broadcasting. So there was no possibility to affect the distribution of speech units in the speech corpus – the distribution was fixed by the recorded sentences. The problem is that even if we use a large natural read-speech corpus, some speech units (especially if we use triphones) do not have to be included in this corpus. Moreover if we use statistical approach to the automatic speech segment database construction [1], statistical modeling of some units is not reasonable because they appear very rarely in the speech corpus. So if we can control the process of sentence selection, it is possible to reach the desired distribution of speech units.

To select suitable sentences, a slightly modified sentence selection procedure than in [3] is used. In [3] sentences were chosen so that the distribution of triphones in these sentences reflects the distribution of triphones in a real speech. Se-

lected sentences were then recorded and used to train a speaker-independent speech recognition system [4]. In an arbitrary TTS system, where no restrictions are put to an input text, an arbitrary triphone could be synthesized. If an infrequent triphone is synthesized, some glitches could be observed in synthetic speech. Improper statistical modeling of this rare triphone can cause these glitches. To avoid this situation, all triphones should appear often enough in the selected sentences. So, the task is to select sentences from a text corpus (that consists of R sentences), so that the selected sentences contain desired number of occurrences of every triphone.

Proposed sentence selection algorithm works in j steps ($j = 0, \ldots, M$) as follows:

1. In step j, there are R_j possible sentences to choose from and $\bar{R}_j$ already selected sentences ($R_0 = R$ and $\bar{R}_0 = 0$).
2. Let us denote D_k as the desired number of occurrences of triphone k in the speech corpus (in general, D_k could be different for different triphones). Given that N_k^j is the number of occurrences of triphone k in sentences selected so far, the equation
$$L_k^j = \max\left[0, D_k - N_k^j\right] \quad (1)$$
denotes how many occurrences of given triphone k are still missing in step j. The function `max` ensures that the missing number of occurrences of triphone k is not negative (if all occurrences of a given triphone have been already selected, no ones are missing).
3. All remaining sentences i ($1 \le i \le R_j$) are evaluated:
$$r_i^j = \sum_{k=1}^{K} \min\left[L_k^j, S_k^i\right], \quad (2)$$
where r_i^j is the rating of the i-th sentence in j-th step, K is the number of triphones in i-th sentence, L_k^j is the missing number of occurrences of triphone k, and S_k^i is the number of occurrences of triphone k in i-th sentence. The rating r_i^j is computed over all different triphones in the sentence i. The function `min` ensures that the sentence rating is increased at most only by the number of occurrences that is missing. So, if there are many occurrences of triphone k in a single sentence, the redundant occurrences are not taken into account (they neither improve, nor worsen the sentence rating).
4. The sentence s^j with the highest rating r_i^j in j-th step
$$s^j = \arg \max_{1 \le i \le R_j} r_i^j \quad (3)$$
is selected and stored to the speech corpus $\bar{R}_{j+1} = \bar{R}_j + \{s^j\}$. Consequently, this sentence is removed from the list of possible sentences $R_{j+1} = R_j - \{s^j\}$ so that it could not be chosen in the next steps.

The whole process should stop when the number of occurrences of every triphone k reaches its desired value D_k. As a result a huge number of sentences could be selected. It could be impossible to record such a number of sentences. Therefore stopping number M is defined to ensure that a reasonable number of sentences is selected (see item 1). The selected sentences are then stored in $\bar{R}_M + 1$. Finally, let us note that if a sentence with a less frequent triphone is chosen, other more frequent triphones in this sentence are chosen too. In this way the distribution of triphones also partially reflects the distribution of triphones in a real speech.

3. Recording of sentences

The next step in a speech corpus building process is the recording itself. The principle of a concatenation-based speech synthesis is to concatenate speech segments from a speech segment database so that the synthetic speech mimics the voice of a speaker who recorded the speech corpus. So it is good to choose a professional speaker with a pleasant voice, good voice quality and possibly time-invariant speech quality. The speaker should be also available to record some extra sentences in case of re-recording some incorrect sentences or speech corpus enlargement.

It is important to keep some recommended conditions during recording all selected sentences, such as a closed silent recording room, using the same microphone and recording device, etc. Ideally, all sentences should be recorded at once, since the voice quality of the speaker can vary from time to time. In case of a large speech corpus it is practically impossible. But time needed to record the corpus (let's denote it as *corpus time*) should be kept as short as possible to minimize the effects of voice variations in time. On the other hand, there is a speaker-dependent time region (we call it *recording time*) in which a speaker can speak continuously while not degrading his/her voice quality. Professional speakers are able to speak continuously for more than two hours, while amateur speakers can find it difficult to speak for more than ten minutes. It is good for a speaker not to record more than his/her recording time to ensure consistent speech recording. A compromise between the requirement of minimum corpus time and maximum recording time should be found for each speaker to build a quality speaker-dependent speech corpus.

Some other requirements can be put on the speaker. The speaker is often forced to speak in a neutral way so that the prosodic features included in the speech are almost monotonous. Standard speech synthesis methods (e.g. TD-PSOLA) can be then used to modify prosodic feature in a reasonable way without audible distortions in a synthetic speech. If a natural speaking (in sense of prosodic characteristics) is demanded when recording the speech corpus, some unit selection algorithm should be employed during speech synthesis to ensure that a unit with prosodic features closed to the target (synthetic) prosody is selected. In this case a very large speech corpus is often used in which each unit should occur more often in a different prosodic context. To take prosodic features into account, sentence selection algorithm described in 2 should be generalized to include prosody as well.

4. Detection of pitch-marks

As mentioned in Section 1, some popular speech synthesis techniques (e.g. PSOLA or some kinds of harmonic/stochastic synthesis) require knowing the moments of glottal closure (so called pitch-marks) during the speech. If such techniques are planned to implement, each sentence in the speech corpus should be accompanied by the proper pitch-mark placement.

4.1. Background

In fact, there are two basic approaches to determination of the moments of principal excitation of vocal tract which differ in the input signal assumed for pitch-mark detection:

I. **Glottal-Based Methods.** To be able to use these methods, glottal signal has to be recorder along with the speech. Glottal signal represents the activity of vocal folds during speech (see Fig. 2c). To measure glottal

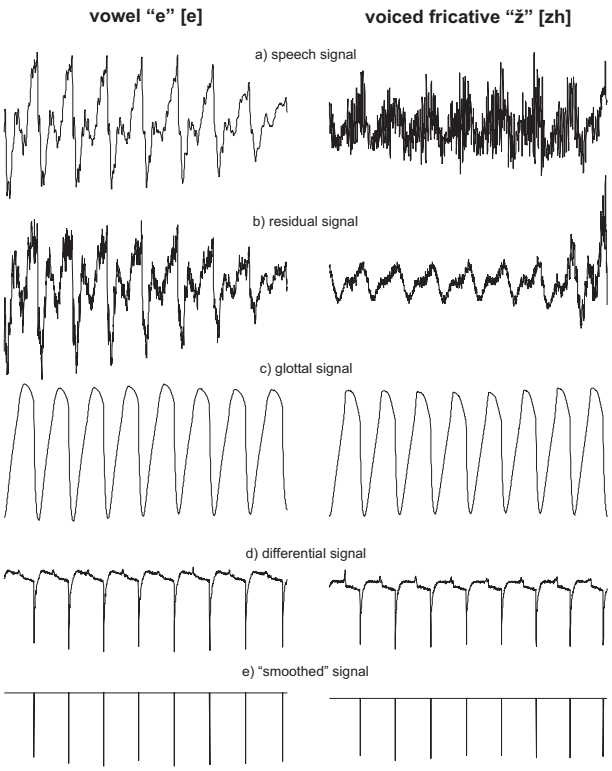

Figure 2: *Speech (a), residual (b), glottal (c), differential (d) and “smoothed” (e) signal of a Czech vowel “e” [e] and Czech voiced fricative “ž” [zh].*

signals, a device called *electroglottograph* (or *laryngograph*) is used. This device enables vocal fold contact (glottis) to be measured in a non-invasive way without affecting the quality of human speech production. An electroglottograph measures the variations in impedance between two electrodes placed across the neck (centered on the larynx) as the area of vocal fold contact changes during voicing.

II. **Residual-Based Methods.** These methods try to estimate pitch-marks from a residual signal of the speech. A residual signal is preferred to a speech signal (see Fig. 2a), since it is more suitable for automatic pitch-mark detection: residual signal usually emphasizes peaks in the signal (see Fig. 2b). These peaks correspond to pitch-marks.

Let’s remark that using glottal-based methods we can expect superior results, since glottal signals are not loaded by the modifications that happen to a flow of speech in vocal tract (compare both glottal and residual signals in Fig. 2 – especially in case of a voiced fricative the preference of glottal signal is evident). On the other hand, there is a need to measure the activity of vocal folds explicitly and this is not always possible (recording over telephone, etc.). In such special cases, pitch-marks have to be detected directly from (processed) speech, preferably from residual signals (e.g. linear prediction residual signal).

4.2. Experiments

To find out which method to use for pitch-mark detection in our speech corpus, some experiments were made. Four techniques

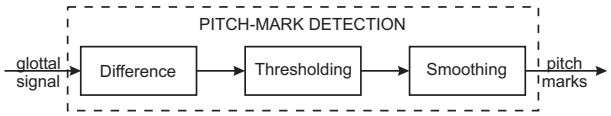

Figure 3: *The scheme of a proposed pitch-mark detection algorithm.*

were examined. Primarily, a program `Epochs` from Entropic Research laboratory was used to determine pitch-marks since the Entropic Signal Processing Package package is available in our laboratory. This program uses a dynamic-programming algorithm and a set of costs and rewards to determine the most likely set of pitch-marks in the input signal [6]. Several input signals were tested. Then, we propose an original algorithm to detect pitch-marks from a glottal signal.

Here are the four methods examined in our experiments:

1. **Residual signal + Epochs (ResE).** This method was applied in [1]. Residual signal of a speech was used as an input signal of `Epochs`.
2. **Glottal signal + Epochs (GlotE).** In this method glottal signal measured by an electroglottograph was used as an input of `Epochs`.
3. **Differential of glottal signal + Epochs (DiffE).** To emphasize the moments of glottal closure, the differential of a glottal signal (see Fig. 2d) can be taken into account, since it reflects the rate of change of the status of the vocal folds. There are usually large sharp peaks in the differential signal which correspond to the moments of glottal closure. Using the differential signal phenomena caused by improper vertical placing of the electrodes across the neck (not centered on the larynx) are also reduced. `Epochs` was used again to find out the position of pitch-marks in the differential signal.
4. **Original method (Orig).** We propose an original method to detect pitch-marks (see Fig. 3). This method uses a processed form of a glottal signal. Firstly, the differential signal is derived from the glottal signal (as in DiffE method). Then, thresholding is performed on the differential signal to remove slow changes of the status of the vocal folds – in our case all positive values and values lower than the global r.m.s. value of the differential signal are removed. The remaining values are chosen as the candidates for the pitch-mark placement. The next procedure performs pitch-mark smoothing. It examines the thresholded signal, removes pitch-mark candidates that are too close to each other and on the other hand inserts pitch-marks which were removed incorrectly in previous steps. There is no reason to use `Epochs` for pitch-mark detection in such a processed signal: after pitch-mark smoothing stage the non-zero samples that remain in the “smoothed” signal correspond to the positions of pitch-marks (see Fig. 2e).

The comparison of the methods described above is given in the next section.

4.3. Results

To be able to compare pitch-mark detection methods described in the previous section, we have to perform a manual pitch-mark identification in a test speech data. The tested speech data includes both speech and glottal signal and comprises 1511 pitch-

marks. Most of the manually determined pitch-marks (about 95%) were easy to identify. Let us call these pitch-mark as *explicit* pitch-marks. The remaining 5% of pitch-marks were not so explicit in the speech data and therefore were difficult to identify. We denote these pitch-marks as *indistinct* pitch-marks.

To compare the sequence of manually determined pitch-marks (we call it *reference* sequence S_R) and the sequence of automatically detected pitch-marks (by one of the method mentioned in the previous section – let's denote it as *test* sequence S_T), a dynamic-programming algorithm (modified Levenshtein distance of sequences of time instants) was proposed. This algorithm searches for the minimum number of transformations needed to derive the sequence of pitch-marks S_R from the sequence S_T. The transformations considered are substitution (S), deletion (D) and insertion (I). Each transformation is assigned a weight that describes how much the transformation modifies the test sequence S_T. Deletion is applied when there is an extra pitch-mark in S_T. On the other hand, insertion is employed when a pitch-mark is missing in S_T (in comparison with the reference sequence S_R). The weights of deletion and insertion were set equal to 1 in all cases. Substitution is used when a pitch-mark in S_T is replaced by a pitch-mark in S_R. If a distance between the substituted pitch-marks is lower than 10% of the local pitch period, no penalty is given (the weight is equal to 0 – in fact, no modification is needed). Otherwise, the weight of substitution is set equal to 1 as well. The threshold 10% was used because the pitch-mark position misplacement in this range does not influence the quality of the synthetic speech [5]. The accuracy of automatic pitch-mark detection is given by

$$Accuracy = \frac{N_R - N_S - N_D - N_I}{N_R} \times 100\%, \qquad (4)$$

where N_R is the number of pitch-marks in the reference sentence S_R, N_S is the number of substitutions, N_D is the number of deletions and N_I is the number of insertions involved in the comparison process.

If we use a glottal-based method for pitch-mark detection (i.e. GlotE, DiffE or Orig method), detected pitch-mark positions have to be adjusted to reflect the time lag between the glottal signal measured at the vocal folds and the speech signal measured in front of lips. The lag is a little bit different for different speakers (it depends on the distance between the two points of measurement – i.e. on speaker's vocal tract dimensions). For our test female speaker the lag was measured typically to be between $440\ \mu s$ and $560\ \mu s$. This range for pitch-mark position shifting has to be taken into account when comparing the sequence of pitch-marks detected by a glottal-based method with a manually determined sequence of pitch-marks.

The experiments were made in two steps. Firstly, all the manually determined pitch-marks (including indistinct ones) were taken into account (the number of these pitch-marks is 1511). The results are presented as the bright columns in Fig. 4. Secondly, the number of all manually determined pitch-marks was still assumed the same, but indistinct pitch-marks were ignored in the comparison process. The results are then shown in Fig. 4 (the dark columns). The results support our presumption that glottal-based methods outperform residual-based methods. The best results were obtained by the method we proposed.

5. Conclusion

In this paper we proposed a way how to design a speech corpus for the use in concatenation-based TTS synthesis applications.

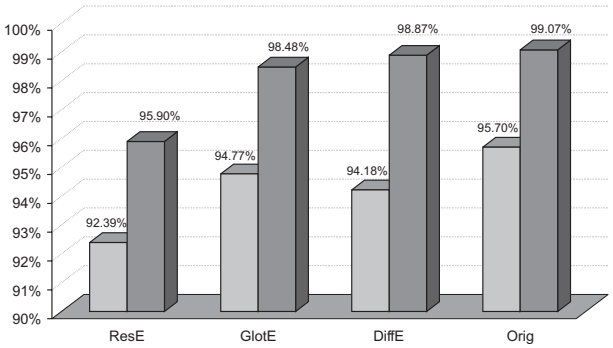

Figure 4: *The pitch-mark detection accuracy. The bright columns show the accuracy when all manually determined pitch-marks were taken into account. The dark columns show the accuracy when ignoring indistinct pitch-marks in the comparison process.*

An algorithm to choose sentences from a text corpus was suggested. The selected sentences include all possible triphones in a sufficient number of occurrences. Some remarks on speech recording were also discussed to ensure a quality speech corpus. The moments of a principal excitation of vocal tract should be a part of the speech corpus, so pitch-mark detection was examined too. Several automatic pitch-mark detection methods were discussed and a comparison test was performed to find out the best method.

In our next work we will build a speech corpus as proposed in this paper. Then, using this new corpus a new speech segment database will be created and applied in our TTS system [1]. The enhanced quality of this new speech corpus should lead to a better quality of the synthetic speech.

6. Acknowledgment

This research was supported by the Grant Agency of Czech Republic no. 102/96/K087 and the Ministry of Education of Czech Republic, project no. MSM235200004.

7. References

[1] Matoušek, J. and Psutka, J., "ARTIC: A New Czech Text-to-Speech Synthesis System Using Statistical Approach to Speech Segment Database Construction", Proceedings of ICSLP2000, vol. IV, Beijing, 2000, pp. 612–615.

[2] Psutka, J., "Communication with Computer by Speech", (in Czech), Academia, Prague, 1995.

[3] Radová, V., "UWB_S01 Corpus – A Czech Read-Speech Corpus", Proceedings of ICSLP2000, vol. IV, Beijing, 2000, pp. 732–735.

[4] Müller, L., Psutka, J., Šmídl, L., "Design of Speech Recognition Engine", Proceedings of TSD2000, Springer Verlag, Berlin, 2000, pp. 259–264.

[5] Moulines, E. and Charpentier, F. J., "Pitch-Synchronous Waveform Processing Technique for Text-to-Speech Synthesis Using Diphones", Speech Communication, vol. 9, 1990, pp. 453–467.

[6] "ESPS Programs", Version 5.1, Entropic Research Laboratory Inc, 1996.