

Slovak Speech Database for Experiments and Application Building in Unit-selection Speech Synthesis

Milan Rusko, Marian Trnka, Sachia Daržágín, and Miloš Cernák

Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia
E-mail: milan.rusko@savba.sk, trnka@savba.sk, darzagin@savba.sk,
milos.cernak@savba.sk

Abstract. After the years of hesitation the conservative Slovak telecommunication market seems to become conscious of the need of voice driven services. In the last year, all the three telecommunication operators have adopted our text to speech system Kempelen in their interactive voice response systems. The diphone concatenative synthesis has probably reached the frontier of its abilities and so the next step is to check for a synthesis method giving more intelligible and more natural synthesized speech with better prosody modelling. Therefore we have decided to build a one speaker speech database in Slovak for experiments and application building in unit-selection speech synthesis. To build such a database, we tried to exploit as much of the existing speech resources in Slovak as possible, to utilize the knowledge from previous projects and to use the existing routines developed at our department. The paper describes the structure, recording and annotation of this database as well as first experiments with unit-selection speech synthesizer.

1 Introduction

At the beginning of this project there was no annotated speech database available for unit-selection speech synthesis in Slovak, there was only a database for Czech language [1]. We decided to design a professional-quality one speaker database for research, experiments and application building in unit-selection speech synthesis. The database being built is extendible, but also downscalable. Smaller parts of the database can be used for simpler synthesizer systems (e.g. limited domain TTS). The database covers the phonetic inventory of Slovak, includes a set of sentences for prosody modeling; it contains naturally spoken spontaneous speech, application-oriented phrases, set of words with embedded diphones and basic numerals.

2 Recording

The database consists of recordings of one male, non-professional speaker, experienced in speech processing. The recording took place in an anechoic room of a

professional studio specialized to speech recording (radio commercials, dubbing, etc.). Typically, the sessions lasted about two hours and were realized in irregular intervals from one week to one month. A Neumann U 87 cardioid condenser microphone with Focusrite Trackmaster pre-amplifier and a hard disk recording system equipped with AARK 20/20+ sound board was used during the sessions. 44.1 kHz sampling frequency and 16 bit resolution were used.

3 Choice of the source text material, database content

In spite of the fact that we plan to extend the speech database in future, the initial structure of the database had to be clearly defined. Our ambition was to design a general-purpose database being at the same time suitable for making experiments in limited domain synthesis. The other contradictive requirement that the database should not be too big, but representative enough from the phonetical, phonological, and other points of view. Therefore we decided to design the database as a combination of several more or less independent parts:

- Phonetically rich sentences
- Set of words covering all Slovak diphones
- Set of sentences covering intonation phenomena
- Spontaneous speech record (General topic story, Application oriented story)
- Set of prompted application-oriented phrases and embedded application commands
- Numerals

3.1 Phonetically rich sentences

For good coverage of phonemes we have chosen the subset of 1000 phonetically rich sentences from a set of nearly 2000 sentences used at our department for the development of the SpeechDat-E Slovak fixed telephone database [2].

The coverage of all Slovak phonemes is guaranteed in this set. To cover the Slovak diphones and triphones as well as possible we have decided that the first future extension of our database will be a set of phonetically rich sentences used in our SpeechDat-SK mobile database, which was designed to cover the majority of Slovak diphones and a considerable number of triphones.

3.2 Set of words covering all Slovak diphones

For experimental purposes and to raise the number of occurrences of Slovak phonemes we decided to include our set of words (part of them being nonsense words) with embedded phonemes. This set of words was originally designed for recording of the set of the synthesis elements of our diphone synthesizer.

3.3 Set of sentences covering intonation phenomena

One of the biggest problems we had faced was the absence of technically oriented research in Slovak prosody, which would give a ground to define sophisticated rules for prosody modeling in Slovak. We asked Dr. Gabriela Mucskova from the Department of Slovak Language of the Comenius University in Bratislava to help us define a set of sentences that would reflect the important phenomena of Slovak prosody with respect to accent and intonation. At first we started building a set of sentences based on syntactical classification; that gave us a basic set of intonation contours. Soon we found out that this set has to be minimized by using a different classification scheme based on melody contour typology which was determined according to the literature [3] and to our own research mentioned above. Several sentences with different number of syllables were generated for every class. An algorithm for division of the text into syllables was designed with the help of Dr. Gabriela Mucskova.

3.4 Spontaneous speech record

General topic story. Spontaneously spoken story - e.g. a story of a film expressed by the speaker naturally in his/her own words represents a typical part of speech databases for synthesis purposes and is believed to increase the level of naturalness in synthesized speech. This is the reason why we included a ten minutes long spontaneously spoken story into the database. The speaker is telling short stories and fairy tales.

Application oriented story. As we wanted to give an opportunity to the researchers to compare the synthesis using general topic story and application-oriented story in limited domain synthesis, we made also a record of a story where the speaker describes his journey by bus, train and air-plane.

3.5 Additional prompted application-oriented phrases

In this part of the database the phrases typical for a supposed application are included (bus, train, air transport, names of stations and places), as well as typical phrases with embedded command words from telecommunication vocabulary. Command words and embedded command phrases used in SpeechDat-E Slovak fixed telephone database are included in the synthesis database.

3.6 Numerals

Combinations of numerals expressing time, money amounts, telephone numbers etc. The basic numerals were recorded in three ways to represent the melodic contours at beginning, center, and end positions of the compound phrase.

4 Annotation

The annotation consists of several levels of information. If necessary, new levels of annotation can be added. Annotation techniques and choice of annotation levels belong to the subjects of research to be accomplished on this database; therefore the above-mentioned annotation levels serve as a reference only, as an initial annotation to start with.

4.1 Annotation levels

There are two text annotation levels:

- orthographic text
- orthoepic text

Signal annotation levels are as follows:

- microsegmental information - pointers to individual pitch periods
- phoneme boundaries information
- diphone boundaries information
- syllable boundaries information
- whole words and phrases information

Suprasegmental annotation level consists of:

- melody contour information - smoothed f0 value, intonation phrase boundaries
- accent information

4.2 Automatic annotation

Orthographic to orthoepic conversion. The text in the orthographic form was transcribed to the orthoepic form by the block of pronunciation developed for earlier versions of our synthesizers [5]. The orthoepic text generated automatically was then manually checked and corrected by an expert with a degree in linguistics.

Microsegmentation, pitch marking. Microsegmentation, pitch period boundaries detection was accomplished by a rule based routine, which works well on a clean studio-quality full range speech signal [6]. Using an orthoepically transcribed text and a rule-based phoneme recognizer [7] correspondence of every microsegment to a particular phoneme can be recognized and its boundaries can be estimated.

Segmentation to diphones. One of the levels of annotation divides the speech signal into parts (elements, mainly diphones) whose inventory matches the set of the elements used in our Kempelen diphone synthesizer. The boundaries of the elements from which the signal was generated are known for the synthesized signal. Making use of the fact that we have a synthesizer with the voice of the same speaker, we applied a DTW algorithm to automatically label element (diphone) boundaries in the recorded signal.

Our automatic segmentation methods are reliable to such an extent that we could develop a PC interactive recording program which asks the user to utter several dozens of words, then automatically finds the required phonemes embedded in these words, and immediately generates a user-voice synthesis elements database [8]. The synthesized speech is fully intelligible, in spite of the fact that the recording session takes only about 10 minutes [9]. To reach high quality and more accurate annotation, after automatic labelling all the levels of annotation of our new speech database are manually checked by our department staff.

5 Experimental synthesizer

5.1 Labeling speech

In this experiment a different annotation technique was used than that mentioned in Sect. 4.2. We used Baum-Welch training to build complete ASR acoustic models from the database. This engine was then used to label the data. The whole labeling was realized in FestVox framework [12], where Carnegie Mellon University's SphinxTrain and Sphinx speech recognition system are used. We used 500 phonetically balanced utterances (see Sect. 3.1) for training and labeling. For phonetic transcription, a lexicon from SpeechDat-E [2] was used. We automatically labeled the syllable boundaries and we assigned stress to the each first syllable in a polysyllabic word (this is typical for Slovak language).

As the lexicon contained only phonetic transcription of isolated words, we corrected the pronunciation at the word boundaries by hand. The complete orthography transcription was then checked by a phonetician. Recorded prompts plus orthography and the described method of producing phone strings from that orthography were then used to create full acoustic HMM models of the recorded data. Finally, these models were used to align the labels against the recorded prompts by Sphinx2 [10] speech recognition system.

Evaluation. We have automatically aligned the boundaries of 18967 phones. The process of automatic labeling was checked by two expert labelers at our department. We can summarize their conclusions into the following list, ordered by importance:

1. The vocals were often labeled shorter than the consonants; sometimes they had just one or two periods assigned.
2. The boundaries of affricates were sometimes misplaced.

3. The label for the beginning of a phone "r" was put 1 – 2 periods later and the label for end was placed 1 – 2 periods sooner.
4. The boundaries of concatenation of two words – one of them ending and the second one beginning with a vowel – were misplaced (the boundary between vowels was not recognized correctly)
5. The labels for phones "v" and "l" were often shifted
6. The phones "m", "n" and "J" also belong to the phones that were often labeled incorrectly.

This automatic labeler for Slovak language failed in labeling speech segments at the places where human labelers often have problems too to correctly assign label positions. Based on labeled data, we further analyzed mean durations of Slovak phones. The results are depicted in Table 1. Firstly, we have done frequency analysis of the used corpus to find out whether it has the properties typical for Slovak phones; this is shown as Incidence1 in the table. We compared this with the work published in [13], where such analysis was done on 1 million of words (7.1 millions graphemes). The results of this study are shown as Incidence2. Comparing these data sets we can conclude that our distribution roughly fits the "general" distribution and so the achieved durations can also be generalized (for this particular speaker). Standard deviations are shown in the last column. We got higher standard deviations for long vowels i:, e:, u:, and diphongs i_ˆa, i_ˆe. On the other hand, consonants have shown smaller standard deviations.

5.2 Speech synthesizer built on the phonetically rich sentences

Within the consecutive step a Slovak corpus-based speech synthesizer was built using the labeled data. It is based on the principles published in [11]. The approach uses CART technique to build a classification tree for each phoneme with questions from NLP block in its nodes. We used duration model with average durations of the phones according to Table 1. Then we applied simple multiplicative factors for phrase final and phrase initial positions. As the first approximation no prosody modelling was done. The prosody obtained is a by-product of unit selection from a large speech corpus, using both contextual and linguistic features to find optimal sequence of speech segments using Viterbi algorithm.

6 Conclusion

The first annotated general purpose speech database in Slovak for experiments and application building in unit-selection speech synthesis was built. The results of previous work and projects of the Department of Speech Analysis and Synthesis of the Institute of Informatics was used for database content definition and in automatic annotation. Some of the annotation layers are still under construction, but the experiments with synthesizer building have already started.

By almost fully automatic process we have built the corpus-based speech synthesizer, which produces intelligible speech. The main drawback of this first

Table 1. Durations of phones

Phone	Incidence1 [%]	Incidence2 [%]	Mean duration [ms]	Standard deviation
o	9.34	9.71	63.46	29.7
a	8.32	8.59	75.33	26.4
e	7.81	6.94	59.23	30.7
i	5.38	5.93	65.36	38.1
t	4.53	3.66	74.63	30.6
s	4.53	5.11	111.22	24.0
r	4.32	4.98	50.17	15.9
n	3.87	3.86	54.87	20.9
v	3.51	3.65	45.99	17.9
k	3.50	3.87	90.53	31.4
m	3.44	3.26	61.39	21.2
p	3.40	3.48	81.95	28.1
a:	2.54	1.94	121.02	39.5
u	2.49	2.38	70.01	43.5
d	2.44	2.03	62.61	21.2
J	2.30	2.27	81.34	31.1
i:	2.29	2.35	93.41	48.1
l	2.14	2.40	67.89	42.7
z	2.00	1.67	93.16	17.1
L	1.66	1.78	64.05	29.5
c	1.64	1.43	91.72	45.7
b	1.58	1.59	71.57	16.8
j	1.50	1.04	77.76	39.9
h	1.40	1.20	60.43	34.4
i_ˆe	1.16	1.08	105.99	51.2
tS	1.11	1.09	113.30	26.3
ts	1.09	1.44	122.42	26.1
u:	1.03	0.87	97.83	50.1
S	1.00	1.02	105.78	28.7
i_ˆ	0.98	0.90	73.22	26.8
Z	0.96	0.71	97.34	25.7
J\	0.93	0.86	72.51	31.3
x	0.78	1.22	87.55	44.4
e:	0.78	0.82	101.68	56.6
u_ˆ	0.75	1.04	80.92	62.0
f	0.70	1.04	72.15	27.4
i_ˆa	0.62	0.66	120.91	46.7
g	0.43	0.32	82.48	32.2
w	0.23	0.08	54.40	31.7
u_ˆo	0.23	0.20	123.12	28.3
r=	0.17	0.17	72.75	23.0
N	0.15	0.13	98.10	33.7
l=	0.14	0.06	79.62	27.7
o:	0.11	0.11	104.63	35.2

version is in prosody modelling, which relies on the variability of the large speech corpus. Within the following step, it is necessary to design a detailed prosody model incorporating data driven methods. As to duration modeling, we plan to build zscores model. For intonation modeling the data set described in Sect. 3.3 will be used.

7 Announcement

This research was supported by the Slovak Agency for Science VEGA, grant No.2/2087/22.

References

1. Matoušek J., Psutka J. and Kruta J.: Design of Speech Corpus for Text-to-Speech Synthesis, Proceedings of the 7th Conference on Speech Communication and Technology EUROSPEECH 2001, vol. 3. Aalborg, Denmark, pp. 2047-2050.
2. Rusko M.: Definition of corpus, scripts and standards for Fixed Networks - Slovak (SpeechDat-E deliverable ED1.2.3), <http://www.fee.vutbr.cz/SPEECHDAT-E>
3. Král A.: Pravidlá slovenskej výslovnosti, Slovenské pedagogické nakladateľstvo Bratislava (1996), pp.163 – 200, ISBN 80-08-00305-7.
4. Rusko M., Trnka M., Daržágín S., Petriska M.: SpeechDat-E, the First Slovak professional-quality telephone speech database, In: Research Advances in Cybernetics., ELFA Publishing House, Košice, (2000) 187–211, ISBN 80-88964-61-X.
5. Daržágín S., Franěková L., Rusko M.: Conversion and Synthesis of the Slovak Speech. (in Slovak), Jazykovedný časopis, **45** (1994) 31–34.
6. Daržágín S., Král A., Rusko M.: Phoneme-oriented Approach to Speech Recognition in Slovak, in D. Mehnert Hrsg.): Elektronische Sprachsignalverarbeitung in der Rehabilitationstechnik, Berlin, (1993) 83–89, ISSN 0940-6832.
7. Daržágín S., Rusko M.: Automatic Labeling of Speech Signal for Slovak Speech Database Building, Proceedings of the 31st Int. Conf. ACOUSTICS-High Tatras, (1997) 124–125.
8. Rusko M., Daržágín S., Trnka M.: Databases for speech recognition and synthesis in Slovak, In: Proceedings of the conference SLOVKO - Slovenčina a čeština v počítačovom spracovaní, Bratislava, VEDA, (2001) 88–97, ISBN 80-224-0692-9.
9. Rusko M., Daržágín S., Trnka M.: Automatic design of the elements database for speech synthesizer in Slovak, (in Slovak), In Proceedings of the conference Noise and vibrations in practice - Kočovce 2002, Slovak Technical University Bratislava, (2002) 75–78, ISBN 80-227-1355-4.
10. Huang, X.D., et. al.: The SPHINX-II Speech Recognition System: An Overview, Computer Speech and Language (1993) 137–148.
11. Black A.W., Taylor P.: Automatically clustering similar units for unit selection in speech synthesis Proc. of the European Conference on Speech Communication and Technology Rhodes, Greece, (1997).
12. Black A.W, Lenzo K.A.: Building Synthetic Voices for FestVox 2.0 Edition, <http://festvox.org> (2003).
13. Štefánik J., Rusko M., Považanec D.: The Frequency of Words, Graphemes, Phones and Other Elements in Slovak, Jazykovedný časopis, **50** (1999) 81–93.