

DEVELOPMENT OF A REAL-TIME ASR SYSTEM FOR SLOVAK SPEECHDAT DATABASE

Milos Cernak and Marian Trnka

Institute of Informatics, Slovak Academy of Sciences
Dubravska 9, 845 47 Bratislava, Slovak Republic
Emails: Milos.Cernak@SAVBA.SK, Trnka@SAVBA.SK

Abstract: *This paper describes development of a real-time speech recognition system in Slovak for the voice-operated telephone services. The system is based on SPHINX2 platform. The decoder using Hidden Markov Models was trained on the SpeechDat-E Slovak database. It is speaker independent, large vocabulary, continuous speech real-time automatic speech recognition system. Test results are given for the test groups of isolated digits, connected isolated digits, application words, phonetically rich words, and city (plus proper) names. Achieved word error rates are in the interval from 3.73% (connected isolated digits, vocabulary of 11 words) to 15.72% (city names, vocabulary of 927 words).*

Keywords: *Real-time systems, speech recognition, speech databases.*

1. INTRODUCTION

The Department of the Speech Analysis and Synthesis of the Institute of Informatics of the Slovak Academy of Sciences has been working on the tasks concerning man-machine communication using speech interface in Slovak since 1989. The main parts of the interface that have been developed so far are:

- monophone speech synthesizer (TTS) in Slovak [1],
- speaker independent word recognizer (DTW based) [2],
- speaker independent phoneme recognizer (knowledge based) [1997, unpublished],
- diphone speech synthesizer (TTS) in Slovak.

All of the above mentioned applications were developed for full range, high-quality speech. For the experiments and development of more up-to-date recognition systems based on Hidden Markov Models a large, correctly designed and annotated database of telephone call recordings was needed. A volume of about 5000 recorded calls of different speakers per database seems to be standard in the bigger European countries. The database has to be balanced phonetically, geographically, and also balanced according to the age and sex of the speakers. In a frame of the SpeechDat-E COPERNICUS project (2000-2001) we have built a telephone database in Slovak, having 1000 recordings for fixed telephone network. [3].

Our aim was to develop a *real-time* ASR system with adequate accuracy, which could be used in the telephone services. The current system, which we describe in this paper, is based on Sphinx2 platform. Sphinx2 achieves real time speeds, and belongs to the best ASR systems available. It is written in C programming language, supporting n-grams and finite state grammars (FSG) language models. Having good experiences in using Sphinx2 for automatic annotation of our synthesis database [4] we decided to use this system also in our recognizer.

This paper describes the developed recognizer in detail. Section 2 introduces the SpeechDat database; section 3 describes training procedure of semi-continuous HMM acoustic models; section 4 describes testing procedure using five different test sets and summarizes achieved results, and finally section 5 concludes the paper with a concise discussion.

2. THE SPEECHDAT-E DATABASE FOR SLOVAK

SpeechDat-E is a set of databases following the standard defined with SpeechDat II [5]. The collection was performed automatically telephone via the ISDN connection (on the recording side). As a compromise between the need and the economical possibilities, it was decided to build a 1000 speakers database for Czech, Polish, Slovak and Hungarian and a 2500 speakers database for Russian. After the preliminary statistical research a set of the so called prompt sheets had to be generated. The prompt sheet is a list of sentences and words to be read by the caller and a set of questions to be answered. The prompt-sheets were formed according to the possible areas of the speech recognizer applications (computers, banking, shopping, marketing, traveling and tourist information, telecommunication etc.). Every of them include:

- isolated digits and its sequences,
- digit / number strings,
- natural number,
- money amounts in Slovak crowns, Dollars and Euro and their smaller units,
- yes/no questions (spontaneous answer),
- dates, prompted phrases with date, relative and general date expression,
- time and time-phrases,
- application words / key phrases,
- word spotting phrase using embedded application word,
- directory assistance names: city of birth (spontaneous), company, agency, surname, forename plus surname, own forename (spontaneous),
- spellings: artificial sequence, city name, own forename (spontaneous),
- phonetically rich words,
- phonetically rich sentences.

To reflex the real-life features the database was statistically balanced according to:

- a) Regional coverage - representation of the main phonetic groups. The repartition of speakers should be proportional to the population in regions with 5% tolerance and with minimum 5% speakers per region.
- b) Age of the callers.
- c) Sex of the callers.

It is the first large telephone speech corpus collected in Slovakia. Speechdat-E Slovak is available for the users now. It is being used in our experiments for training of several types of recognizers. The companies which are members of the SpeechDat-E Consortium have already started to develop commercial recognizers using this database. We also hoped that the database will be useful not only for the universities and academic institutions, but primarily for companies in the telecommunications and teleservices. The liberalization of the Slovak telecommunication market, hand in hand with recent boom in speech processing technology, will lead to a competition among operators and also other companies in the field of voice-driven teleservices. The created database can be the first step to the professional design of such services.

A new database intended for building the speech synthesis systems in Slovak has been recently published, as well [4].

3. TRAINING PROCEDURE

The first step in training procedure is training data preparation. This stage includes training session selection, removing files with undesired label content and feature files generation. Following files are imported from SpeechDat-E database: A-law speech files, the SAM format label files, lexicon and a list of training sessions. A-law speech sample files are necessary to convert to 16-bit linear speech samples to process with Sphinx system. The training set consisted of 32191 utterances, selected from 800 recording sessions. Rest of 200 recording sessions was later used for system evaluation.

The **SphinxTrain** tool [5] has been used for training of acoustic models. Sub-phonetic acoustic models have been trained using 5 state HMMs. The HMM topology used in this system was a strict left-to-right Bakis topology, allowing the HMMs to skip states. The total number of shared state distributions in our final set of trained HMMs (our acoustic models) was set to 6000. The ratio of the difference in likelihood between the current and the previous iteration of Baum-Welch to the total likelihood in the previous iteration was set to 0.04. We used minimal 7 iterations of the Baum-Welch training. Four feature streams has been used: 12 MFCC cepstrum coefficients, 24 differential cepstrum coefficients, 3 power coefficients, and 12 second order differential cepstrum coefficients.

The training procedure starts with training of the Context-Independent (CI) models for the sub-word units in the dictionary. Flat initialization of the CI model parameters is used. Training continues with the training of the models for Context-Dependent sub-word units (triphones) with untied states. These are called CD-untied models and are necessary for building decision trees in order to tie states. Next, decision trees for each state of each sub-word unit are built. Decision trees are then pruned, and the states are tied. Following this, the final models for the triphones in the training corpus are trained. These are called CD-tied models. We have trained also three non speech events:

- [sil] – Silence (pause).
- [fil] – Filled pause. Examples of filled pauses are uh, um, er, ah, mm.
- [spk] – Speaker noise. All kinds of sounds and noises made by the calling speaker that are not part of the prompted text, e.g. lip smack, cough, grunt, throat clear, tongue click, loud breath, laugh, loud sigh.

In a semi-continuous model, all the senones share a single *codebook* of Gaussian distributions, but each senone has its own set of *mixture weights* applied to the codebook components. We have used a codebook size of 256 vectors.

4. TESTING PROCEDURE

We have used five different test sets for evaluation of the system, selected from 200 recording sessions of the database. For each of the test set, a trigram model has been created using **CMU-Cambridge Statistical Language Modeling Toolkit** [7]. Fig. 1 shows the picture of the testing procedure. A noisedict consisted of two items: fil, and spk. The test set control file listed all

utterances to be recognized, and language models were trigram grammars in ARPA format¹. Recognized words and reference transcription (created from the SAM format label files) have been formatted into MLF (Master Label Format) files, and evaluated by **HTK HResults** tool [8]. The percentage numbers of labels correctly recognized is given by

$$Correctness = \frac{H}{N} \times 100\%,$$

where H is the number of correct labels, and N is the total number of labels. The difference $(N - H)$ is the number of deletions and substitutions. The accuracy is computed by

$$Accuracy = \frac{H - I}{N} \times 100\%,$$

where I is the number of insertions. Table 1 gives the summary of achieved results.

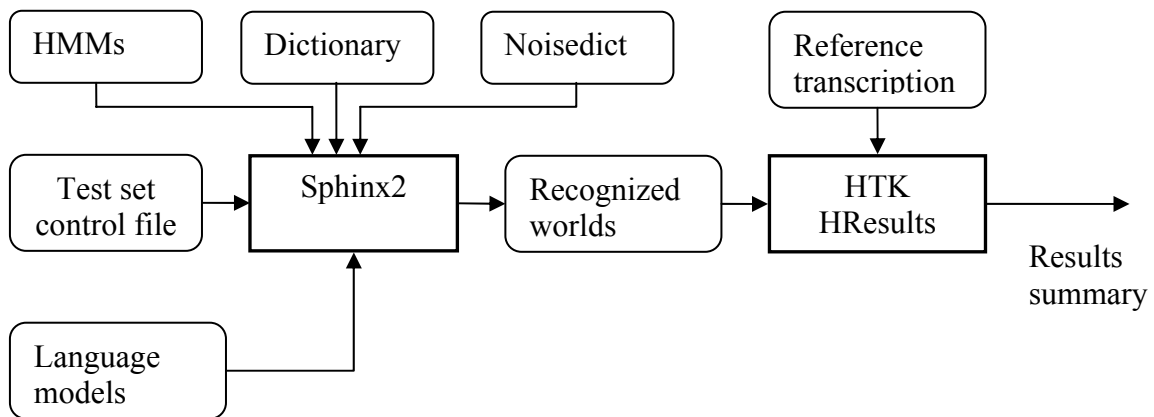


Figure 1. Testing procedure of the ASR system.

Used trigram language model primarily consisted of the following:

- *Unigrams*: The entire set of words in this LM, and their individual probabilities of occurrence in the language. The unigrams included the special *beginning-of-sentence* and *end-of-sentence* tokens: $\langle \mathbf{s} \rangle$, and $\langle / \mathbf{s} \rangle$ respectively.
- *Bigrams*: $P(\text{word}_2 | \text{word}_1)$.
- *Trigrams*: $P(\text{word}_3 | \text{word}_1, \text{word}_2)$.

During speech decoding in batch mode, *global best-path* search of the word lattice has been used. Number of codewords computed per frame was set to 4. Both of those parameters increase recognition accuracy. Table 1 shows also the perplexity of the used LM for each test set. The perplexity can be roughly interpreted as the geometric mean of the branching factor of the text when presented to the language model. It is generally true that lower perplexity correlates with better recognition performance.

Table 2 shows achieved recognition results in word error rates (in %), as compared to other SpeechDat-compliant databases, and used reference recognizer version 0.95 [9].

¹ The format is simply P(N-gram sequence) sequence BP(N-gram sequence). These (the numbers associated with unigrams and bigrams) are actual probabilities. The format of "sequence" is A B C D := D after C after B after A (as spoken or written in the language).

Table I: Recognition results achieved on the test sets with the used vocabularies, and the perplexity (cross-entropy) of the used grammar. The row “used subcorpora” shows the codes of the sessions used in the test sets, with the number of selected of utterances in the brackets.

	Isolated digits (I-test)	Application words (A-test)	Digit strings (BC-test)	City names (O-test)	Phon. rich words (W- test)
Accuracy [%]	94.62	95.92	96.27	84.28	86.86
Correctness [%]	97.85	98.24	97.58	90.64	91.24
Used subcorpora	I1 (186)	A1 – A6 (1166)	B1, C1 – C4 (788)	O1, O2, O3, O5, O7, O8 (1152)	W1 – W4 (784)
Vocabulary (words)	11	32	11	927	734
Grammar perplexity	2.28	3.08	3.57	12.44	76.86

Table II: Recognition results in word error rates (in %) for SpeechDat-compliant databases, reference recognizer version 0.95 [9]. The last line in the table represents our achieved results.

Test language/ Database	Isolated digits (I-test)	Phonetically rich words (W-test)	City names (O-test)	Application words (A-test)	Digit strings (BC-test)
Danish	1.04	64.38	15.82	2.36	2.30
English	1.69	38.74	12.64	2.62	3.93
German	0.80	8.70	6.00	2.40	2.70
Swedish	2.56	35.21	12.37	1.52	3.78
Swiss German	0.51	24.26	6.29	1.06	3.10
Norwegian	2.31	34.73	17.31	4.43	5.87
Slovenian	4.15	19.25	9.33	4.86	6.14
Slovak	0.54	14.16	3.70	1.72	2.66
Slovak SPHINX- II	5.38	13.14	15.27	4.08	3.37

5. CONCLUSION

The Slovak SpeechDat database has been used for speech recognition purposes already, using reference recognition system from the COST 249 community [9]. That system used continuous mixture density HMMs, and did not achieve real-time speech recognition speed. Our developed system achieves word error rates around 15 % for 1000 vocabulary domains. The system might be improved. Firstly, we used only flat initialization of the CI model parameter. In the future we plan to initialize our phone model parameters using automatic segmented and hand-labeled segments from the small part of the database. Secondly, for

constrained tasks we plan to use FSG grammars, which might be more suitable for those tasks than statistical n-grams. We deduce it from our results, where using n-grams, WER for isolated digits was higher than WER for connected isolated digits.

ACKNOWLEDGEMENTS

We would like to thank Dr. S. Lihan from TU Kosice, who provides us with the test set control files for the evaluation of the system, as were used in the evaluation of the COST 249 reference recognizer. This work was supported by the Slovak Agency for Science VEGA, grant No. 2/5124/25.

REFERENCES

- [1] S. Darzagin, L. Franekova, and M. Rusko: *Conversion and Synthesis of the Slovak Speech*. (in Slovak), Jazykovedný časopis, 45, Bratislava, 1994, No. 1, pp. 31-34.
- [2] S. Darzagin and M. Trnka: *Speaker independent speech recognition system in Slovak*. Proceedings of the international conference Telecommunications'95, Bratislava, Dom techniky ZSVTS, pp. 118-123.
- [3] H. van den Heuvel et al.: *SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed*. Proceedings EUROSPEECH 2001, Aalborg, Denmark, Vol 3., pp 2059-2062.
- [4] M. Rusko, M. Trnka, S. Darzagin, and M. Cernak: *Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis*. In Sojka et al. (Eds.), Proceedings TSD 2004, pp. 457-464.
- [5] R. Winski: *Definition of corpus, scripts and standards for fixed networks*. Technical report. SpeechDat-II, January 1997, Deliverable SD 1.1.1., workpackage WP1, <http://www.speechdat.org>.
- [6] SpeechTrain: acoustic modeler, Speech Software at CMU, March 2005, <http://www.speech.cs.cmu.edu/SphinxTrain/>.
- [7] P. Clarkson and R. Rosenfeld: *Statistical Language Modeling using the CMU-Cambridge Toolkit*, Proceedings EUROSPEECH 1997, Rhodes, Greece, Vol 5., 2707-2710.
- [8] HTK Speech Recognition Toolkit, Speech Vision and Robotics Group at CUED, March 2005, <http://htk.eng.cam.ac.uk/>.
- [9] M. Marcinek, J. Juhar, and A. Cizmar: *Phoneme ASR System for Slovak Speech Database*, V. Matousek et al. (Eds.), Proceedings TSD 2001, LNAI 2166, pp. 237 – 241, 2001.