Fast Phonetic/Lexical Searching in the Archives of the Czech Holocaust Testimonies: Advancing Towards the MALACH Project Visions

Josef Psutka, Jan Švec, Josef V. Psutka, Jan Vaněk, Aleš Pražák, and Luboš Šmídl

Department of Cybernetics, West Bohemia University, Pilsen, Czech Republic
{psutka,svec,psutka_j,vanekyj,aprazak,smidl}@kky.zcu.cz,
http://www.kky.zcu.cz

Abstract. In this paper we describe the system for a fast phonetic/lexical searching in the large archives of the Czech holocaust testimonies. The developed system is the first step to a fulfillment of the MALACH project visions [1,2], at least as for an easier and faster access to the Czech part of the archives. More than one thousand hours of spontaneous, accented and highly emotional speech of Czech holocaust survivors stored at the USC Shoah Foundation Institute as video-interviews were automatically transcribed and phonetically/lexically indexed. Special attention was paid to processing of colloquial words that appear very frequently in the Czech spontaneous speech. The final access to the archives is very fast allowing to detect segments of interviews containing pronounced words, clusters of words presented in pre-defined time intervals, and also words that were not included in the working vocabulary (OOV words).

1 Introduction

The principal goal of MALACH (Multilingual Access to Large Spoken Archives) project was to develop methods for improved access to large multilingual spoken archives collected by the Survivors of the Shoah Visual History Foundation (VHF) between 1994 and 1997. Today these archives of video-interviews are located in the Shoah Foundation Institute at the University of Southern California and contain approximately 52,000 interviews (testimonies) in 32 languages of personal memories of survivors of the World War II Holocaust (116,000 hours of video). More than 550 testimonies of this collection is in Czech (almost 1,000 hours of video). The MALACH project was carried out between 2002–2007 (in cooperation with the VHF, IBM, JHU Baltimore, University of Maryland, CU in Prague and UWB in Pilsen) with the financial support of the NSF. The objective of the MALACH project was to develop and verify techniques for speech recognition of spontaneous, accented and highly emotional speech of holocaust survivors. The plan was to use an output of the recognizer for automatic indexing of pronounced testimonies and automatic searching for keywords and topics. Using multilingual thesaurus this process should work across different languages. Although a great deal of work was done not all objectives were fully fulfilled [3,4]. This paper describes our two years effort to fulfill the MALACH project visions at least for access to the Czech part of archives.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 385–391, 2010.

The state-of-the-art techniques of acoustic and language modeling were applied to build up a LVCSR system which overcomes the former one [5] in recognition accuracy up to 9% absolutely. More than one thousand hours of speech of Czech holocaust survivors stored as video-interviews were then automatically transcribed and phonetically/lexically indexed. Special attention was paid to processing of colloquial words that appear very frequently in the Czech spontaneous speech. The final access to the archives is very fast, allowing to detect segments of interviews containing pronounced words, clusters of words presented in pre-defined time intervals, and also words that were not included in the working vocabulary (OOV words).

2 Characteristics of the Corpora

Testimonies of the Czech holocaust corpus as well as other languages are stored at the Shoah Foundation Institute (SFI) digital library as video interviews. The speech of each interview participant (the interviewer and interviewee) was usually recorded in a quiet rooms via lapel microphones that recorded speech on separate channels. The speech quality in individual interviews is however very poor from the ASR point of view, as it contains whispered or emotional speech with many disfluencies and non-speech events as crying, laughter etc. The speech was also often affected by using many colloquial (non-grammatical) words. The speaking rate (measured as the number of words uttered per minute) varies greatly depending on the speaker (the average age of all speakers was about 75 years), changing from 64 to 173 with the average of 113 [words/minute].

The average length of a Czech testimony is 1.9 hours. Each testimony was divided and stored at SFI as half-hour parts in MPEG-1 video files. For the further processing the audio streams were extracted. The audio track was stored at 128kb/sec stream in 16-bit resolution and 22.05 kHz sampling rate.

For preparing the acoustics 400 speakers were randomly selected. However, only 15 minute segment was transcribed for training purposes per each speaker. This training set contains 42% males and 58% females speakers (it corresponds to the whole database). Another entire 20 testmonies (10 males and 10 females) were transcribed for test a developement sets.

3 Building LVCSR

3.1 Annotation

The audio files were divided into segments and annotated using the special annotation software Transcriber 1.4.1, which is a tool for assisting the creation of speech corpora. Transcriber is freely available from the Linguistic Data Consortium (LDC) web site `http://www.ldc.upenn.edu/` (for details of the annotation process see [2]).

3.2 Acoustic Modeling

The acoustic training portion consisted of 100 hours of Czech speech. The data was parameterized as 15 dimensional PLP cepstral features including their delta and delta-delta derivatives (3 15=45 dimensional feature vectors) [6]. These features were

computed at rate of 100 frames per second. Cepstral mean subtraction was applied per speaker. The resulting triphone-based model was trained using HTK Toolkit. The number of clustered states and number of Gaussians mixtures per state was optimized using development test set and had more than 6k states and 16 mixtures per state (almost 100k Gaussians). A silence model was trained by borrowing Gaussians from all non-speech HMMs in proportion to their state and mixture occupancies. The resulting silence model contained 128 mixtures per state and was found to be useful in rejecting non speech events during recognition.

Speaker-adaptive models (SAT) were trained via fMLLR, for each training speaker. After fMLLR transforms for training speakers were computed against the original speaker-independent model, the original model was then re-estimated using the affinely transformed features. This process was repeated few times to converge. The DT model was developed from SAT model via four training iterations based on MMI-FD objective function [7]. Because the speaker identity is available, it can be used to improve the recognition. All training data were split to three clusters (male-speakers female-speakers and interviewer) for DT adaptation. This DT adaptation was done via two iterations DT-MAP on SAT-DT acoustic model [8].

3.3 Language Modeling

Since it is impractical to create enough language model training data by transcribing the speech, we investigated the use of other text collections to complement the transcriptions (see [9] for details). Two basic language models were trained. The first language model was trained on 5.6MB (1.1M tokens) of training set transcriptions. One of the most important issues that had to be decided before the transcription process started is the way of transcription of colloquial words. As explained in details in [4] the best performance of ASR is obtained by using the colloquial forms during acoustic model training while restricting the language model to the formal forms both in the lexicon and in the LM estimation process.

The second language model was trained from the selection of the Czech National Corpus (CNC). This corpus is relatively large (approximately 400M words) and is extremely diverse. Therefore we investigated the possibility of using automatic methods to select sentences from the CNC that are similar in language usage, lexicon and style to the sentences in the training set transcriptions. This in-domain selection from CNC contains 82MB of text (16M tokens). An interpolated language model has been created with the ratio 2:1 (transcriptions to the CNC). The resulting trigram language model with modified Kneser-Ney smoothing contains 252k words (308k phonetical variants). Language models were estimated using the SRI Language Modeling Toolkit (SRILM) [10].

3.4 Word and Phoneme Lattices

Due to the stereo speech signal (the interviewer on one channel and interviewee on the other) and to the conversation character of the data the special algorithm was introduced in order to reduce the huge amount of speech data. Only parts, where at least one speaker was talking, were further processed. In the case, where interviewer

and interviewee were cross-talking, both channels were recognized separately. This task was quite challenging because there occurred echoes even though the speakers had lapel microphones. During recording, the speech of interviewer and interviewee mixed together so that each speaker was recorded in both channels, only with different level of energy.

The LVCSR system was designed to work in two passes. In the first pass, clustered DT adapted acoustic models was automatically adapted to each of 550 speakers, using a bigram language model. This automatic iterative fMLLR+MAP adaptation [11] used only speech segments with posterior probabilities over 0.99. Word lattices were then generated based on information about word transitions performed during the second pass recognition by LVCSR system. The lattices were built up retroactively, from the last recognized word by adding other most probable word hypotheses (alternatives) of the recognized words according to the desired depth (number of concurrent hypotheses) of word lattices. For searching through word lattices, the posterior probability computed by forward-backward algorithm was assigned to each hypothesis in the word lattice. Normalized acoustic likelihoods and a trigram language model were used during the lattices computation. Due to the effect of the segmentation of the word graph [12], posterior probabilities for different hypotheses of the same word were summed.

Phoneme lattices were generated in the same manner, based on information about phoneme transitions performed during the recognition by phoneme recognizer without use of any language model. This recognizer was built for each speaker on its acoustic model adapted by the first pass of LVCSR system.

The parameters for the LVCSR system were optimized on the development data (whole testimonies of 5 male and 5 female speakers). The recognition results depicted in the Table 1 show the phoneme recognition accuracy as well as recognition accuracy for LVCSR system. This results were enumerated on the test set (another 5 male's and 5 female's testimonies). The total number of words in the test set was 63,205 with 2.39% out-of-vocabulary (OOV) words.

Table 1. The results of recognition experiment

Recognition level	Acc [%]
LVCSR	71.44
Phonemes	70.38

4 Indexing and Searching

4.1 Indexing

To achieve a very good responsiveness of the searching system, we decided to create an index of both the word lattices and the phonetic lattices. The index was created using the SQL database that was able to store the huge number of records. During the indexing procedure the structural properties of lattices were omitted from the database and only the word or phone occurrences were stored.

Word lattice index was relatively simple. Every lattice arc represented one word and the word with corresponding time was added into the index if the score of the arc exceeded some threshold. The index of phonetic lattice was more complicated because indexing single phones was not effective – the database query was not very specific and search would return many unrelated results. Therefore we decided to index trigrams of the subsequent phones. The trigrams are overlapped and can be simply generated from the phonetic lattice by virtue of using the breadth-first search. The score of the trigram was obtained as an average score of the three phones in the trigram. Again, only trigrams with the score larger then some threshold (different from the word lattice threshold) were added into the index.

In total, the database contains about 100M records and there is about 27 records for every second of the indexed audio signal. The index contains 63,329 unique trigrams of phones. For instance, the five most frequent are: `sem` (140,248 occurrences), `bil` (125,977), `tak` (120,972), `ese` (117,515) and `oto` (117,252).

4.2 Searching

Unlike the laboratory (off-line) system for full-scale IR such as [13] this system searches only keywords and/or key-phrases but is extremely fast and fully interactive. The searching algorithm depends on the searched word. First, a lemma of the searched word is generated. Then if the lemma is found in the vocabulary the word lattice search is performed. It is also possible to search for all possible forms (from the lemma) of the searched vocabulary word. This behavior can be optionally disabled if the user wants to search only the exact word form.

If the searched word is an out-of-vocabulary word the phonetic lattice search is performed. The phonetic transcription of the word is generated and the overlapping trigrams of phones are computed. Then the database query selects the records corresponding to these trigrams. The results are grouped by a corresponding audio track and ordered by the time. If the searched word occurs in the audio track at the given time there must be a cluster of the trigrams. The algorithm we use does not strictly require the presence of all trigrams from the searched word. The score of the word occurrence is computed from the score of indexed trigrams and the total number of found trigrams.

The time required to search through the whole archive strongly depends on the searched word itself, mainly on the number of occurrences. For in-vocabulary words the time needed to search the whole archive is typically between 5 and 10 seconds. The out-of-vocabulary words are typically searched between 30 and 60 seconds.

5 GUI Description

The graphical user interface is designed to be as simple as possible. We suppose that the users of our searching tool will be mostly non-technicians. The user enters the searched word into the text box and selects the channel or channels to be searched (the reporter and/or the witness). The user can also enter a phrase instead of a single word. There are also used some operators in the search engine to modify the number of results: to tag the word as required (it must occur in every result) the user should write the plus sign in

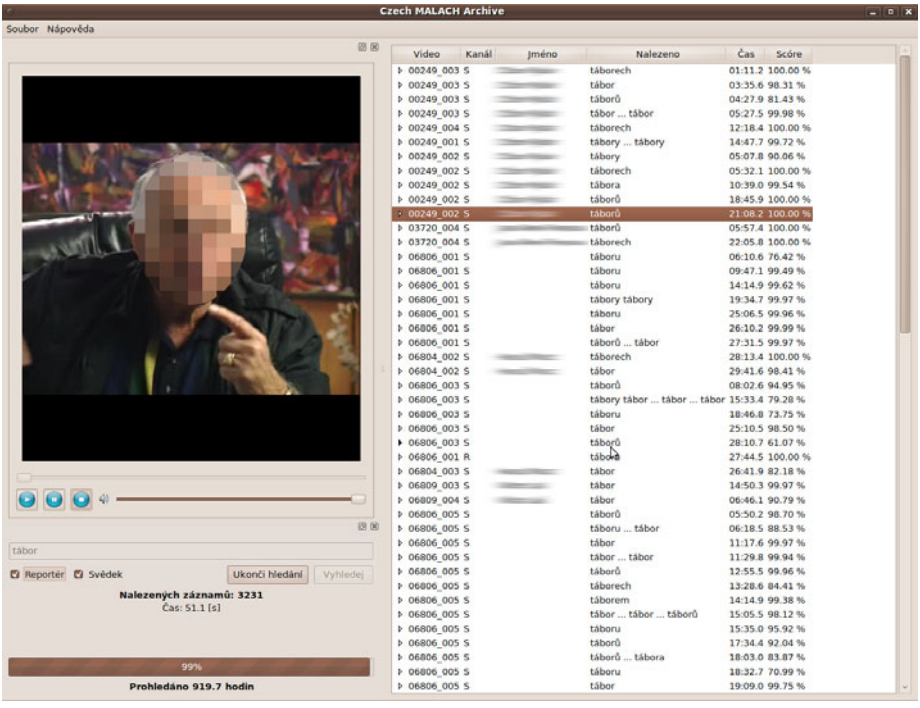

Fig. 1. Searching of the word *"tábor"*

front of it and to find the exact form of the word it should be enclosed into parenthesis. For example the searching of the word *"tábor"* (*camp*) can be seen in Figure 1.

6 Conclusion

This paper presents the system for a fast phonetic/lexical searching in the large archives of the Czech holocaust testimonies. Nearly 1,000 hours of interviews are searched typically up to 10 seconds for in-vocabulary words and up to 60 seconds for out-of-vocabulary words. Without this tool searching for a specific event or situation was complicated, since this huge amount of data had to be searched manually. This searching tool made the interviews more accessible to the historians, students, researchers, and actually to the whole public. By now is the Czech part of the holocaust survivors database stored at the SFI, the only one with such searching engine.

Acknowledgements

This research was supported by the Ministry of Education of the Czech Republic, projects No. MSM235200004 and No. LN00A063 and by the grant of The University of West Bohemia, project No. SGS-2010-054.

References

1. Byrne, W., Doerman, D., Franz, M., Gustman, S., Hajič, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.: Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. IEEE Transactions on Speech and Audio Processing 4, 420–435 (2004)
2. Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Gustman, S., Ramabhadran, B.: Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2002. LNCS (LNAI), vol. 2448, pp. 253–260. Springer, Heidelberg (2002)
3. Psutka, J., Ircing, P., Psutka, J.V., Hajič, J., Byrne, W., Mírovský, J.: Automatic Transcription of Czech, Russian and Slovak Spontaneous Speech in the MALACH Project. In: Interspeech Lisboa 2005, pp. 1349–1352. ISCA, Bonn (2005)
4. Psutka, J., Ircing, P., Hajič, J., Radová, V., Psutka, J., Byrne, W., Gustman, S.: Issues in Annotation of the Czech Spontaneous Speech Corpus in the MALACH Project. In: Fourth International Conference on Language Resources and Evaluation, pp. 607–610. European Language Resources Association, Lisbon (2004)
5. Psutka, J., Hajič, J., Byrne, W.: The Development of ASR for Slavic Languages in the MALACH Project. In: Acoustics, Speech, and Signal Processing, pp. 749–752. IEEE, Piscataway (2004)
6. Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. J. Acoustic. Soc. Am. 87 (1990)
7. Povey D.: Discriminative Training for Large Vocabulary Speech Recognition. Ph.D. thesis, Cambridge University, Department of Engineering (2003)
8. Vaněk, J., Psutka, J.V., Zelinka, J., Pražák, A., Psutka, J.: Discriminative Training of Gender-Dependent Acoustic Models. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS (LNAI), vol. 5729, pp. 331–338. Springer, Heidelberg (2009)
9. Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Mírovský, J., Gustman, S.: Large Vocabulary ASR for Spontaneous Czech in the MALACH Project. In: EUROSPEECH 2003 Proceedings, pp. 1821–1824. ISCA, Geneva (2003)
10. Stolcke, A.: SRILM – An Extensible Language Modeling Toolkit. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA (2002)
11. Zajíc, Z., Machlica, L., Müller, L.: Refinement Approach for Adaptation Based on Combination of MAP and fMLLR. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 274–281. Springer, Heidelberg (2009)
12. Wessel, F., Schlüter, R., Macherey, K., Ney, H.: Confidence Measures for Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing 9(3) (2001)
13. Ircing, P., Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 759–765. Springer, Heidelberg (2007)