EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive

Josef Psutka, Jan Švec, Josef V Psutka, Jan Vaněk, Aleš Pražák, Luboš Šmídl and Pavel Ircing*

Abstract

The main objective of the work presented in this paper was to develop a complete system that would accomplish the original visions of the MALACH project. Those goals were to employ automatic speech recognition and information retrieval techniques to provide improved access to the large video archive containing recorded testimonies of the Holocaust survivors. The system has been so far developed for the Czech part of the archive only. It takes advantage of the state-of-the-art speech recognition system tailored to the challenging properties of the recordings in the archive (elderly speakers, spontaneous speech and emotionally loaded content) and its close coupling with the actual search engine. The design of the algorithm adopting the spoken term detection approach is focused on the speed of the retrieval. The resulting system is able to search through the 1,000 h of video constituting the Czech portion of the archive and find query word occurrences in the matter of seconds. The phonetic search implemented alongside the search based on the lexicon words allows to find even the words outside the ASR system lexicon such as names, geographic locations or Jewish slang.

1 Introduction

The whole story of the cultural heritage archive that is in focus of our research and development effort began in 1994 when, after releasing "Schindler's List", Steven Spielberg was approached by many survivors who wanted him to listen to their stories of the Holocaust. Inspired by these requests, Spielberg decided to start the Survivors of the Shoah Visual History Foundation (VHF) so that as many survivors as possible could tell their stories and have them saved. In his original vision, he wanted the VHF (which later eventually became the USC Shoah Foundation Institute [1]) to perform several tasks, including collecting and preserving the Holocaust survivors' testimonies and cataloging those testimonies to make them accessible.

The "collecting" part of the mission has been completed, resulting into what is believed to be the largest collection of digitized oral history interviews on a single topic: almost 52,000 interviews of 32 languages, a total of 116,000 h of video. About half of the collection is in English, and about 4,000 of English interviews (approximately 10,000 h, i.e., 8% of the entire archive) have been extensively annotated by subject-matter experts (subdivided into topically coherent segments, equipped with a three-sentence summary and indexed with keywords selected from a pre-defined thesaurus). This annotation effort alone required approximately 150,000 h (75 person-years) and proved that a manual cataloging of the entire archive is unfeasible at this level of granularity.

This finding prompted the proposal of the MALACH project (Multilingual Access to Large Spoken Archives–years 2002-2007) whose aim was to use automatic speech recognition (ASR) and information retrieval techniques for access to the archive and thus circumvent the need for manual annotation and cataloging. There were many partners involved in the project (see the project website [2]), each of them possessing expertise in a slightly different area of the speech processing and information retrieval technology.

The goal of our laboratory was originally only to prepare the ASR training data for several Central and Eastern European languages (namely Czech, Slovak, Russian, Polish and Hungarian); over the course of the project, we gradually became involved in essentially all the research areas, at least for the Czech language. After the project has finished, we felt that although a great deal of work has been done (see for example [3-5]), some of the original project objectives still remained somehow

* Correspondence: ircing@kky.zcu.cz
Department of Cybernetics, University of West Bohemia, Plzeň, Czech Republic

unfulfilled. Namely, there was still no complete end-to-end system that would allow any user to type a query to the system and receive a ranked list of pointers to the relevant passages of the archived video recordings. Thus, we have decided to carry on with the research and fulfill the MALACH project visions at least for the Czech part of the archives. The portion of the testimonies that was given in Czech language is small when compared to the English part (about 550 testimonies, 1,000 h of video material), yet the amount of data is still prohibitive for complete manual annotation (verbatim transcription) and also poses a challenge when designing a retrieval system that works in (or very near to) real time.

The big advantage that our research team had when building a system for searching the archive content was that we had a complete control over all the modules employed in the cascade, from the data preparation works through the ASR engine to the actual search algorithms. That way we were well aware of inherent weaknesses of individual components and able to fine tune the modules to best serve the overall system performance.

The following sections will thus describe the individual system components, concentrating mainly on the advancements that were achieved after the original MALACH project was officially finished. But first let us briefly introduce the specific properties of the Czech language

2 Characteristics of the Czech language

Czech, as well as other Slavic languages (such as Russian and Polish, to name the most known representatives), is a richly inflected language. The declension of Czech nouns, adjectives, pronouns and numerals has 7 cases. Case, number (singular or plural) and gender (masculine, feminine or neuter) are usually distinguished by an inflectional ending; however, sometimes the inflection affects the word stem as well. The declension follow 16 regular paradigms but there are some additional irregularities.

The conjugation of Czech verbs distinguishes first, second and third person in both singular and plural. The third person in the past tense is marked by gender. The conjugation is directed by 14 regular paradigms but many verbs are irregular in the sense that they follow different paradigms for different tenses.

Word order is grammatically free with no particular fixed order for constituents marking subject, object, possessor, etc. However, the standard order is subject-verb-object. Pragmatic information and considerations of topic and focus also play an important role in determining word order. Usually, topic precedes focus in Czech sentences.

In order to make a language with such free word order understandable, the extensive use of agreement is necessary. The strongest agreement is between a noun and its adjectival or pronominal attribute: they must agree in gender, number and case. There is also agreement between a subject (expressed by a noun, pronoun or even an adjective) and its predicate verb in gender and number, and for pronouns, also in person. Verbal attributes must agree in number and gender with its related noun, as well as with its predicate verb (double agreement). Possessive pronouns exhibit the most complicated type of agreement–in addition to the above mentioned triple attributive agreement with the possessed thing, they must also agree in gender and number with the possessor. Objects do not have to agree with their governing predicate verb but the verb determines their case and/or preposition. Similarly, prepositions determine the case of the noun phrase following them [6].

It stems from the highly inflectional nature of the Czech language that the size of the ASR lexicon grows quite rapidly and the ASR decoder must be designed in such a way that it is able to cope with large vocabularies. Our recognition engine is indeed able to handle a lexicon with more than 500 thousand entries [7]. Unlike in the case of Turkish or Finnish, where the problem of vocabulary growth caused by agglutination was successfully addressed by decomposing words into morphemes [8], similar attempts for Czech have brought inconclusive results [9].

An interesting phenomenon occurring in the Czech language is a considerable difference between the written form of the language (Standard or Literary Czech) and the spoken form (Common Czech). This difference occurs not only on the lexical level (usage of Germanisms and Anglicisms), but also on the phonetic and morphological level. Some of the differences can even be formalized (see for example [10]). We had to address this issue during the development of both acoustic and language models (see the “Language model” section below).

3 Automatic speech recognition

3.1 Data preparation

The speech contained in the testimonies is very specific in many aspects, owing mostly to the very nature of the archive. The speakers are of course elderly (note that the recording started in 1995), and due to the character of their stories, their speech is often very emotional and contains many disfluences and non-speech events such as crying or whimpering. The speaking rate also varies greatly depending on the speaker, which again is frequently an issue related to age (some interviewees were so old that they struggled with the mere articulation, while others were still at the top of their rhetorical abilities) and/or the language environment where the speakers spent the last decades (as those living away from the Czech Republic naturally stopped to search for the correct expression more often).

Consequently, the existing annotated speech corpora were not suitable for training of the acoustic models, and we have to first prepare the data by transcribing a part of the archived testimonies.

We have randomly selected 400 different Czech testimonies from the archive and transcribed 15-min segment from each of them, starting 30 min from the beginning of the interview (thus getting past the biographical questions and initial awkwardness). Detailed description of the transcription format is given in [11]; let us only mention that in addition to the lexical transcription, the transcribers also marked several non-speech events. That way we have obtained 100 h of training data that should be representative of the majority of the speakers in the archive. Another 20 testimonies (10 male and 10 female speakers) were transcribed completely for the ASR development and test purposes.

3.2 Acoustic modeling

The acoustic models in our system are based on the state-of-the-art hidden Markov models (HMM) architecture. Standard 3-state left-to-right models with a mixture of multiple Gaussians in each state are used. Triphone dependencies (including the cross-word ones) are taken into account. The speech data were parameterized as 15-dimensional PLP cepstral features including their delta and delta-delta derivatives (resulting into 45-dimensional feature vectors) [12]. These features were computed at the rate of 100 frames per second. Cepstral mean subtraction was applied per speaker. The resulting triphone-based model was trained using HTK Toolkit [13]. The number of clustered states and the number of Gaussians mixtures per state were optimized using a development test set and had more than 6 k states and 16 mixtures per state (almost 100 k Gaussians).

As was already mentioned, non-speech events appearing in spontaneous speech of survivors were also annotated. We used these annotated events to train a generalized model of silence in the following manner:

We took the sets of Gaussian mixtures from all the non-speech event models including the standard model for a long pause (silence–**sil**–see [13]). Then we weighted those sets according to the state occupation statistics of the corresponding models and compounded the weighted sets together in order to create a robust "silence" model with about 128 Gaussian mixtures. The resulting model was incorporated into the pronunciation lexicon, so that each phonetic baseform in the lexicon is allowed to have either the short pause model (**sp**) or the new robust **sil** model at the end.

The described technique "catches" most of standard non-speech events appearing in running speech very well, which improved the recognition accuracy by eliminating many of the insertion errors.

The state-of-the-art speaker adaptive training and discriminative training [14] algorithms were employed to further improve the quality of the acoustic models. Since the speaker identities were known, we could split the training data into several clusters (male interviewees, female interviewees and interviewers) before the actual discriminative training adaptation (DT–see [15] for details) to enhance the method's effectiveness.

3.3 Language modeling

The language model used in the final system draws from the experience gained from the extensive experiments performed over the course of the MALACH project [16]. Those experiments revealed that even though the transcripts of the acoustic model training data constitute a rather small corpus from the language modeling point of view (approximately one million tokens), they are by far more suitable for the task than much larger, but "out-of-domain" text corpora (comprising, for example, newspaper articles). However, if a more sophisticated technique than just throwing in more data is used for extending the language model training corpus, it is possible to further improve the recognition performance. We have also found out that the spontaneous nature of the data brought up a need for careful handling of colloquial words that are abundant in casual Czech speech. It turned out that the best results were achieved when the colloquial forms are employed in the acoustic modeling stage only and the standardized forms are used as the "surface" forms in the lexicon of the decoder and in the language model estimation process (see [17] for details). In other words, the recognizer produces text in the standardized word forms, while the colloquial variants are treated as pronunciation variants inside the decoder lexicon.

In concordance with those findings, we have trained two basic language models. The first one was estimated using only the acoustic training set transcripts, and the second was trained from the selection of the Czech National Corpus (CNC). This corpus is relatively large (approximately 400 M words) and extremely diverse. Therefore, it was impractical to use the whole corpus, and we investigated the possibility of using automatic methods to select sentences from the CNC that are in some way similar to the sentences in the training set transcriptions. The method that we have used is based on [18] and employs two unigram language models–one of them (P_{CNC}) is estimated from the CNC collection, and the other (P_{Tr}) was estimated from the acoustic training set transcripts. A likelihood ratio test was applied to each sentence in the CNC, using a threshold *t*: a sentence *s* from the CNC was added to the filtered set (named CNC-S) if $P_{\mathrm{CNC}}(s) < t.P_{\mathrm{Tr}}(s)$. This is a simple way of assessing whether sentences from the CNC are closer to the testimony transcriptions than to the bulk of the CNC corpus itself. The test threshold effectively allowed us to determine the size of selected sub-corpus CNC-S. Gradually decreasing the threshold yields smaller and smaller subcorpora that, ideally, are more and more similar to the testimony transcriptions. A threshold of 0.8 created a CNC-S containing about 3% of the CNC (approximately 16 M tokens). Merging the lexicons from both CNC-S and acoustic training set transcripts and consequently interpolating corresponding language models yielded WER improvement of 2% absolute [16]. The interpolation ration 3:1 (transcriptions to the CNC-S) was used in the presented system as this factor gave the best recognition performance in the experiments [16]. The lexicon of the resulting trigram language model contains 252 k words (308 k pronunciation variants), 3.6 M bigrams and 1.3 M trigrams. Language models were estimated using the SRI Language Modeling Toolkit (SRILM) [19] employing the modified Kneser-Ney smoothing method [20].

3.4 Speech recognition: generation of word and phoneme lattices

There was an important issue to solve even before the actual speech recognition process started. That is, what speech signal should be actually recognized. The problem was that the signal extracted from the archive video recordings was stereo, one channel containing the speech of the interviewer and the other the speech of the interviewee. However, there were frequent echoes despite the fact that the speakers were wearing lapel microphones. This was particularly challenging in the event of cross-talking when the speech of both dialogue participants was mixed together in both channels and we have to design an algorithm for separating the speech that was based on the levels of energy. Also, to save the computational power and storage, we have omitted from recognition all the portions of the signal that did not contain any speech.

Then the processed signal was streamed into our in-house ASR system [21] that was used in two recognition passes. The first pass employs the trigram language model described in Section 3.3 and clustered DT adapted acoustic models that are automatically gradually adapted to each individual speaker. This unsupervised iterative speaker adaptation algorithm employs both fMLLR and MAP methods (see [22] for details) and uses for adaptation only the speech segments with confidence measure (expressed in our case in terms of posterior probabilities) exceeding 0.99, thus ensuring reliable estimates of the transformation matrices.

The speaker adapted models are then employed in the second pass to generate the lattices to be used in the search engine. In order to help the search algorithm, the lattices were equipped with a confidence scores computed as the posterior probabilities using the forward-backward algorithm. Both word and phoneme lattices were generated in this manner, important distinction being that the phoneme recognizer did not use any language model for the lattice generation.

The parameters of the ASR system were optimized on the development data (complete testimonies of 5 male and 5 female speakers). The recognition results listed in the Table 1 show the (one-best) phoneme recognition accuracy as well as recognition accuracy of the word-based system. This accuracy was computed on the test set comprising 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. Note that the accuracy of the Czech ASR reported just after the MALACH project completion was about 10% absolute lower (see [23]).

Using the lattices for searching is an important step away from the oversimplifying approach to speech retrieval on the same archive that was adopted by all teams participating in the CLEF campaign CL-SR tracks in 2006 [24] and 2007 [25], where the problem of searching speech was reduced to a classic document-oriented retrieval by using only the one-best ASR output and artificially creating "documents" by sliding a fixed-length window across the resulting text stream. The lattice-based approach, on the other hand, allows to explore the alternative hypotheses about the actual speech content–note that the one-best error rate is still rather high. Dropping the artificial segmentation into the (quite long) fixed-length documents then enables much more finely grained time resolution when looking for the relevant passages. This could save the users of the search engine a lot of browsing through unrelevant bits of the archive. Furthermore, the presence of phoneme lattices enables for searching of out-of-vocabulary terms (see more details in the following sections).

4 Indexing and searching

The general goal of the search system is rather clear and well defined. The task is to:

Table 1 Test set ASR results

Recognition units	Accuracy (%)
Words	72.89
Phonemes	70.38

1. Identify appropriate replay points in the recordings–that is, the moments where the discussion about the queried topics starts.
2. Present them in some user-friendly manner to the searcher.

However, there are many ways to approach this tasks. One of them is essentially a standard text retrieval that was used in the aforementioned CLEF campaign. The approach adopted in the presented work conforms to the definition of spoken term detection (STD) as given for example in [26]. This method does not care about the somehow abstract topic of the document (like traditional IR does or at least claims to) but instead it just looks for the occurrences of query terms. Unlike the keyword spotting methods, the STD uses a pre-built index for the actual query searching, making the search faster; it also means that the queries need not to be known beforehand.

4.1 Indexing

Separate indexes were built from the word and the phoneme lattices.

4.1.1 Word index

The construction of the word index was the easier task. In the word lattice, every arc represents one word and the weight of the arc denotes the confidence measure (expressed as posterior probability) associated with the given word. In order to reduce the size of the resulting index, two stages of pruning were applied. The first stage takes place at the beginning when all the arcs whose posterior probability is lower than a threshold θ_w are discarded (θ_w = 0.05). Each of the remaining arcs is represented by a 5-tuple:

$$(start\_t, end\_t, word, score, item\_id)$$

where *start_t* and *end_t* are the beginning and end time, respectively, *word* is the word (ASR lexicon item) associated with the arc, *score* is the aforementioned posterior probability and finally *item_id* is the identifier of the original video file (*start_t* and *end_t* represent the offset relative to the beginning of this file). The index is further pruned by removing similar items. If there are two arcs labeled with the same word that are either overlapping or are being less than Δt_w apart (Δt_w is set to 0.5 s), only the arc with the higher score is retained. It follows from the description that the indexing procedure omits the structural properties of the original lattice but, on the other hand, makes a compact and efficient representation of the recognized data. The total number of items in the resulting word index is approximately 12 M.

4.1.2 Phoneme index

The building of the phoneme index is more complicated. Having single phones as the index items was found to be ineffective as it produced a lot of false alarms. Therefore, the proposed algorithm traverses the lattice and collects triplets of adjacent arcs (i.e., trigrams of the subsequent phonemes) and immediately discards those trigrams that meet any one of the following conditions:

- one or more of the phonemes is a silence
- any two adjacent phonemes are identical
- posterior probability of any phoneme is lower than a threshold θ_p (θ_p = 0.05)

Each remaining trigram is then labeled with appropriate start and end time and with a combined score that is computed as the geometric mean of posterior probabilities of the individual phonemes. The geometric mean is used because it assigns, in comparison with the arithmetic mean, much lower probability to the trigrams where one of the phonemes has distinctively lower confidence score. This leads to the desirable elimination of the least promising paths in the lattice. The usage of geometric mean also facilitates the computation as the posterior probabilities of the lattice arcs are given in the logarithm form.

In the next step of the indexing algorithm, all the trigrams whose combined score falls below a threshold θ_C (θ_C = 0.1) are discarded. The remaining trigrams are then ordered on the time axis–if there are more triplets labeled with the same phoneme trigram within the window of the length Δt_p ($\Delta t_p = 0.03s$), only the triplet with the highest score is included in the index. All the algorithm steps naturally again cause the structural properties of the lattice to be omitted. Finally, the same 5-tuples representing each remaining arc as in the case of the word index are stored in the database, and only now the word is replaced with the numeric ID representing given phoneme trigram. There are approximately 63 k different phoneme trigrams in the final index, and the number of items exceeds 88 M.

4.2 Searching

When searching the word index, all possible phonetic transcriptions (phoneme representations) of the query word are found in the lexicon. Then those phoneme sequences are mapped back to all corresponding word forms from the lexicon. This allows to search simultaneously, for example, for both the English and Czech spelling variants of a word (e.g., **Shoah** and the Czech transliteration **Šoá**). The system also makes possible to search for all inflected forms of a given word. If this feature is enabled, the lemma is also found for each of the query words. Consequently, the set of query words is extended with all possible word forms found in the vocabulary for each of the lemmas (these linguistic processing steps are done using a method described in [27]).

Search in the phoneme index takes place when the query word is an OOV or when it is forced by the user. The query word is again transcribed into a sequence of phonemes (we have a rule-based system for phonetic transcription, and thus, the phoneme representation can be obtained easily even for an OOV word). Then for each of the phoneme strings, the following steps are performed:

1. The consecutive phone trigrams are generated–e. g., the word '**válka**' (the war) is decomposed into '**v aa l**', '**aa l k**' and '**l k a**'.
2. All those trigrams are simultaneously searched for in the phoneme index and ordered according to the video file ID and the starting time.
3. For each video file ID, the found trigrams are clustered together on the time axis so that the time gap between clusters is at least equal to θ_{search} ($\theta_{\text{search}} = 0.2s$).
4. Every such cluster is then assigned a score that is computed as

$$score_{\text{comb}} = (1 - \lambda)score_{\text{ACM}} + \lambda\ score_{\text{hit}}$$

where $score_{\text{ACM}}$ is the arithmetic mean of scores of the phoneme index items in the cluster and $score_{\text{hit}}$ is the ratio between the number of trigrams that were correctly found in the given cluster and the number of trigrams representing the searched word. This implies that the algorithm does not strictly require the presence of all trigrams from the query. The interpolation coefficient was tuned using development data and consequently set to $\lambda = 0.6$. The $score_{\text{comb}}$ then serves as the ultimate relevance score.

The presented system also provides some functionality that allows searching for phrases of several words. Every word in the query phrase can be marked as either mandatory or optional. The search algorithm then:

1. Looks for individual words and orders the results on the time axis (separately for each video file)
2. Clusters the results so that the time gap between the clusters is at least $\theta_{\text{phrase-search}} = 10s$
3. Discards all clusters that do not contain all mandatory words.
4. Assigns each cluster a score that is computed as the arithmetic mean of the individual word scores.

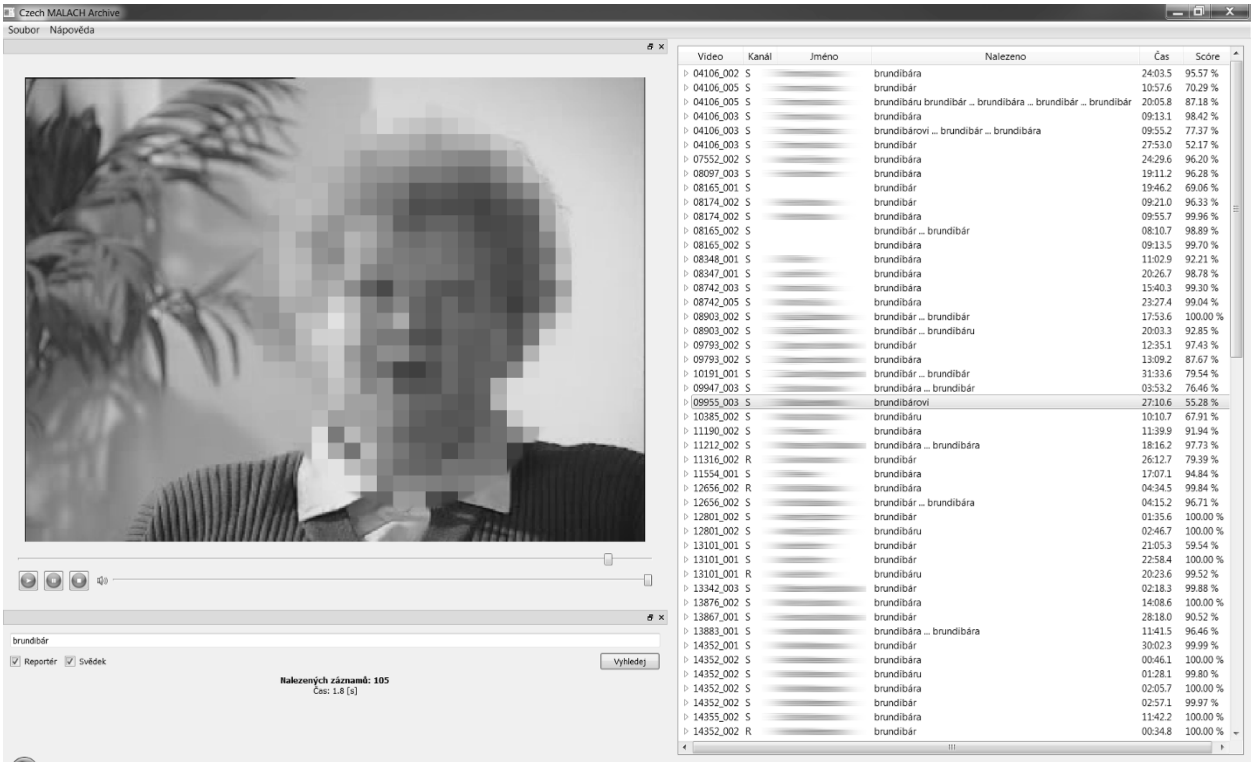
4.3 GUI description and HW/SW implementation details

The graphical user interface is designed with the IT non-professional in mind and is therefore as simple as possible (see the Figure 1). In the lower left corner, it has a text box for entering query word/phrase and check boxes for selecting the channels to be searched (interviewer and/or interviewee). The query can be modified using a set of simple operators–the plus sign is used to mark mandatory words and enclosing a word in parentheses tells the search engine that it should look for the exact word form only (i.e., the default expansion to all possible word forms is disabled). The retrieved results are shown in the right half of the GUI window. Each item shows the unique video file ID, the channel, the speaker's name, the exact form of the word or the phrase that was found, the time when the word/phrase occurs and the relevance score. The upper left corner then contains the multimedia player with the usual controls that allows to immediately replay any video file listed in the result window, starting several second prior to the query occurrence.

The search engine was implemented with a specific focus on the retrieval speed and on the system scalability. We also wanted to run the search algorithm on a portable equipment so that we can disseminate the research results at various forums. Thus, we have decided to employ SQL database server architecture for storage of both word and phoneme indexes in order to ensure fast system response (as the SQL access algorithms are well optimized for speed). The speed is further improved by storing the database on the 64 GB SSD drive instead of the conventional HDD. Other parameters of the hardware are rather moderate (HP EliteBook 8730w with Intel® Core™2 Duo Processor 2.80 GHz, 4 GB RAM). The video files with the actual testimonies are stored on two external USB hard drives (1 TB each). The system architecture supports remote access to the database which enables to run the search algorithm on different portions of the archive in parallel using several CPUs and therefore allows to scale the system to much larger archives rather seamlessly.

5 Evaluation

The quality of the STD was evaluated using two sets of queries whose occurrences have been manually annotated in the test portion of the video data. The first set (SetIn) contains 20 words that are present in the ASR lexicon (and consequently also in the word index); the total number of occurrence of SetIn words in the test data is 374. The second set (SetOut) consists of 108 words that are not included in the ASR lexicon and thus can be found only using the phoneme-based search; those words occur in the test data 414 times in total. The detection results are shown in Figure 2 (DET plot). In addition, the figure-of-merit (FOM) and equal-error-rate (EER) values are given in Table 2.



brundibár

☑ Reportér ☑ Svědek

Nalezených záznamů: 105
Čas: 1.8 [s]

Video	Kanál	Jméno	Nalezeno	Čas	Scóre
04106_002	S	[redacted]	brundibára	24:03.5	95.57 %
04106_005	S	[redacted]	brundibár	10:57.6	70.29 %
04106_005	S	[redacted]	brundibáru brundibár ... brundibára ... brundibár ... brundibár	20:05.8	87.18 %
04106_003	S	[redacted]	brundibára	09:13.1	98.42 %
04106_003	S	[redacted]	brundibárovi ... brundibár ... brundibára	09:55.2	77.37 %
04106_003	S	[redacted]	brundibár	27:53.0	52.17 %
07552_002	S	[redacted]	brundibára	24:29.6	96.20 %
08097_003	S	[redacted]	brundibára	19:11.2	96.28 %
08165_001	S		brundibár	19:46.2	69.06 %
08174_002	S	[redacted]	brundibár	09:21.0	96.33 %
08174_002	S	[redacted]	brundibára	09:55.7	99.96 %
08165_002	S		brundibár ... brundibár	08:10.7	98.89 %
08165_002	S		brundibára	09:13.5	99.70 %
08348_001	S	[redacted]	brundibára	11:02.9	92.21 %
08347_001	S	[redacted]	brundibára	20:26.7	98.78 %
08742_003	S	[redacted]	brundibára	15:40.3	99.30 %
08742_005	S	[redacted]	brundibára	23:27.4	99.04 %
08903_002	S	[redacted]	brundibár ... brundibár	17:53.6	100.00 %
08903_002	S	[redacted]	brundibár ... brundibáru	20:03.3	92.85 %
09793_002	S	[redacted]	brundibár	12:35.1	97.43 %
09793_002	S	[redacted]	brundibára	13:09.2	87.67 %
10191_001	S	[redacted]	brundibár ... brundibár	31:33.6	79.54 %
09947_003	S	[redacted]	brundibára ... brundibár	03:53.2	76.46 %
09955_003	S	[redacted]	brundibárovi	27:10.6	55.28 %
10385_002	S	[redacted]	brundibáru	10:10.7	67.91 %
11190_002	S	[redacted]	brundibára	11:39.9	91.94 %
11212_002	S	[redacted]	brundibára ... brundibára	18:16.2	97.73 %
11316_002	R	[redacted]	brundibár	26:12.7	79.39 %
11554_001	S	[redacted]	brundibára	17:07.1	94.84 %
12656_002	R	[redacted]	brundibára	04:34.5	99.84 %
12656_002	S	[redacted]	brundibár ... brundibára	04:15.2	96.71 %
12801_002	S	[redacted]	brundibár	01:35.6	100.00 %
12801_002	S	[redacted]	brundibáru	02:46.7	100.00 %
13101_001	S	[redacted]	brundibár	21:05.3	59.54 %
13101_001	S	[redacted]	brundibár	22:58.4	100.00 %
13101_001	R	[redacted]	brundibáru	20:23.6	99.52 %
13342_003	S	[redacted]	brundibár	02:18.3	99.88 %
13876_002	S	[redacted]	brundibára	14:08.6	100.00 %
13867_001	S	[redacted]	brundibár	28:18.0	90.52 %
13883_001	S	[redacted]	brundibára ... brundibára	11:41.5	96.46 %
14352_001	S	[redacted]	brundibár	30:02.3	99.99 %
14352_002	S	[redacted]	brundibára	00:46.1	100.00 %
14352_002	S	[redacted]	brundibáru	01:28.1	99.80 %
14352_002	S	[redacted]	brundibára	02:05.7	100.00 %
14352_002	S	[redacted]	brundibár	02:57.1	99.97 %
14355_002	S	[redacted]	brundibára	11:42.2	100.00 %
14352_002	R	[redacted]	brundibár	00:34.8	100.00 %

Figure 1 Screenshot of a GUI window.

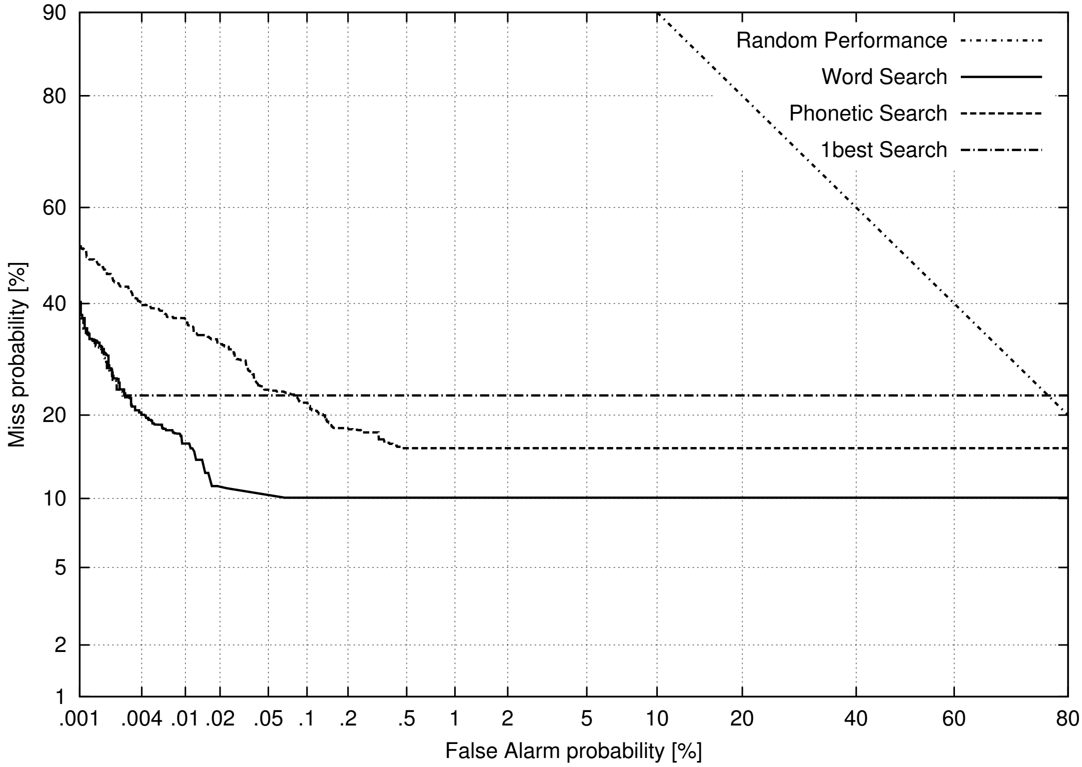


Figure 2 Detection error tradeoff (DET) plot.

Table 2 Spoken term detection results

	SetIn (lexicon words)		SetOut (OOVs)	
Lattice type	FOM (%)	EER (%)	FOM (%)	EER (%)
1-best ASR	79.60	20.86	-	-
Word lattice	93.73	19.79	-	-
Phoneme lattice	73.69	45.99	75.41	52.17

The plots reveal that the number of false alarms is essentially the same for search in 1-best ASR output (i.e., the situation when only the best path through the lattice is retained) and search in the word lattice, up to the point where the 1-best system is not able to provide any more correct hits and the lattice search becomes the clear winner. Searching in the phoneme lattice, on the other hand, produces substantially more false alarms than both the word-based algorithms, yet its big advantage lies in the ability to search for the words that are missing from the ASR lexicon. Those missing words are often some rare personal and place names and just as they are underrepresented in the language model training data, they are also very important to the searchers of the collection [28] (in fact, two-thirds of the requests specified named entities in the preliminary MALACH user studies). The phonetic search mode also allows users to type only an "approximate" spelling of the searched query which is extremely helpful especially in the case of foreign words or even words that are transliterated from different alphabets and it is not clear what spelling variant (if any) appears in the ASR lexicon.

One of the key considerations of our STD engine design was the focus on the quick response of the system. The following section is therefore devoted to the evaluation of the retrieval speed. Figure 3 depicts the histograms of search times for both the lexical and phonetic searches (we define the search time as the period between entering the query and the moment when all the found segments are presented to the user). It shows that the statistical mode of the lexical search time is only 0.5 s and the vast majority of the searches is finished in less than 5 s. For the search in the phonetic lattices, the statistical mode is 7.5 s and the majority of the searches take less than 20 s which we find still very reasonable. Figure 4 further reveals that in the case of

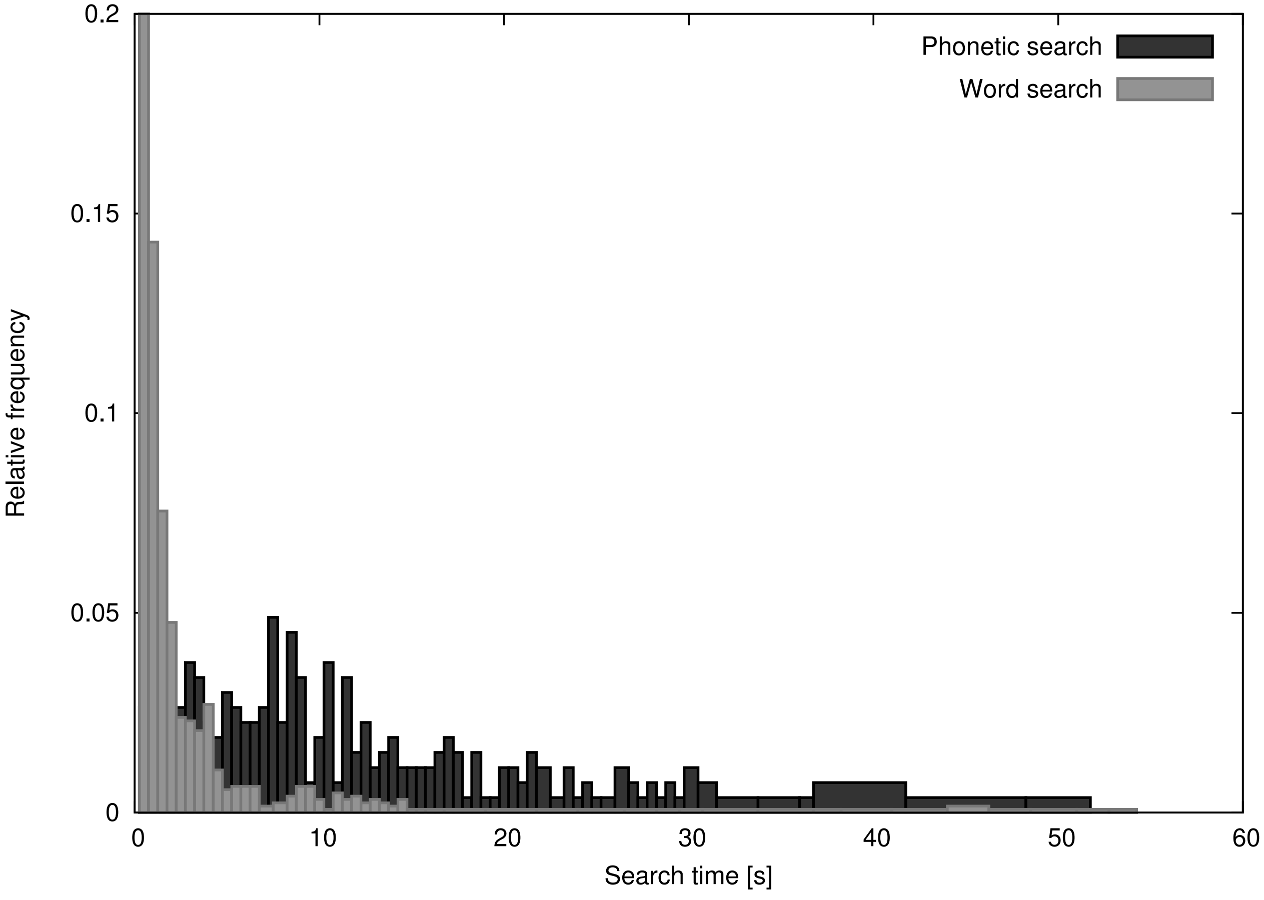


Figure 3 Histogram of search times.

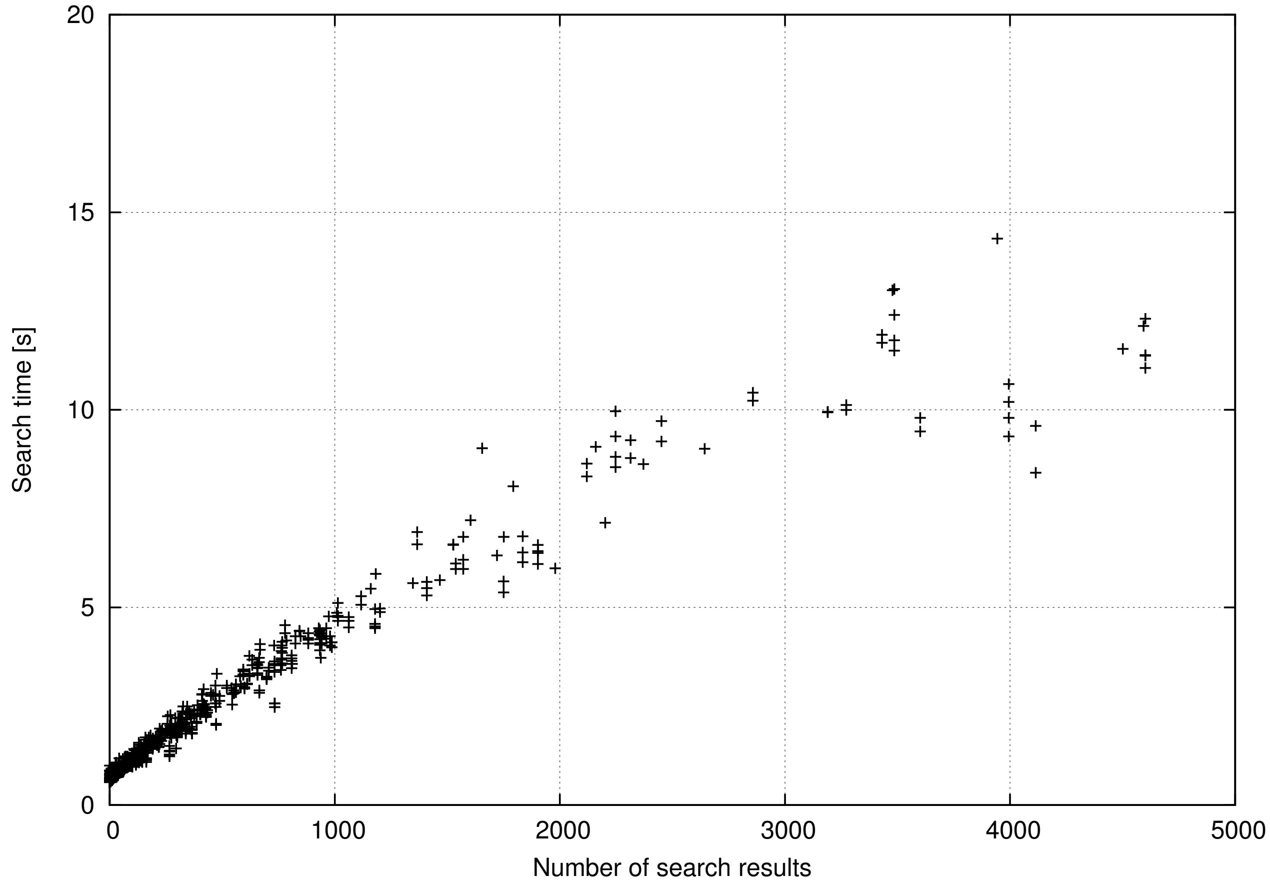


Figure 4 Dependency of a lexical search time on a number of search results.

searching the word lattices, the search time is more or less linear to the number of retrieved results, It is because retrieving one word hit requires just a single SQL query that takes always the same time and no further processing is necessary. On the other hand, the phonetic search time dispersion that could be observed in Figure 5 is attributed to the fact that large number of individual phoneme trigrams is retrieved first for all queries and those trigrams are then clustered and filtered to produce the list of relevant results. The search time then depends mainly on the query length (see Figure 6) because this is the factor that influences the number of individual phoneme trigrams that are returned in the first retrieval step.

6 Conclusions

The paper introduced the system for searching spontaneous speech data that was built in an effort taken to advance toward the ultimate goals of the MALACH project. The novelty of our contribution lies mainly in the fact that we have managed to develop end-to-end system that incorporates all the state-of-the-art components necessary for processing audio archives to the form that is directly searchable in real time. The methodology developed within the research effort described in this paper will be directly applicable in the related tasks of indexing various audiovisual archives. The demand for such a technology is currently in the upswing as the volume of unstructured audio data available in digital form is growing with an unprecedented pace. Actually, we are already trying to address a similar research issue in a joint project with Czech Television (national public broadcaster) that needs a technology for assisted cataloging of its evergrowing archive.

Both the objective evaluation presented in this paper and the positive feedback that the researchers were receiving during several live demonstrations suggest that the work was successful and that we have created fast and efficient system. It could make the Czech interviews more accessible to the historians, filmmakers, students and of course also to the general public. Negotiations concerning the system deployment in the Malach Center for Visual History in Prague are currently taking place, as well as the preparation of the joint project with the USC Shoah Foundation Institute that would aim at the development of the English version of the system.

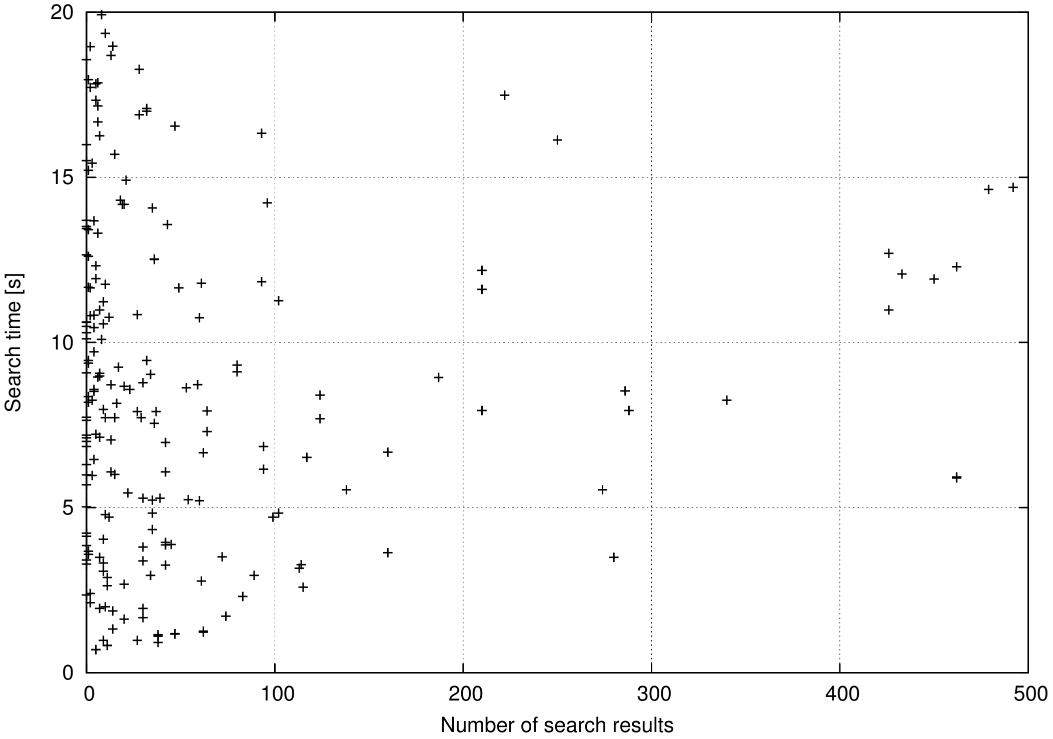


Figure 5 Dependency of a phonetic search time on a number of search results.

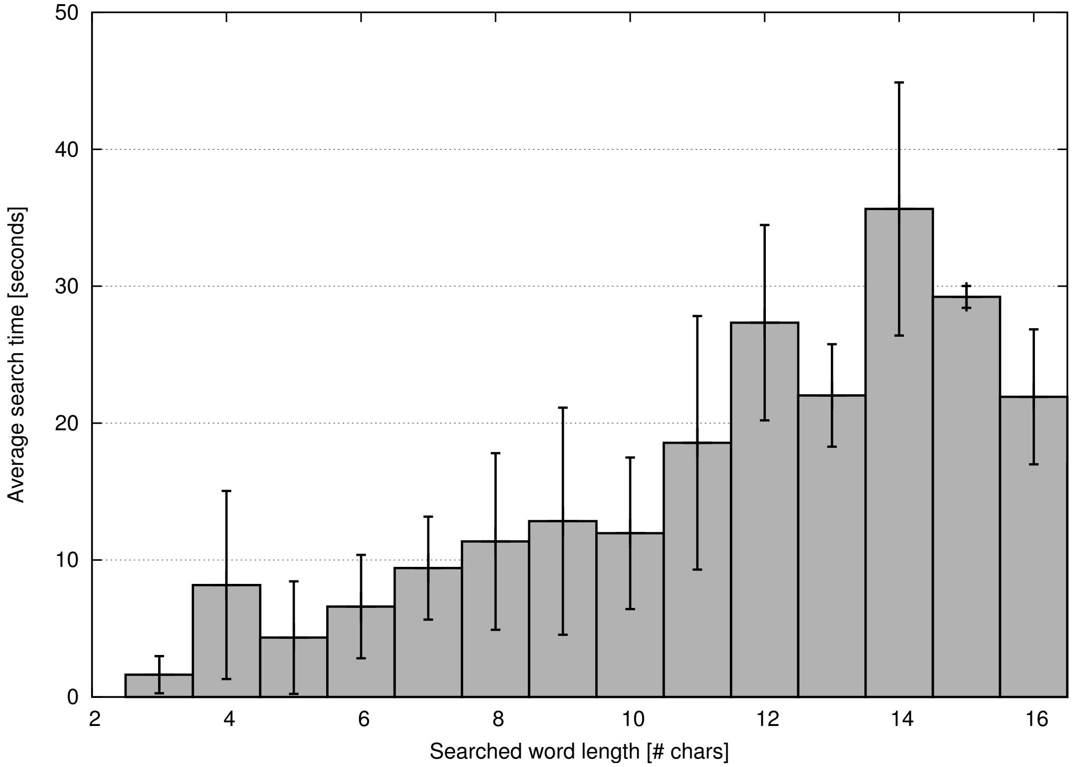


Figure 6 Dependency of a phonetic search time on a searched term length.

Endnotes

[a]The Czech speech resources were scarce anyway in the time when the MALACH project started, not to mention their domain incompatibility.[b]Some keyword spotting algorithms build specific acoustic models for each query and then process entire archive using those models, which is of course intractable in real time. [c]Note that Czech is a language with highly inflectional morphology.

Acknowledgements
This work was supported by the Grant Agency of the Czech Academy of Sciences project No. 1QS101470516 and the Ministry of Education, Youth and Sports of the Czech Republic project No. LM2010013 (LINDAT-Clarin). The authors would like to thank the USC Shoah Foundation Institute for assistance with getting all the necessary archive files and also for inviting us to present our results at the USC Digital Access, Outreach, and Research Conference in March 2010.

Competing interests
The authors declare that they have no competing interests.

Received: 26 July 2011 Accepted: 5 December 2011
Published: 5 December 2011

References

1. USC Shoah Foundation Institute http://college.usc.edu/vhi/ (2011)
2. MALACH Multilingual Access to Large Spoken Archives http://malach.umiacs.umd.edu/ (2007)
3. W Byrne, D Doermann, M Franz, S Gustman, J Hajič, D Oard, M Picheny, J Psutka, B Ramabhadran, D Soergel, T Ward, WJ Zhu, Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans Speech Audio Process. **12**(4), 420–435 (2004). doi:10.1109/TSA.2004.828702
4. P Ircing, J Psutka, J Vavruška, What can and cannot be found in Czech spontaneous speech using document-oriented IR methods–UWB at CLEF 2007 CL-SR track, in *Advances in Multilingual and Multimodal Information Retrieval, Lecture Notes in Computer Science*, vol. 5152, ed. by Peters C, Jijkoun V, Mandl T, Müller H, Oard D, Peńas A, Petras V, Santos D (Springer, Berlin, 2008), pp. 712–718. doi:10.1007/978-3-540-85760-0_90
5. P Ircing, JV Psutka, J Psutka, Using Morphological Information for Robust Language Modeling in Czech ASR System. IEEE Trans Audio Speech Lang Process. **17**(4), 840–847 (2009)
6. M Grepl, Z Hladká, M Jelínek, P Karlík, M Krčmová, M Nekula, Z Rusínová, D Šlosar, *Příruční mluvnice češtiny* (NLN, Praha, 1996)
7. A Pražák, P Ircing, J Švec, J Psutka, Efficient combination of N-gram language models and recognition grammars in real-time LVCSR decoder, in *Proceedings of ICSP 2008*, China, 587–591 (2008)
8. M Creutz, T Hirsimäki, M Kurimo, A Puurula, J Pylkkönen, V Siivola, M Varjokallio, E Arisoy, M Saraçlar, A Stolcke, Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Trans Speech Lang Process. **5**, 1–29 (2007)
9. W Byrne, J Hajič, P Ircing, P Krbec, J Psutka, Morpheme based language models for speech recognition of Czech, in *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 1902, ed. by Sojka P, Kopecek I, Pala K (Springer, Berlin, 2000), pp. 139–162. doi:10.1007/3-540-45323-7_24
10. P Sgall, J Hronek, A Stich, J Horecký, (eds.), *Variation in Language: Code Switching in Czech as a Challenge for Sociolinguistics* (John Benjamins, Amsterdam, 1992)
11. J Psutka, J Hajič, W Byrne, The development of ASR for Slavic languages in the MALACH project, in *Proceedings of ICASSP 2004*, Montreal, Canada, 749–752 (2004)
12. H Hermansky, Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am. **87**(4), 1738–1752 (1990). doi:10.1121/1.399423
13. S Young, D Kershaw, J Odell, D Ollason, V Valtchev, P Woodland, *The HTK Book* (Entropic, Cambridge, 2000) http://htk.eng.cam.ac.uk/
14. D Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, PhD thesis (University of Cambridge, Cambridge, 2003).
15. J Vaněk, J Psutka, J Zelinka, A Pražák, J Psutka, Discriminative training of gender-dependent acoustic models, in *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 5729, ed. by Ma-toušek V, Mautner P (Springer, Berlin, 2009), pp. 331–338. doi:10.1007/978-3-642-04208-9_46
16. J Psutka, P Ircing, JV Psutka, V Radová, W Byrne, J Hajič, J Mírovský, S Gustman, Large vocabulary ASR for spontaneous Czech in the MALACH project, in *Proceedings of Eurospeech 2003* Geneva, Switzerland, 1821–1824 (2003)
17. J Psutka, P Ircing, J Hajič, V Radová, JV Psutka, W Byrne, S Gustman, Issues in annotation of the Czech spontaneous speech corpus in the MALACH project, in *Proceedings of LREC 2004* Lisbon, Portugal, 607–610 (2004)
18. R Iyer, M Ostendorf, H Gish, Using out-of-domain data to improve in-domain language models. IEEE Signal Process Lett. **4**(8), 221–223 (1997). doi:10.1109/97.611282
19. A Stolcke, SRILM–an extensible language modeling toolkit, in *Proceedings of ICSLP 2002*, Denver, USA, 901–904 (2002)
20. SF Chen, J Goodman, *An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98* (Computer Science Group, Harvard University, Cambridge, MA, 1998)
21. A Pražák, L Müller, JV Psutka, J Psutka, LIVE TV SUBTITLING–fast 2-pass LVCSR system for online subtitling, in *Proceedings of SIGMAP 2007*, Barcelona, Spain, 139–142 (2007)
22. Z Zajíc, L Machlica, L Müller, Refinement approach for adaptation based on combination of MAP and fMLLR, in *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 5729, ed. by Matoušek V, Mautner P (Springer, Heidelberg, 2009), pp. 274–281. doi:10.1007/978-3-642-04208-9_39
23. P Ircing, JV Psutka, J Psutka, A Pražák, Z Tychtl, Automatic speech recognition and information retrieval techniques for facilitating access to video archives of cultural heritage, in *Proceedings of DHMS 2008*, Athens, Greece, 323–328 (2008)
24. P Ircing, L Müller, Benefit of proper language processing for Czech speech retrieval in the CL-SR task at CLEF 2006, in *Evaluation of Multilingual and Multi-modal Information Retrieval, Lecture Notes in Computer Science*, vol. 4730, ed. by Peters C, Clough P, Gey F, Karlgren J, Magnini B, Oard D, de Rijke M, Stempfhuber M (Springer, Berlin, 2007), pp. 759–765. doi:10.1007/978-3-540-74999-8_95
25. P Pecina, P Hoffmannová, G Jones, Y Zhang, D Oard, Overview of the CLEF-2007 cross-language speech retrieval track, in *Advances in Multilingual and Multimodal Information Retrieval, Lecture Notes in Computer Science*, vol. 5152, ed. by Peters C, Jijkoun V, Mandl T, Müller H, Oard D, Peńas A, Petras V, Santos D (Springer Berlin, 2008), pp. 674–686. doi:10.1007/978-3-540-85760-0_86
26. NIST Spoken Term Detection Portal http://www.itl.nist.gov/iad/mig//tests/std/ (2006)
27. J Kanis, L Müller, Automatic lemmatizer construction with focus on OOV words lemmatization, in *Text, Speech and Dialogue, Lecture Notes in Computer Science*, vol. 3658, ed. by Matoušek V, Mautner P, Pavelka T (Springer, Berlin, 2005), pp. 132–139. doi:10.1007/11551874_17
28. S Gustman, D Soergel, D Oard, W Byrne, M Picheny, B Ramabhadran, D Greenberg, Supporting access to large digital oral history archives, in *Proceedings of JCDL 2002*, Portland, USA, 18–27 (2002)

doi:10.1186/1687-4722-2011-10
Cite this article as: Psutka *et al.*: System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. *EURASIP Journal on Audio, Speech, and Music Processing* 2011 **2011**:10.