Study on Phrases Used for Semi-Automatic Text-Based Speakers' Names Extraction in the Czech Radio Broadcasts News

Michaela Kuchařová[1], Svatava Škodová[2], Ladislav Šeps[1], and Marek Boháč[1]

[1] Institute of Information Technology and Electronics, Technical University of Liberec, Studentskáa 2, 461 17, Liberec, Czech Republic
{michaela.kucharova1,ladislav.seps,marek.bohac}@tul.cz

[2] Department of the Czech Language and Literature, Technical University of Liberec, Studentská 2, 461 17, Liberec, Czech Republic
{svatava.skodova}@tul.cz

Abstract. In this paper we introduce a methodology leading to the extension of speakers' database used in the process of automatic transcription of spoken documents stored in the largest Czech Radio audio archive. We address the issue of the conversion of spoken speech to written texts – the automatic detection of speakers and their names. We work with a subset of the archive that consists of 8,020 hours of broadcasting news and 58,914,179 words within the years 1968–2011. We observed the occurrence of thousands of speakers' names during the period and therefore it is necessary to use their automatic or semi-automatic identification. Another investigated issue leading to the extension of speakers' database is the co-occurrence of a speaker's name in a specific phrase in the text transcription linked with the speaker's change in the audio recording.

Keywords: audio archive processing, spoken formal speech, radio broadcast news, automatic speech recognition, text-based speakers' names extraction

1 Introduction

The research described in this paper comes from the needs of a project which aims at making the Czech Radio audio archive publicly accessible and searchable [1]. This task demands many speech recognition technologies: i) a Large-Vocabulary Continuous Speech Recogniser (LVCSR) able to recognise Czech and Slovak (and to automatically determine which language is being spoken), ii) a document segmentation module able to split the document into speaker-homogenous segments, iii) a speaker identification tool and iv) a post-processing module, which makes the recognised text comfortably readable. The third mentioned module – the speaker identification tool – must be provided with a database of speakers. This database can be viewed at two levels – firstly, as a list of important speakers and persons (speakers who appear or are mentioned often enough) with some basic information (e.g. gender, language) and secondly, as the training data and corresponding model linked to one of the enlisted speakers.

As the oldest recordings are 90 years old and the amount of the recordings approaches 100,000 hours, the number of speakers occurring in the broadcasts is

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 416–423, 2014.

enormous. Hence we need an unsupervised way to perform the following tasks: i) to find names in the recognised text (this part covers the problem of recognition of surnames originating in common words, i.e., *Jan Medvěd* (bear) is recognised as *Jan medvěd*), ii) to decide whether the person is a speaker or he/she is only mentioned by the actual speaker (e.g. *"reporter X.Y. is speaking from the country Z"* vs. *"president X.Y. visited the country Z"*) and iii) to estimate if there is a recording segment pronounced by the found speaker (e.g. *"our reporter X.Y. is reporting from…"*).

To improve the automatic transcripts we plan to use a speaker adaptation technique [2] that can improve the recognition accuracy by approx. 4%. For this technique we need to identify speakers and we have to prepare acoustic data for each speaker.

2 Technical Background – Systems and Programs

For our research, we needed several systems and programs described in this part. First, an automatic speech recognition system, its description is given in Sec. 2.1; next, a program NanoTrans [3] to correct several transcripts that we use for training and testing. Above that, we had to prepare several scripts to find names, phrases and to count the number of their occurrences. All these scripts were especially designed to work with a vast amount of data – the transcripts of the recordings are stored in 18,239 files.

2.1 Automatic Speech Recognition System

The transcriptions of recordings of spoken documents stored in the Czech Radio audio archive were prepared by the standard LVCSR system [1]. This system was developed at the Technical University of Liberec and processes 16 kHz audio data, parameterises them into 39 MFCC features, and then applies global or floating CMS. The acoustic model uses a context-dependent triphone HMMs to represent 41 Czech phonemes and seven types of noises. It has been trained on 320 hours of speech (microphone and broadcasting speech). The decoder works in real time with a vocabulary whose size is about 500,000 lexical words. The language model is based on bigrams and is smoothed by the Kneser-Ney method.

2.2 Document Segmentation and Speaker Identification Modules

Document segmentation is a part of diarisation. It follows a standard framework consisting of three parts – voice activity detection, speaker turn detection and speaker clustering. You can see more in [2]. This module splits transcript into segments based on distinct speakers detected locally in a given recording.

The speaker identification module assigns the speaker (his/her ID or name) to the locally identified segments. Input to this module are speaker models that are trained from an acoustic data labelled by speaker ID. For our purpose we used the first mentioned module, and we attempted to find adequate acoustic data for the second one.

2.3 Phrase Patterns Used for the Proper Names Search

Experiments performed previously [4] indicated that the limiting factor for crawling through the transcriptions is the actual read speed of a hard drive, not the speed of a CPU. The more complex nature of present experiments described in this paper prevented us from using simple text-based searches because more complex pattern- matching techniques were required. The regular expressions proved to be too slow for the given task (as one experiment took more than 3 hours). To improve the search speed, we implemented our own simple pattern-matching algorithm. To make it as fast as possible, we limited the smallest searchable unit to the whole word, instead of a character used in the regular expressions. Additionally, by restricting patterns to start with a word, we reduced the time of one experiment below 15 minutes.

Table 1. Examples of the phrase patterns and retrieved phrases

Phrase patterns	Examples of retrieved phrases
více ? * * * * $ (m/f) *more ? * * * * $*	**více** už ale prozradí XY (m/f) **více** už o tom naše zpravodajka XY (f)
telefonuje * * **zpravodaj** (m) **telefonuje** * * **zpravodajka** (f) *is phoning * * rapporteur*	**telefonuje** z Moskvy **zpravodaj** (m) **telefonuje** náš pařížský **zpravodaj** (m) **telefonuje** naše moskevská **zpravodajka** (f)
sleduje * * **zpravodaj** (m) **sleduje** * * **zpravodajka** (f) *monitors * * rapporteur*	**sleduje** náš **zpravodaj** (m) **sleduje** stálý londýnský **zpravodaj** (m) **sleduje** naše spolupracovnice **zpravodajka** (f)
pokračuje ? * * * * $ (m/f) *continues ? * * * * $*	**pokračuje** jablonecká zpravodajka XY (f) **pokračuje** redaktor XY (m)
potvrdil to ? * * * * $ (m) **potvrdila to** ? * * * * $ (f) *confirmed by ? * * * * $*	**potvrdil to** našemu zpravodajovi hlavní inženýr XY (m) **potvrdila to** její tisková mluvčí XY (f)

These patterns have been assembled on the basis of a previous experimental probe that we have performed to delimit key words and phrases preceding proper names of speakers. Each pattern consists of one or two key words (usually a verb of speaking [5], [6] and a designation of the speaker's role, i.e., *moderator*, *reporter*, *correspondent*, etc., and several defined wildcards. As the basis of the patterns we used valence patterns with the verbs of speaking [7], to deconcretize them we use combinations of the following wildcards: "?" for any word in the phrase pattern stream; "*" for zero or one word; "$" for a word starting with capital character. Example of the phrase pattern would be *"popsala českému rozhlasu ? * * * * $"* (*"described to Czech Radio? * * * * $"*). The number of asterisks in each wildcard was determined according to the results of the tested hypothesis. Overall, we used a maximum of six wildcards in one phrase.

Table 1 illustrates some examples of the patterns and their specific occurrences within the texts. The column Phrase Patterns shows examples of the patterns in Czech and their English translation; the disproportion of phrases in Czech and English is caused by the highly inflective character of Czech, which specifies the information in various lexemes or morphemes. The examples of retrieved phrases are not translated

to English as far as their meaning is not significant to the text; but they are used to provide illustration of the words variability represented by wildcards. All real names in the illustrational phrases were replaced by the signs XY for space considerations. We present only several phrase patterns with a limited amount of illustrative phrases. The symbols (f/m) in the Tables 1 and 2 are used to determine masculine or feminine gender reflected in a phrase.

3 Methodology

In Part 3, we describe in detail the complete scheme of the speakers' database extension in the following steps: i) finding all speakers´ names (name entities which are parts of speaker indicating phrase); ii) counting the speaker occurrences in the documents of interest (potentially the whole archive or some sub-period). The speakers of interest are chosen and added to the database (if not already included); iii) finding the utterances of important speakers, checking it and using it for the training of the speakers' models.

As long as the method is based on the automatic process and its successive manual correction, it is not necessary to evaluate the error rate of the method.

For this project, we need well-built acoustic and language models for automatic recognition. To improve the models, we use the specially written program NanoTrans for semi-manually rewriting the recordings from broadcasting; labelling the speakers and creating the speakers' database.

The term semi-manual rewriting is used for a combination of automatic and manual work: the automatically-generated transcriptions are corrected by people. We have 375 recordings (1 recording = 1 daily news broadcast) that store 160 hours of Czech Radio (from which there are approximately 152 hours of Czech speech, the rest being mainly Slovak, with small amounts of both Russian and English speech). These were recorded through the time period 1968–2005.

By rewriting these recordings we get a basic speaker database and labelled acoustic data to train some models to recognise speakers. We cannot use the complete set of data from rewritten recordings to train an acoustic model for speakers because there were a lot of data from telephone, some parts were spoken over music, or the recordings were of very poor quality. Also, in the entire basic speaker database not all of the speakers are relevant and some are missing. For the time period selected for the transcription (1923–2014), there are hundreds of relevant speakers and not all of them are present in our semi-manually rewritten recordings.

Therefore we decided to find all possible speakers in the recordings, choose the most frequent ones, and then train the acoustic model for them.

To search the speakers' names, we have selected 196 introductory phrases that frequently co-occur with speaker names. Inasmuch as our system lacks the lemmatisation, it was necessary to supply all indispensable morphological variations of the key words in the phrases, i.e., morphological forms of selected verbs for the present and past tense, singular and plural forms of nomina agentis. Examples of the wildcards and their realisations are seen in Table 1.

As stated in Sec 2.1, we used phrase patterns for speakers´ name search, or the co-occurrence of a speaker´s name in a specific phrase in the text transcription connected with the speaker´s change in the audio recording.

As far as the amount of data in the period between 1923 and 1966 is negligible from the statistical point of view, we analysed the data starting in 1966 for our purposes.

Table 2. Speakers change

	Before phrase		Number of occurrences	**After phrase**	
	Average distance	Standard deviation		Average distance	Standard deviation
telefonuje * * zpravodaj (m) telefonuje * * zpravodajka (f) *is phoning * * rapporteur*	7.7 8.9	45.1 58.8	1877 416	32.2 46.0	27.1 42.9
telefonoval o tom (m) telefonovala o tom (f) *phoned about*	7.9 6.9	40.8 60.2	78 11	28.7 35.9	22.0 29.0
z ? * * se přihlásil * * zpravodaj (m) z ? * * se přihlásila * * zpravodajka (f) *from ? * * has entered * * rapporteur*	13.9 11.4	41.9 27.9	708 45	26.6 26.2	30.2 14.9
jak dodává * * reportérka (f) *how adds * * rapporteur*	20.6	58.4	198	59.3	40.6
jak dodává mluvčí (m/f) *how adds speaker*	15.7	48.2	43	43.0	27.6
blíže k tomu náš zpravodaj (m) *in more details on the topic our rapporteur*	7.8	25.0	153	21.7	14.1

We assumed that the broadcast news follow typical text patterns using stereotypical phrases when introducing each speaker who will be next to speak. We proceeded as follows: i) we searched the data; seeking the typical phrases connected with the new speaker's occurrence, ii) we created patterns combining particular words with variable positions that hypothetically combine with the occurrence of a speaker's name, iii) we found all the occurrences of these phrases with names in all transcriptions, iv) we chose names (actually they were proper names, we had to filter them) and added the important speakers' names to the speaker database.

Among the steps leading to the extension of the database we included the findings of acoustic data for the frequent speaker (we assume that we will find the data mostly for the reporters). We took a phrase, found it in all transcripts and counted the distance between this phrase and the change of the speaker. We assume that if the change of speaker is close, there are some phrases that indicate the name of the changed speaker.

Table 2 shows the distance (count in words) from the phrase to the change of the speaker. In the first column you can see the average distance from phrase to change of a speaker before the phrase, the fourth column shows the distance between the phrase and the change of a speaker after the phrase. The second and fifth columns give the standard

deviation. In the third column is the number of occurrences. You can also see that there are differences between occurrences of male and female speakers.

Table 3. Examples of patterns occurrence

	1966-9	**1970-4**	**1975-9**	**1980-4**	**1985-9**	**1990-4**	**1995-9**	**2000-4**	**2005-9**
u mikrofonu ? * * * * $	112	247	645	823	1010	369	431	178	201
u mikrofonu $	15	17	121	304	331	204	186	69	49
at the microphone									
sleduje * * zpravodajka	0	0	0	0	0	1	64	183	50
sleduje * * zpravodaj	0	1	0	6	22	40	194	379	246
sledovala * * zpravodajka	0	0	1	0	0	6	60	142	48
sledoval * * zpravodaj	6	1	13	25	36	48	146	164	105
sledovala $	0	0	7	18	17	97	280	175	10
sledoval $	6	4	24	47	217	297	542	176	44
monitors (monitored)									
více ? * * * * $	445	493	639	704	787	1377	3726	3191	1704
více * * zpravodajka	0	1	0	0	0	1	79	191	7
více * * zpravodaj	0	0	3	9	11	27	138	271	116
more									

From the results shown in Table 2 we decided to choose acoustic data for the speakers in several steps: i) find a phrase with a speaker's name, ii) count the distance between the phrase and the following speaker and decide if it is close enough, iii) extract the speaker name, iv) automatically split the audio recording and the transcript, and v) manually check to make sure the speaker was correctly identified.

Table 3 presents the examples of patterns for three key words incorporated in the appropriate phrases and the normalised number of their occurrences through the years in the time spacing of five years.

Figure 1 illustrates selected nomina agentis (the names of agent) occurrence in the time span from 1966 to 2009 in the time spacing of five years. From the total amount of 17 nomina agentis closely related to the speakers appearing in the radio broadcasting we have chosen the following ones: spolupracovník – associate (m), spolupracovnice – associate (f), zpravodaj – rapporteur (m), zpravodajka – rapporteur (f), redaktor – editor (m), redaktorka – editor (f). The graph shows not only the popularity of the names used across the period but also the male/female speaker occurrences in broadcasting.

4 Speakers

Based on the phrase patterns, we have found 60,172 potential speakers' names. There were not only names of actual speakers but also proper names (e.g. Brno, Radiožurnál - the names of Czech cities and the name of broadcasting stations). We downloaded the list of first names and surnames from Ministry of the Interior of the Czech Republic. By

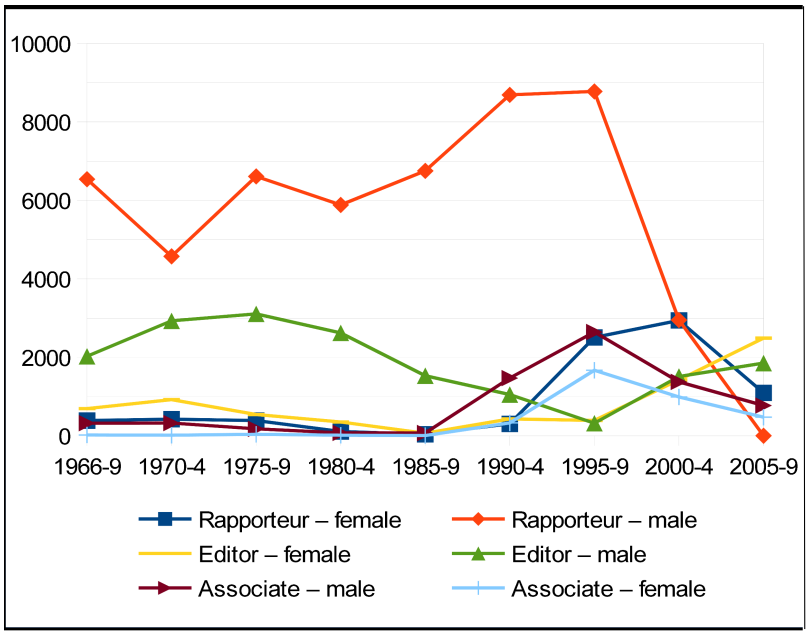

Fig. 1. Selected nomina agentis occurrence

comparing these lists with the found names, we get a resulting list of 20,313 speakers' names.

In this list were speakers about whom they were talking on the radio. We need only the list of speakers that speak on the radio. We know that almost all editors are among the important speakers and because they should be named in the broadcasting, we found how many times the name of the speaker occurred in the transcripts. There were several speaker names that occurred more than 500times (15 speakers' names). We decided to add all speaker names that occurred 10 or more times to the speaker database (1,478 speaker names).

5 Conclusions and Future Work

In this paper we investigated a methodology for automatic extension of a speaker database based on automatic transcription. We want to add the most frequent speakers to our speaker database; we assume that these speakers are publicly important persons (presidents, politicians, famous artists, etc.) and reporters or frequent guests of the radio – they are contemporary speakers. With the introduced tools and methodology we extended our speaker database by 1,478 most frequent speakers.

We prepared a methodology showing how to extract acoustic data necessary to train models for speaker recognition. In our future work we will apply it. The extraction is not fully automated, there will be necessary manual work (mostly with the decision and

confirmation that the data are chosen for the correct speaker). Hence we will try to lower the demands on human work (by automatic data clustering).

It has been observed that the performance of speaker recognition depends on the number of potential speakers. Therefore, we would like to investigate the benefit of restricting the potential set of speakers to the respective time period in which they could be active.

Acknowledgments. This work was supported by project no. DF11P01OVV013 provided by The Czech Ministry of culture in research program NAKI.

References

1. Nouza, J., at al.: Making Czech Historical Radio Archive Accessible and Searchable for Wide Public. Journal of Multimedia, vol. 7/2012, issue 2, pp. 159–169, Academy Publisher (2012)
2. Cerva, P., Silovsky, J., Zdansky, J., Nouza, J., Seps, L.: Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives. Speech Communication, Volume 55, Issue 10, pp. 1033-1046, (2013)
3. Seps, L.: NanoTrans — Editor for orthographic and phonetic transcriptions. Tel. and Signal Processing (TSP), 36th International Conference, pp.479–483, (2013)
4. Kucharova, M., Skodova, S., Seps, L., Labus, V., Nouza, J., Bohac, M.: On the quantitative and qualitative speech changes of the Czech radio broadcasts news within years 1969-2005. 16th International Conference on Text, Speech, and Dialogue, Czech Republic (2013)
5. Soltys, O.: Verba dicendi a metajazyková informace. Ústav pro jazyk český, Praha (1983)
6. Hirschova, M.: Česká verba dicendi v performativním užití: Příspěvek ke zkoumání komunikativních funkcí výpovědi. FF UPOL, Olomouc (1988)
7. Lopatkova, M., Zabokrtsky, Z., Kettnerova, V. Valenční slovník českých sloves. Praha, Karolinum (2008)