Dealing with Bilingualism in Automatic Transcription of Historical Archive of Czech Radio

Jan Nouza, Petr Cerva, and Jan Silovsky

Institute of Information Technology and Electronics, Technical University of Liberec
Studentska 2, 461 17 Liberec, Czech Republic
jan.nouza@tul.cz
https://www.ite.tul.cz/itee/

Abstract. One of the biggest challenges in the automatic transcription of the historical audio archive of Czech and Czechoslovak radio is bilingualism. Two closely related languages, Czech and Slovak, are mixed in many archive documents. Both were the official languages in former Czechoslovakia (1918-1992) and both were used in media. The two languages are considered similar, although they differ in more than 75 % of their lexical inventories, which complicates automatic speech-to-text conversion. In this paper, we present and objectively measure the difference between the two languages. After that we propose a method suitable for automatic identification of two acoustically and lexically similar languages. It is based on employing 2 size-optimized parallel lexicons and language models. On large test data, we show that the 2 languages can be distinguished with almost 99 % accuracy. Moreover, the language identification module can be easily incorporated into a 2-pass decoding scheme with almost negligible additional computation costs. The proposed method has been employed in the project aimed at the disclosure of Czech and Czechoslovak oral cultural heritage.

Keywords: oral archives, automatic speech-to-text transcription, language identification.

1 Introduction

In 2011 we launched works on an ambitious project supported by the Czech Ministry of Culture whose aim is to disclose the historical audio archive of Czech Radio (and its predecessor Czechoslovak Radio) to researchers (historians, media experts, linguists, phoneticians) as well as to wide public [1]. The archive contains several hundreds of thousands of spoken documents (with the total duration exceeding 100.000 hours) and covers 90 years of public broadcasting in Czechoslovakia and Czechia. During the last decade, the archive records have been digitized and now, within the project, they are to be transcribed, indexed and made accessible for search and listening via a special web portal.

For the transcription task, we have been adapting and enhancing a large-vocabulary continuous speech recognition (LVCSR) system developed previously

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 238–246, 2013.

in our lab. During the first two years of the 4-year project, we have implemented most of the required functionalities and utilized the system to process, transcribe and index more than 75.000 documents broadcast since 1993 to present [2]. That period did not pose a particular challenge for our research as we could employ the existing system trained for contemporary Czech.

When moving backwards in time, the situation becomes more complicated. Many older spoken documents contain not only Czech but also Slovak, as both the languages were the official ones in former Czechoslovakia (which split into the Czech and Slovak republics in 1993) and both were used in broadcasting. For example, in news programs, they were arbitrarily mixed as each speaker used his or her own native tongue. This may occasionally happen also in recently broadcast Czech Radio programs if there is an interview with a Slovak person.

Several years ago, we have adapted our LVCSR system also to Slovak [3]. Hence, the remaining problem to be solved is how to recognize automatically which language is spoken. This task is known as language identification (LID). One of its classic techniques is based on statistical modeling of phoneme sequences, which vary from one language to another (phonotactic approach, [4]). In this paper, we propose and evaluate a method that performs better, particularly for languages that are acoustically and lexically similar. The method is applicable if we already have a LVCSR system with lexicons and language models for each of the languages. In that case, instead of just phoneme sequences, the method takes into account words and word sequences. Moreover, the method can be made fast as only a smaller part of the lexicon from each language is really needed, which is shown in the experimental part.

2 Related Work

Recently, there has been an intensive research towards multilingual and bilingual LVCSR systems. The main reason for their use and development is the benefit of sharing resources, namely the data needed for training acoustic models [5]. Bilingual systems have been designed and tested both for pairs of tongues from different language families, e.g. English and Tamil [6], or English and Mandarin [7], as well as for those closely related, e.g. Spanish and Valencian [8], or Croatian and Slovenian [9]. In the last mentioned paper, the authors focused also on the automatic language identification task, though only in a narrow domain of weather forecast reports. For these two similar Slavic languages, the LID scheme based on identifying language-specific words performed better than the classic phonotactic approach. The limitations of the phonotactics (even if complemented by some recently proposed improvements) was demonstrated in the LID system evaluation campaign organized by NIST in 2011 [15], where Czech and Slovak were reported among the most confusable language pairs.

3 Czech and Slovak as Related Languages

Czech and Slovak belong to the West-Slavic branch of European languages. They are considered very similar and closely related because in the past both were the official languages used within one state. Anyway, since 1993, when Czechoslovakia split, a new generation of young people has grown in the succession states who have difficulties to understand the language of the other nation. This indicates that the difference is larger than it was commonly thought.

3.1 Difference in Lexical Inventories

To quantify the degree of difference, we compared two lexicons used in the Czech and Slovak versions of our LVCSR system. For each language, we created a subset made of 100,000 most frequent words. The two subsets contained 25,325 items with the same orthography, from which 2,802 differed in pronunciation. It means that only 23 % of the lexical inventory is exactly same in the two languages. In [3] we arrived at the same figure (also 23 %) by comparing several parallel corpora - EU documents published in Czech and Slovak.

Despite this 77 % difference in the lexicons, the perceptual level of dissimilarity will look not that high if we perform a more detailed comparison of corresponding word pairs. Many differ only in one or two letters, often in prefixes and suffixes.

3.2 Difference in Morphology

Czech and Slovak are languages with rich morphology and a high degree of inflection. There exist several thousands of words in the two languages that have the same lemma but differ in some inflected forms. Where Czech nouns take suffix -*em*, the Slovak ones would use -*om*, where Czech adjective use suffix -*ém*, the other language take -*om*, etc. In Table 1, we give several examples of these related morphological patterns. This phenomenon can be utilized, if we need to transfer a list of words (proper names, particularly) from one lexicon to the other.

Table 1. Examples of some regular differences between suffixes used in Czech and Slovak (demonstrated on proper name 'Barack' and adjective 'political')

Word type	**Czech [CZ]**	**Slovak [SK]**
Proper names	Barack*em* Barack*ův* Barack*ova*	Barack*om* Barack*ov* Barack*ovho*
Adjectives	politick*ém* politick*é* politick*ou*	politick*om* politick*ej* politick*ú*

3.3 Difference in Phonetics

In LVCSR systems, Czech phonetic inventory is usually composed of 41 basic phonemes, while the Slovak one uses 48. For their SAMPA symbols, see Table 2.

In [3] we have shown that for initial experiments with Slovak speech recognition and also for bootstrapping a Slovak acoustic model (AM) trained on speech records that are not phonetically annotated, we can map the Slovak specific phonemes on the closest Czech ones. The mapping proposed in [3] is useful also for a bilingual Czecho-Slovak LVCSR system as it can operate (if desired) with one phonetic inventory, which allows for fast and efficient switching between Czech, Slovak and Czecho-Slovak AMs.

Table 2. Czech and Slovak phonemes represented by their SAMPA symbols (language specific ones are printed in bold)

Groups	**Czech [CZ]**	**Slovak [SK]**
Vowels	a, e, i, o, u, a:, e:, i:, o:, u:, **@** (schwa)	a, e, i, o, u, a:, e:, i:, o:, u:, **{**
Consonants	p, b, t, d, c, J\, k, g, ts, dz, tS, dZ, r, l, **Q**, **P** f, v, s, z, S, Z, X, j, h\, m, n, N, J, F	p, b, t, d, c, J\, k, g, ts, dz, tS, dZ, r, l, **r=**, **r=:**, **l=**, **l=:**, **L** f, v, s, z, S, Z, X, j, h\, **w**, **U_^**, **G**, **I_^** m, n, N, J, F

4 Speech Transcription System

The LVCSR system used for the transcription of archive documents employs a two-pass strategy. The output of the first decoder pass is used for a) segmentation to speech and non-speech parts, b) synchronization of speaker change detector with word and noise boundaries [10], c) speaker clustering [11], and d) speaker adaptation via the CMLLR technique [12]. The first pass is usually performed with a smaller lexicon to reduce computational costs and time. In the second pass, the decoder processes the already separated segments, uses the adapted acoustic model, and utilizes the full lexicon with the corresponding language model.

The acoustic front-end takes the archive data and converts them into 16 kHz, 16 bit, PCM WAV format. A signal is parameterized into a stream of 39 mel-frequency cepstral coefficient (MFCC) feature vectors computed every 10 ms in 25-ms-long frames. Using a 2-second long moving window, the MFCC features are normalized by the cepstral mean subtraction (CMS) technique. The final step is the HLDA (Heteroscedastic Linear Discriminant Analysis) transform performed by multiplying each feature vector by a 39 x 39 HLDA matrix determined during the acoustic model training procedure. These features are employed in all the following modules and in both the passes.

The acoustic model is a triphone-based one covering 41 phonemes and 7 types of noise. The Czech AM has been trained on 320 hours of (mainly broadcast) data. The amount of speech available for training the Slovak AM was smaller, 107 hours. In the experiments described in section 5, we used also a Czecho-Slovak (CZ+SK) AM trained on 120 hours of Czech and 107 hours of Slovak.

The lexicon for contemporary Czech contains 551K words. For Slovak, the lexicon is smaller (due to smaller text corpora), its size is 303K words. The language models are based on bigrams. However, as both the lexicons contain several thousands multi-word expressions (frequently collocated word strings), a significant part of bigrams covers sequences that are three-, four-, five- or even six-word long. This feature helps to improve the recognition rate by 2 %. The unseen bigrams are backed-off by the Kneser-Ney smoothing technique [13].

5 LID Scheme for Czech and Slovak

In the situation, when we need to distinguish between two closely related languages, for which we already have lexicons and LMs, the most reliable approach to language identification is to employ the existing LVCSR system. The scheme can be efficiently incorporated within the first pass.

5.1 LVCSR with Merged Lexicons and Language Models

The LID module has been designed in the following way: A Czecho-Slovak (CZ+SK) lexicon is created by merging L most frequent Czech words with the same number of words from the Slovak lexicon. (The total size of the merged lexicon is $2L$.) Each word gets a label saying whether the word is Czech or Slovak. Using the available text corpora we compute word-pairs counts separately for the Czech part of the merged lexicon and for the Slovak one. Before merging the two word-pair lists, we label their items in the same way as in the lexicon. (The only common item in the two lists is the $START$ symbol used for the beginning of an utterance.) After that, the CZ+SK LM is computed using the standard Kneser-Ney smoothing technique, which assigns some small probabilities also to transitions between Czech and Slovak words.

Now, the LID task can become a part of the first pass, without influencing the other goals required on that level. The only modification is that the LVCSR runs with the CZ+SK acoustic model and CZ+SK language model. The output of the recognizer contains words with either Czech (CZ) or Slovak (SK) labels. A special label (COM) is assigned to those Czech and Slovak words that share the same orthography and pronunciation. For each speech segment (determined by the speaker change point detector [10]), we get the numbers of recognized Czech words (N_{CZ}), Slovak words (N_{SK}) and the common ones (N_{COM}).The utterance in the segment is identified as Czech or Slovak according to the higher of counts N_{CZ} and N_{SK}.

The performance of the proposed scheme is illustrated in Table 3. It shows a transcription of an initial part of evening news where two speakers, a Czech and a Slovak one, talk. In the first row, there is the manual transcript that can be used for comparison. The second row shows the languages used. The third and fourth rows indicate the recognized words (from a joint CZ+SK lexicon) and their labels. In the last row, there is information from the speaker change detector that provides boundaries for speaker and language identification modules. We can see that the utterance of the first speaker included four words that are common to both the languages (i.e. N_{Com}=4). Anyway, it was identified as Czech because N_{CZ}=1 and N_{SK}=0. The second fragment would be labeled as Slovak since the Slovak specific words prevail.

Table 3. Illustration of the LID scheme performance on evening news. (English translation is: In Czech: Good evening, we broadcast radio news. In Slovak: We welcome listeners ...)

Manual	Dobrý	večer	vysíláme	rozhlasové	noviny	Pri	počúvaní	vítáme	poslucháčov
Lang.	Czech					Slovak			
Auto	dobrý	večer	vysíláme	rozhlasové	noviny	pri	počúvaní	vítané	poslucháčov
Label	COM	COM	CZ	COM	COM	SK	SK	COM	SK
Speaker	1	1	1	1	1	2	2	2	2

5.2 Experimental Evaluation

The proposed LID scheme has been evaluated on a test set that included 1000 Czech and 1000 Slovak speech segments. Their total duration was 228 minutes (31,214 words). In average, each utterance was 6.8 seconds long and contained 15.6 words. (The minimum length was 6 words).

Table 4. Language identification error and Real Time factor as function of lexicon size

Lexicon size **[L words in each language]**	**LID error rate** **[%]**	**RT** **factor**
L = 1,000	8.75	0.43
L = 5,000	3.03	0.51
L = 10,000	2.02	0.53
L = 20,000	1.51	0.56
L = 30,000	1.31	0.60
L = 40,000	1.31	0.64
L = 50,000	1.15	0.69

We conducted several experiments with different size of the merged lexicons. The results are summarized in Table 4. From the above results we can see, that if L is chosen 20,000 or higher, the two languages are identified with an acceptable error rate smaller than 1.6 % and the time required for processing and decision is slightly above one half (0.56) of the signal duration. The same lexicon size (20,000 words) was found sufficient also for the unsupervised speaker adaptation scheme proposed in [12]. This means that the inclusion of the LID module to the complete transcription system requires almost negligible additional computation costs.

We have also run a complementary series of experiments in which we investigated the influence of the *acoustic model* on the LID results. We compared the performance of the above described Czecho-Slovak (CZ+SK) AM with those trained only on Czech or Slovak speech data. From the diagram in Fig. 1 we can see, that the impact of the acoustic models becomes significantly smaller when the lexicon size increases. This confirms the strength of the information provided by the lexicon and LM in the task of discrimination between closely related languages.

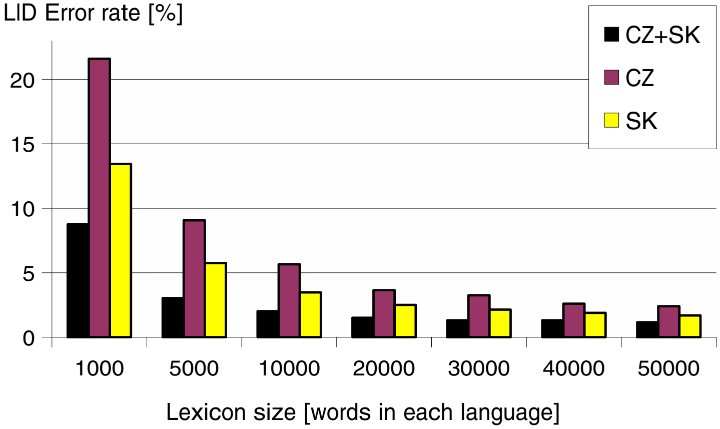

Fig. 1. Comparison of LID performance for increasing lexicons and 3 acoustic models

6 Evaluation of Complete Bilingual Transcription System

Recently, the proposed LID method has been tested on a set of archive documents representing the last eight years of Czech Radio broadcasting (2005-2012). We chose this particular period from which we have official transcripts provided by a media monitoring company for some relevant radio programs, like evening news or political talk-shows. This allows us to compare the human-made transcriptions with those achieved by our system in a fully automated way. For experimental evaluation we selected 24 complete news (3 from each year). The duration of each was 30 minutes.

The LID module worked with 40,000-word CZ+SK lexicon (i.e. L=20,000). It identified 58 segments spoken in Slovak whose total duration was 15.2 minutes (approx. 2 % of all audio data). That language was used by Slovak correspondents of the Czech Radio and by Slovak politicians who had been interviewed. One of the 58 found segments was wrongly labeled as Slovak. It was a short Czech utterance that contained Slovak proper names. From the same reason also an opposite error occurred when one Czech sentence was identified as Slovak. These results as well as those presented in section 5 seem much better the results reported for Czech and Slovak in [15].

The successful separation of the two languages in the first pass allowed for running the second pass with the proper (Czech or Slovak) full-size lexicon and the corresponding language model. The transcription accuracy of the Czech part of the news was 89.1 %. For the Slovak segments, it was lower, only 83.6 %. There were several reasons for the worse latter result: a) most Slovak utterances were recorded out of studio, b) they had a character of either planned or spontaneous (but not read) speech, and c) the Slovak lexicon is smaller (303K words) when compared to the Czech one (551K words) - due to a much smaller amount of text resources available for Slovak.

7 Conclusion and Future Work

In this paper, we present a LID method suitable for distinguishing between languages that are acoustically and lexically similar. The method utilizes a large-vocabulary speech recognition system operating with the merged lexicon and language model composed from size-optimized lexicons and LMs of the individual languages. Unlike the classic phonotactic technique, this approach takes into account real words and their N-gram probabilities and hence it provides a better discriminative strength. Moreover, the proposed LID module can be incorporated into a two-pass decoding scheme with minimum additional computation costs.

So far, the method has been tested on contemporary spoken documents from the Czech radio archive. It identified almost all utterances spoken in Slovak language and allowed for automatic switching between two language specific speech recognition modules.

Recently, we prepare its application also to the historical part of the archive, which is the main goal of the project. Especially before 1993 (i.e. in times of former Czechoslovakia) the Slovak language will occur more frequently. Before doing it, we have to adapt the lexicons so that they better fit speech of previous historical epochs. For Czech, it has been already done [14].

Acknowledgments. The research was supported by the Czech Ministry of Culture -project no. DF11P01OVV013 in program NAKI.

References

1. Nouza, J., Blavka, K., Bohac, M., Cerva, P., Zdansky, J., Silovsky, J., Prazak, J.: Voice Technology to Enable Sophisticated Access to Historical Audio Archive of the Czech Radio. In: Grana, C., Cucchiara, R. (eds.) MM4CH 2011. CCIS, vol. 247, pp. 27–38. Springer, Heidelberg (2012)
2. Nouza, J., Blavka, K., Zdansky, J., Cerva, P., Silovsky, J., Bohac, M., Chaloupka, J., Kucharova, M., Seps, L.: Large-scale processing, indexing and search system for Czech audio-visual cultural heritage archives. In: IEEE 14th International Workshop on Multimedia Signal Processing (MMSP), pp. 337–342 (2012)
3. Nouza, J., Silovsky, J., Zdansky, J., Cerva, P., Kroul, M., Chaloupka, J.: Czech-to-Slovak Adapted Broadcast News Transcription System. In: Proc. of Interspeech 2008, Australia, pp. 2683–2686 (2008)

4. Navratil, J., Zuhlke, W.: An efficient phonotactic-acoustic system for language identification. In: Proc. of ICASSP, Seattle, USA, vol. 2, pp. 781–784 (1998)
5. Uebler, U.: Multilingual speech recognition in seven languages. Speech Communication 35(1-2), 53–69 (2001)
6. Kumar, C.S., Wei, F.S.: A Bilingual Speech Recognition system for English and Tamil. In: Proc. of ICICS PCM, pp. 1641–1644 (2003)
7. Zhang, Q., Pan, J., Yan, Y.: Mandarin-English bilingual speech recognition for real world music retrieval. In: Proc. of ICASSP, Las Vegas, USA, pp. 4253–4256 (2008)
8. Alabau, V., Martinez, C.D.: A Bilingual Speech Recognition in Two Phonetically Similar Languages. Jordanas en Tecnologia del Habla, Zaragoza, pp. 197–202 (2006)
9. Zibert, J., Martincic-Ipsic, S., Ipsic, I., Mihelic, F.: Bilingual Speech Recognition of Slovenian and Croatian Weather Forecasts. In: Proc. of EURASIP Conf. on Video/Image Processing and Multimedia Communications, Zagreb, Croatia, pp. 957–960 (2000)
10. Silovsky, J., Zdansky, J., Nouza, J., Cerva, P., Prazak, J.: Incorporation of the ASR output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams. In: Proc. of IEEE workshop on Multimedia Signal Processing (MMSP), Banff, Canada, pp. 118–123 (2012)
11. Silovsky, J., Prazak, J.: Speaker Diarization of Broadcast Streams using Two-stage Clustering based on I-vectors and Cosine Distance Scoring. In: Proc. of ICASSP, Kyoto, pp. 4193–4196 (2012)
12. Cerva, P., Palecek, K., Silovsky, J., Nouza, J.: Using Unsupervised Feature-Based Speaker Adaptation for Improved Transcription of Spoken Archives. In: Proc. of Interspeech 2011, Florence, pp. 2565–2568 (2011)
13. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Detroit, pp. 181–184 (1995)
14. Chaloupka, J., Nouza, J., Kucharova, M.: Using Various Types of Multimedia Resources to Train System for Automatic Transcription of Czech Historical Oral Archives. In: Petrosino, A., Maddalena, L., Pala, P. (eds.) ICIAP 2013 Workshop. LNCS, vol. 8158, pp. 228–237. Springer, Heidelberg (2013)
15. Brümmer, N., et al.: Description and analysis of the Brno276 system for LRE2011. In: Proc. of Speaker Odyssey Workshop, Singapur, pp. 216–223 (2012)