Adapting Slovak ASR for native Germans speaking Slovak

Štefan Beňuš [1,2], Miloš Cerňak[1], Milan Rusko[1], Marián Trnka[1], Sachia Darjaa[1]
[1] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Constantine the Philosopher University, Nitra, Slovakia

sbenus@ukf.sk, {Milos.Cernak, Milan.Rusko, Marian.Trnka, Sachia.Darzagin}@savba.sk

Abstract

We explore variability involved in speech with a non-native accent. We first employ a combination of knowledge-based and data-driven approaches for the analysis of pronunciation variants between L1 (German) and target L2 (Slovak). Knowledge gained in this two-step process is then used in adapting acoustic models and the lexicon. We focus on modifications in the pronunciation dictionary and speech rate. Our results show that the recognition of German-accented Slovak is significantly improved with techniques modeling slow L2 speech, and that the adaptation of the pronunciation dictionary yields only insignificant gains.

1 Introduction

Automatic recognition of non-native accented speech represents a complex problem, especially since this type of variability becomes more common even in languages with a relatively small number of speakers due to globalization and increased mobility of people. The methods most commonly used for dealing with this type of speech variability include pronunciation modeling, acoustic modeling, or topological modeling (Oh, Yoon and Kim, 2007, Tomokiyo, 2000). This paper presents an approach that starts with an analysis of the pronunciation variability of nonnative speech taking into account most salient differences between L1 language (in our case German) and L2 target language (Slovak). Following this knowledge-base step, a semi-automatic data-driven approach analyzes the pronunciation variants on a subset of a training corpus is proposed. The knowledge gained in this two-step process is then used to adapt our state-of-the-art ASR system for Slovak in an effort to improve the baseline recognition of this system in German accented Slovak. We primarily experiment with adapting the pronunciation dictionary and speech rate. In short, we test the acoustic model and lexicon adaptation based on the analysis of pronunciation proximity between the German-accented and standard varieties of Slovak.

The paper is structured as follows. Section 2 describes the corpora used for testing and training. Section 3 discusses differences between Slovak and German pronunciation by analyzing the phonological systems of the two languages (3.1) and by analyzing the errors Germans make when speaking Slovak (3.2). Section 4 presents the setup and results of experiments in adapting our state-of-the-art ASR system for Slovak to German-accented pronunciation of Slovak focusing on speech rate manipulation and appending pronunciation dictionary. Section 5 discusses the findings and concludes the paper.

2 Description of the databases

Our testing corpus consists of Slovak sentences read by 18 native speakers of German. The sentences were selected or created to represent four types of variability: dialectological (100), foreign accent (100), phonetic richness and balance (300), and prosody (90). The first type was based on common differences among Slovak dialects, the second specially designed for problematic areas of native German speakers speaking Slovak.

Depending on the L2 proficiency level of the subjects, they were divided into two groups: Beginner – Intermediate (A1-B1), and Upper-intermediate – Advanced (B2-C2). The subjects were evenly distributed into these two groups with 9 speakers each. The first group read sentences for the dialectological and accent tests accompanied by 100 phonetically rich and balance sentences, and the second group read all 590 sentences. In total, the testing corpus represents 8010 sentences (9*300 + 9*590).

3 Features of Slovak with German accent

3.1 Knowledge-based approach

One of the most common ways of predicting differences between native (L1) and foreign-accented (L2) speech is to compare the sound systems of L1 and L2. Here we present a brief overview of most robust pronunciation differences between German and Slovak.

In terms of segmental inventories, Slovak does not have front rounded vowels and has only one front mid vowel quality while German has two. Also, both languages have phonemically distinct short and long vowels, but the length distinction in German robustly affects vowel quality (short vowels being lax and more centralized), while this tendency for Slovak is much less salient and a mid central schwa is missing in the Slovak inventory (Beňuš and Mády 2010). Additionally, a major difference comes from Slovak palatal consonants (stops, nasal, and lateral) that are missing in German. Finally, /r/ is an apical trill in Slovak in all positions while it commonly has uvular or vocalized qualities in German.

Many allophonic processes are different in the two languages. The most perceptually salient include the aspiration of voiceless stops and the glottalization of initial vowels in German and its absence in Slovak. German lacks a so called dark /l/ quality in syllable codas while most /l/s in Slovak have this quality. In terms of phonotactics, Slovak has a richer set of potential onset clusters than German. Additionally, Slovak syllabic nuclei might be formed by liquids (/l/, /r/) that also participate in lengthening alternations, which is not the case in German. While both languages have pervasive voicing assimilation and neutralization, voicing neutralization in obstruent coda consonants is slightly more salient in German than in Slovak.

Finally, most salient prosodic differences include a fixed left-most word stress in Slovak (cf. variable in German). Slovak in general also reduces the length and quality of unstressed vowels minimally, while in German, unstressed vowels tend to be shortened and centralized.

3.2 Analysis of accent sentences

In this section we test the theoretical predictions of pronunciation problems in Slovak with German accent stemming from interferences between L1 and L2 described in the previous section. We took a subset of our corpus, 100 accent sentences read by all 18 speakers and asked trained annotators to mark all perceptually salient markers of accented speech at the level of segments together with word stress differences. Different annotators (N=6) were given identical instructions and labeled different subsets of the data. A single expert then checked all annotations for mistakes and inconsistencies.

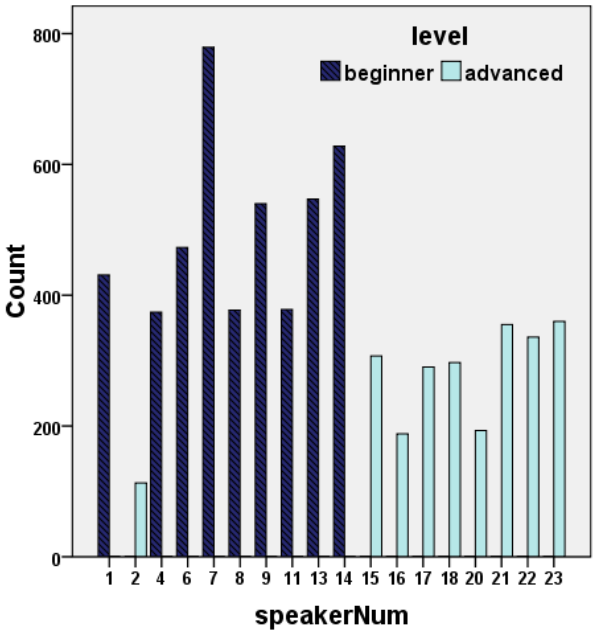

Figure 1. Error counts for all subjects divided by their L2 proficiency level (there were 2540 reference phonemes for each speaker)

The annotators found 6966 segmental differences between 'standard' and German accented Slovak, which represents 15.2% of all 45720 phonemes in the 1800 accent sentences. Roughly half of the differences involved syllable

nuclei including liquids (53.1%) and the rest involved onset and coda consonants. The assignment to proficiency levels showed a fairly reasonable correspondence with the number of segmental problems in the accent sentences, as can be seen in Figure 1 above.

Given the discussion in Section 3.1, we noticed several expected and unexpected patterns in the distribution of pronunciation deviations. Table 1 below lists the most frequent groups of pronunciation problems. The expected problems involved differences in the palatalization of alveolar consonants (15.6%), and the presence of aspiration with voiceless plosives (3.3%). Two notable unexpected patterns were observed. First, despite some differences in the short and long syllabic nuclei, described in 3.1, the overall frequency of deviations in phonemic length was surprising: almost one third (31.6%) of all marked differences involved either the shortening of long nuclei or lengthening of short ones. Additionally, despite the clear and predictable placement of Slovak word stress, 13.7% of differences involved an incorrect placement of word stress. The production of German vowel quality (such as front rounded vowels or schwa) was relatively low (1.8%). Hence, prosodic and metrical features of vowels were perceived as far more problematic than the features related to their quality.

Type of error	Count	%
Vowel shortening	1164	16.7
Palatalization	1090	15.6
Obstruent voicing	1078	15.5
Vowel lengthening	1038	14.9
Nucleus stress	954	13.7
Rhotic	537	7.7
Aspiration	227	3.3
German vow. quality	123	1.8

Table 1: Most common errors in accent sentences

The second unexpected pattern was a relatively high frequency of differences in the voicing of obstruent consonants (15.5%). The majority of these cases included the devoicing of consonants that, in regular fluent Slovak, would be produced as voiced. This pattern is related to pervasive coda voicing neutralization in German mentioned in section 3.1. Voicing of canonically voiceless consonants was observed as well, especially in the voicing of /s/ to /z/.

It is worth noting that both of the unexpected patterns relate to speech rate. A generally slower rate of L2 speakers results in frequent pauses between words thus creating an environment that meets the description for obstruent devoicing in German and prevents across-the-word voice assimilation that is pervasive in Slovak. Additionally, the presence of these pauses facilitates so called pre-boundary lengthening (e.g. Delattre, 1968 for German), in which the rime of the pre-pausal syllable is elongated. Finally, a generally slower rate may result in vowels intended as short to be perceived as long especially in the speech that is slowed down locally (for example with unknown words for L2 speakers).

4 ASR experiment

The analysis of accent sentences in the previous section revealed a potential impact of slower speaking rate of L2 speakers on the frequency of pronunciation deviations. We test the effects of speaking rate and variability in the pronunciation dictionary on the recognition of German accented Slovak in the following experiment.

4.1 Test setup

The training audio database contained 130 hours of phonetically rich sentences, gender balanced, from domains such as news and articles from various magazines, recorded from 140 speakers with Sennheiser ME3 headset microphone with Sennheiser MZA 900 P in-line preamplifier and EMU Tracker Pre USB audio interface. Database was annotated using the Transcriber annotation tool (Barras et al., 2000), twice checked and corrected. Recordings were split on segments if possible not bigger than 10 sec.

The training text corpora contained a total of about 92 million sentences with 1.25 billion Slovak words. A general-domain trigram language model (LM) was created with a vocabulary size of 350k unique words (400k pronunciation variants) which passed the spell-check lexicon and subsequently were also checked manually. Similarly to other recognizers in Slovak (Staš, Hládek and Juhár, 2010) the modified Kneser-Ney algorithm was used as a smoothing technique. The general LM

was adapted with all 590 sentences from the target domain.

The Julius decoder (Lee, Kawahara, and Shikano, 2001) was used as a reference speech recognition engine, and the HTK toolkit was used for word-internal acoustic models (AMs) training. We trained AMs using the triphone mapping as described in (Darjaa et al., 2011), with 32 Gaussian densities per each HMM state.

Experiments have been performed using AMs and LM trained from the training databases, and the 8010 sentences from the testing corpus as described in Section 2.

4.2 Results

To estimate the potential effect of slow L2 speech on the recognition accuracy, we first performed signal level acceleration directly on the recorded waveforms. The Praat speech analysis system (Boersma and Weenink 2011) was used, particularly its functionality of adjusting the time-domain of a sound file with a fixed conversion factor used in subsequent PSOLA resynthesis of the resulting file. We resynthesized all test sentences using the factors 0.9, 0.7, and 0.5 (the last one corresponding to 50% duration of the original file) and performed recognition with an unadapted LM that had the baseline WER of 55%. The results showed that the acceleration factor with the highest accuracy gain was 0.7, which improved the baseline to 43.4% WER. Factor 0.9 lowered WER to 49.5% while factor 0.5 showed the worst result (54.1% WER).

Following this encouraging initial result, feature level acceleration was performed by simple change of frame shift in the ASR front-end. The original features were calculated from 25 ms frame durations and a 10 ms frame shift. While keeping the frame durations constant, we increased the frame shift to 14 ms. This corresponds to the acceleration factor of 0.714, approximately identical to the best performing factor in the signal modulation experiments.

Table 2 shows achieved recognition results based on the adapted LM used as the baseline. This refers to the performance of the system on German accent sentences without any rate modifications. Unfortunately, we don't have a corpus of these sentences produced by Slovak speakers to provide a system baseline for non-accented speech but in a similar, albeit larger, corpus of 18 speakers reading 380 sentences this system's WER was 21.3% (Beňuš et al., 2011).

Speaker rate was accelerated at the signal and feature levels. We see that both signal and feature adaptation of speech rate significantly improved the accuracy of recognition with the latter outperforming the former. The extent of the improvement is rather surprising and suggests that speech rate in read sentences is a major factor when recognizing German-accented Slovak.

Test	**WER %**
Baseline	40.58
Alternate dictionary	40.48
Signal-adapted speech rate	28.67
Signal-adapted rate+alt. dictionary	28.13
Feature-adapted speech rate	25.79
Feature-adapted rate+alt. dictionary	25.33

Table 2: Word error rates (WER) for signal and feature adaptations (speech rate accelerations).

The analysis in section 3 also identified two common patterns: devoicing of consonants of German speakers that, in regular fluent Slovak, would be produced as voiced, and vowel shortening of German speakers. We tried to use this knowledge for improving the speech recognition system. In order to better match the pronunciation of German speakers in Slovak ASR system, we added alternative pronunciations to each entry of Slovak dictionary according to Table 3. For example, the entry 'Aachene' with pronunciation /a: x e J e/, was extended with an alternative pronunciation /a x e n e/ by the application of the rules in the 1st and 4th rows.

Original phones	**Phones used in alternative pronunciations**
/J/, /n/	/n/
/c/, /t/	/t/
/J\/, /d/	/d/
/a:/ /e:/ /i:/ /o:/ /u:/	/a/, /e/, /i/, /o/, /u/

Table 3: Rules for generation of alternative pronunciations (/J/, /c/, /J\/ are Slovak SAMPA symbols for palatal variants of /n/, /t/, and /d/ respectively).

The results in Table 2 show that the changes to the dictionary resulted in only insignificant improvements on top of the rate adjustment.

Finally, we compared the average WER for individual speakers in the baseline system with the adapted systems. For 17 out of 18 speakers the improvement was greater than 5% and ranged up to 34%; only one speaker's results showed deterioration (2%). Interestingly, despite a relatively good correspondence between the proficiency level and the number of pronunciation errors showed in Figure 1, neither the recognition accuracy of the adapted model, nor the extent of improvement after feature adaptation, showed a consistent relationship to the perceived proficiency of our subjects. This may be due to the greater number and complexity of test sentences used for advanced speakers compared to the beginners.

5 Discussion and conclusion

Our results showed that adjusting the rate of non-native speech to resemble the rate of the native training corpus significantly improves the recognition of speech with foreign accent. Moreover, we showed that feature-based acceleration outperforms signal-based acceleration. This is important since feature-based acceleration is much easier to perform, and an ASR system runs faster as it processes less frames. Furthermore, it is plausible that speech rate variability will be similar in non-native accents of multiple L1 languages, which cannot be expected for the pronunciation variants. Hence, although the acceleration of the signal or features does not account for all of the phonetic interference phenomena described in Section 3.2, sophisticated speech rate modeling that includes the combination of phone rate, syllable rate, and word rate promises to provide a robust technique for dealing with variability stemming from non-native accents.

6 Acknowledgments

This work was supported by the European Project of Structural Funds, ITMS: 26240220064.

References

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. 2000. Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication, 33 (1–2).

Beňuš, Š., Cerňak, M., Rusko, M., Trnka, M., Darjaa, S., and Sabo, R. 2011. Semi-automatic approach to ASR errors categorization in multi-speaker corpora. *Proceedings of the International Conference Slovko*.

Beňuš, Š, and Mády, K. 2010. Effects of lexical stress and speech rate on the quantity and quality of Slovak vowels. *Proceedings of the 5th International Conference on Speech Prosody*.

Boersma, P., and Weenink, D. 2011. Praat: doing phonetics by computer [Computer program, http://www.praat.org/].

Darjaa, S., Cerňak, M., Trnka, M., Rusko, M., Sabo, R. 2011. Effective Triphone Mapping for Acoustic Modeling in Speech Recognition. *Proceedings of Interspeech 2011 Conference.*

Delattre, P. 1968. A Comparison of Syllable Length Conditioning Among Languages. *International Review of Applied Linguistics*, IV:183-198.

Gusfield, D.. 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK.

Lee, A., Kawahara, T., and Shikano, K. 2001. Julius – an Open Source Real-Time Large Vocabulary Recognition Engine. In *Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH).*

Oh, Y.R., Yoon, J.S., Kim, H.K. 2007. Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. Speech Communication 49(1), 59-70.

Staš, J., Hládek, D., Juhár, J. 2010. Language Model Adaptation for Slovak LVCSR. In *Proc. of the Intl. Conference on AEI*, pp. 101–106.

Tomokiyo, L.M., 2000. Lexical and acoustic modeling of non-native speech in LVCSR. In: Proc. ICSLP, Beijing, China, pp. 346–349.