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

RESEARCH Open Access

A cross-lingual adaptation approach for rapid development of speech recognizers for learning disabled users

Marek Bohac[1*†], Michaela Kucharova[1†], Zoraida Callejas[2†], Jan Nouza[1†] and Petr Červa[1†]

Abstract

Building a voice-operated system for learning disabled users is a difficult task that requires a considerable amount of time and effort. Due to the wide spectrum of disabilities and their different related phonopathies, most approaches available are targeted to a specific pathology. This may improve their accuracy for some users, but makes them unsuitable for others. In this paper, we present a cross-lingual approach to adapt a general-purpose modular speech recognizer for learning disabled people. The main advantage of this approach is that it allows rapid and cost-effective development by taking the already built speech recognition engine and its modules, and utilizing existing resources for standard speech in different languages for the recognition of the users' atypical voices. Although the recognizers built with the proposed technique obtain lower accuracy rates than those trained for specific pathologies, they can be used by a wide population and developed more rapidly, which makes it possible to design various types of speech-based applications accessible to learning disabled users.

Keywords: Automatic speech recognition; Cross-lingual adaptation; Assistive technology; Speech technology; Atypical voices; Learning disabled; Intellectual disability; Dysarthria

1 Introduction

Millions of individuals suffer from learning disabilities that also affect their speech production. These conditions result in atypical voices that are very difficult to understand even for human listeners, as they may affect one or more of the major language subsystems, including phonology, morphology, syntax and semantics. Focusing on phonology, impaired speech can affect voice timing, pitch, volume, fluency and articulation [1].

Different studies have focused on the nature of such mispronunciations and their impact in intelligibility. For example, [2] shows that impaired speakers have a good control of tone but a diminished discrimination between stressed and unstressed vowels, as well as abnormal production of extremely long or short vowels. In [3], the authors focus on how to measure the intelligibility of atypical voices objectively along different perceptual dimensions.

Speech technology enhances the functional and affective experience of technology for many user groups, including people with reading difficulties, hearing and visually impaired, older adults, and people with learning disabilities [4]. One of the main applications of speech technology is voice therapy. For example, [5] presents the PreLingua tool, which aims to train skills such as intensity, tone, vocal onset, phonation time and vocalization.

However, when the phonological disorder is severe, it can be useful to complement the voice therapy with other applications for augmentative and alternative communication. For example, VIVOCA [6] is a voice-input voice-output augmentative communication aid for people with severe impairment. Joode et al. present a detailed study on assistive technologies for people with cognitive deficits including different uses of speech technology [7], and Lancioni et al. provide a review of speech generating devices for augmented communication [8]. Recently, there was a special issue on speech and language processing as assistive technologies [9], which shows the potential interest of this area.

*Correspondence: marek.bohac@tul.cz
†Equal contributors
[1]Institute of Information Technology and Electronics, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic
Full list of author information is available at the end of the article

Despite their high potential to help users, these technologies usually address specific disorders. For example, [10] studied the best configuration of parametrization, feature selection and classification techniques for the recognition of stuttered events, while [3] studied different measures of vocal quality, articulation, nasality and prosody of spastic dysarthria. This means that most of the systems have been tailored to specific population groups, which makes them more effective for those users but not so adequate for people suffering from other related disorders.

In this paper, we present an approach to develop speech recognizers for learning disabled users aimed at a general population. Also, we were particularly interested in defining a procedure which is rapid and cost-effective. The existence of affordable assistive technology is an effective mean to ensure full and equal enjoyment of all human rights and fundamental freedoms [11]. To make the development fast and efficient, and to speed up the transition from an experimental tool to a professional program, we decided to employ a modular automatic speech recognition (ASR) system designed at the Technical University of Liberec during the last decade. It can be easily adapted to various tasks, including on-line and off-line speech-to-text transcription [12] and robust voice-command control based on real-time keyword spotting [13]. The system has been originally developed for the Czech language, but later it was ported to other languages such as Slovak, Polish, Croatian or Russian, using a cross-lingual adaptation approach [14].

It is well known that ASR depends on large amounts of transcribed speech recordings in order to estimate the parameters of the acoustic model. Recording such large speech corpora is time-consuming and expensive; as a result, there do not exist sufficient quantities of data for disabled users.

Our proposal uses a cross-lingual adaptation approach in which we use most of the available resources in order to recognize atypical speech. This approach is based on an idea similar to the one used to recognize poorly resourced languages: to use data from a well-resourced source language to estimate the acoustic models for a recognizer in a poorly resourced target language [15,16]. In our case, to use data from typical voices in different languages in order to recognize impaired speech using a small amount of training data from disabled people.

Thus, the general idea is to achieve acceptable results reducing the cost of adapting a model to atypical voices. This can contribute to fostering the development of speech applications and help disabled users to be more actively involved in choosing their assistive technology from a wider range of options.

The rest of the paper is structured as follows. 'Related work' section presents related works, 'Proposed method' section presents our proposal for cross-lingual adaptation of speech technologies, the 'A case study for Spanish disabled users' section shows a case study developing our proposal in which we ported a speech recognizer from Czech to Spanish and then adapted it to Spanish disabled users. The experimental results with this example are discussed in 'Experimental evaluation' section. Finally, in 'Conclusions', we present the conclusions and propose future improvements which may increase the performance of the recognizers developed following our proposal.

2 Related work

As discussed in [6], traditional automatic speech recognition techniques are unsuitable for impaired speech for several reasons: the amount of training material is limited, the training samples are highly variable, and they are very different from voices corresponding to non-disabled users.

Due to their very reduced intelligibility, some authors have addressed the problem of automatic recognition of disabled users by carrying out in-depth studies of the most salient features of some types of atypical voices. For example, [17] studied the predictability of articulatory errors and trained a Bayesian network in order to build an augmented ASR system that considered the statistical relationships between vocal tract configurations and their acoustic consequences. Similarly, [18] focused on aspects of syllabic strength for moderate hypokinetic dysarthric speech.

Other work is not so much focused on the recognition itself, but on facilitating the correction of the errors that the recognizer will presumably make. This way, some authors focus on allowing the users to select between alternative word candidates using *n*-best lists [19], while others propose methods that use different approaches to compute the most probable mismatch taking into account the peculiarities of certain pathologies [20].

Despite their high performances, these approaches are focused on particular disabilities or demand a detailed study of the characteristics of the target users. Other authors have addressed more general-purpose approaches. For example, Hawley et al. proposed an incremental approach in which they collected an initial corpus from each user and employed it to train models of words in a reduced vocabulary of commands [6]. Then, they re-estimated the model using the initial training examples and subsequent examples collected from the users while they were employing an application. The advantage of this proposal is that it does not require expert knowledge about the pathologies or a high amount of atypical voice samples, for which the resources are very limited.

We believe that the challenge of limited resources for atypical voices is similar to the case of under-resourced languages: it is costly and time-consuming to gather and process speech material in both cases, which is one of the major limiting factors for speech-enabled application development [21]. Cross-language approaches allow the common exploitation of acoustical similarities between languages in order to be able to use resources available in different languages for the recognition of a less-resourced one. In fact, this approach has also been used to recognize other types of atypical voices such as non-native or accented speech [22]. There are different ways in which existing models can be used along with new data in a target language, mainly training on multilingual data [23], or cross-lingual adaptation of the acoustic [24] and language [25] models.

Our recent experiments show that it is possible to make a cost-efficient cross-lingual adaptation of speech recognition technologies [14]. The continuous speech recognizer may achieve an accuracy between 65% and 75% when using an acoustic model of related languages (e.g. recognition of Croatian using Slovak, Polish, Russian or Czech acoustic models) and between 80% and 85% when the acoustic model is enriched by (semi-automatically obtained) training data of the target language (e.g. recognition of Croatian using Czech + Croatian mixed acoustic model). In this paper, we propose a method to exploit the benefits of cross-lingual adaptation of speech technologies for the recognition of the atypical speech of learning disabled users.

Our previous work in the development of assistive speech technologies for motor handicapped users [13] showed that assistive speech technologies can improve the living conditions of disabled people (and even help them to find job opportunities). Thus, the availability of methods for rapid and cost-efficient prototyping of speech applications, such as the one proposed in this paper, can be of great help for people who suffer from communication disorders.

3 Proposed method

As the development of speech recognition technologies starting from scratch is a very time- and resource-consuming process, we propose to avoid these costs by means of cross-lingual adaptation. The general idea is to use the resources already created for a source language to ease and accelerate the production of the target language resources. The same idea can be used to adapt existing models for common speakers to the needs of handicapped speakers.

3.1 Cross-lingual adaptation

The proposed method for cross-lingual adaptation consists of three partially-independent steps: grapheme-to-phoneme conversion (G2P), building the acoustic model, and building a vocabulary and the corresponding language model. A scheme of the procedure is shown in Figure 1, where the source language (SL) is the one for which the

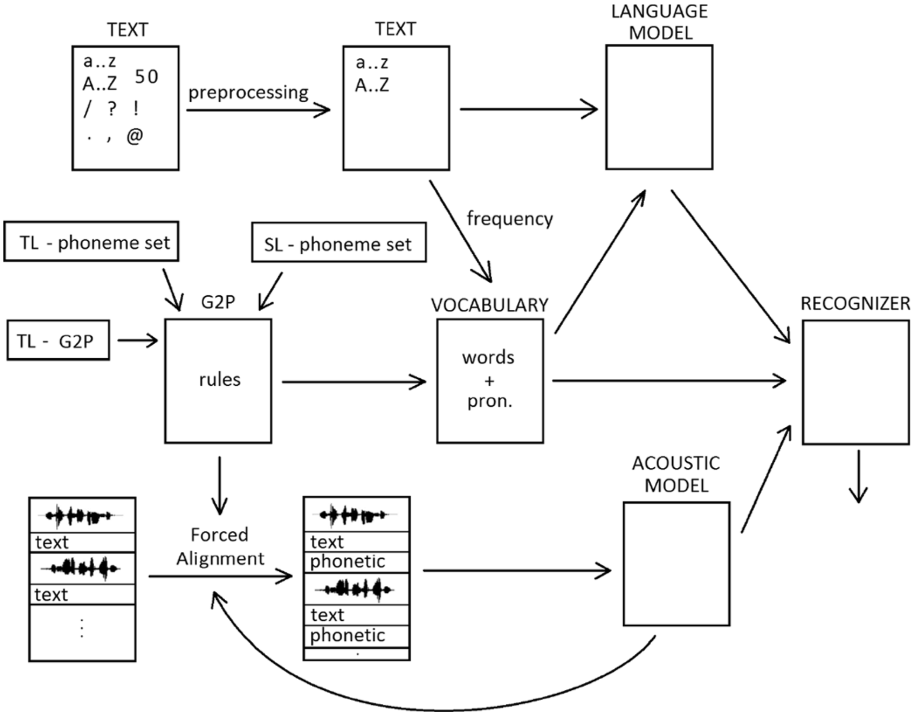

Figure 1 Proposed steps for cross-lingual adaptation from a target language (TL) to a source language (SL).

main resources are available, and the target language (TL) is to be ported to the source language in order to use such resources.

As can be observed, the inputs to the ASR are as follows: the language model in the target language, the (adapted) acoustic model, and a vocabulary in the target language.

G2P conversion determines how the target language words sound in terms of the source language phonetic inventory (*SL - phoneme set* in the figure). It is carried out in two main stages. Firstly, we convert between the target language orthographic form and target language phonetic form (this process is denoted as *TL - G2P* in the figure). Secondly, we decide how to map between the target language phoneme inventory (*TL - phoneme set*) and that of the source language. This involves determining which phoneme pairs (or maybe phoneme groups) are the most similar and possibly which phonemes remain unused.

When the G2P conversion is defined, a fitting acoustic model can be prepared. There exist three alternatives: i) to use the data already available in the SL and mix the model from one or more source languages, ii) mix the SL recordings with some amount of the TL data, and iii) use the TL data only. In the latter case, we can exploit the SL for the development of support technologies in order to lower the demands of expert work (e.g. forced alignment of the TL data using SL models).

Finally, a language model can be built. To do so, it is necessary to gather and pre-process a sufficient amount of textual data in the target language and analyse it to choose a suitable vocabulary. Part of this vocabulary is the phonetic form of the items obtained by the G2P conversion, which can be manually corrected.

3.2 Acoustic and language model

The preparation of training data for the acoustic model is the most time-consuming phase of the adaptation. To reduce the time requirements, we propose to use forced alignment, which consists in assigning time stamps to every word in the input orthographic transcription according to the corresponding audio recording. Forced alignment has many other applications, such as indexing a spoken document for searching, timing subtitles automatically or testing data preparation [26-28]. Processing a document with a forced alignment algorithm requires an audio recording, its corresponding text transcription, an acoustic model (fitting the phonetic inventory used in the phonetic transcription) and a vocabulary containing all words in the document or a G2P conversion module. Eventually, there can be some more input resources to process more complicated tasks (e.g. processing of numerals, physical units, degrees and titles, and special symbols like @, &, or %).

The forced alignment tool that we use is described in [26] and is based on the continuous speech recognizer in the source language. The language model is very constrained as all words must appear strictly in the correct order. They can either fit one of its acoustic forms or be skipped. After the recording is 'recognized' a special post-processing takes place which corrects the different errors that may arise (e.g. when the textual transcript somehow differs from the audio content). As this is a complex process, we encourage the readers to see the details in [26].

Forced alignment is particularly advantageous for the adaptation to atypical voices, as the quality of the acoustic model required can be much lower than the model demanded by the continuous speech recognizer and still it is possible to obtain very accurate results. Moreover, our approach was developed to process inaccurate transcriptions as it can handle 'low quality' acoustic models, i.e. non-robust models trained with a small amount of data [26].

The preparation of the language model does not demand much manual work. The most sensible step is to find a suitable source for the text in the language model. Also, usually there are many characters that are not appropriate for training a language model, there are several ways in which they can be handled, for example, some of them may be erased (e.g. '.', '?'…), others can be rewritten (e.g. numerals) and some of them can be unified and then rewritten (e.g. brackets). Once the text has been processed, a bigram statistical language model is automatically created.

3.3 Adaptation to atypical voices

Every application for disabled users demands a high level of adaptation and customization. Some of these enhancements can be done by the system developers, while others demand active cooperation between the final user and the developers (or a trained assistant). The main enhancements we can enlist include the following: i) training the acoustic model using the data recorded with a subset of final users, ii) general changes in the G2P so it covers some typical speech disorders, iii) adaptation of the acoustic model for the concrete speaker or environment, and iv) reasonable vocabulary limitation (usually context-dependent).

Some users are able to cooperate further (e.g. motor handicapped) but some are not (e.g. severe intellectual disability). If the users are able to cooperate, it is possible to enhance the described adaptation scheme by recording 10 to 60 min of additional recordings in the environment in which they will more frequently use the application and redefine the keywords so that they are easier to pronounce [13]. Finally, users can be offered some training to help them master the assisting technology and improve the performance [29].

Once the cross-lingual adaptation is carried out, a second step is performed to adapt the recognizer trained with common voices to detect impaired speech. Manually sorting and transcribing the recordings is very time-consuming. There is no other choice in the case of longer utterances (sentences), but we propose an automated approach to choose and prepare the isolated word recordings (even if it may imply losing some data).

Our solution requires the audio recording, expected transcription of the recording, and a G2P tool. The solution must be robust to face different phenomena. For example, as some of the users may have reading difficulties, they can be prompted by an assistant to repeat a word that he/she has previously read aloud. This may lead to recording the annotated word more than once. Also, the observed pronunciation may strongly differ from the G2P one. All these obstacles are solved by setting the forced alignment tool properly. We suggest changing the forced aligner language model so the words may be repeated. The phonetic inventory of the forced aligner can also be enhanced by a set of rules modelling the most common pronunciation distortions observed in the data set, as well as the totally mispronounced words (e.g. when omitting groups of phonemes).

Once the recordings are aligned, we select the data suitable for training. If the detected pronunciation differs slightly from the presumed one, we use it or we do not use it otherwise. The similarity is checked using the minimum edit distance (MED) algorithm [30]. The MED algorithm aligns the reference and result sequences in the terms of hits, substitutions, deletions or insertions needed to transform one sequence to the other. We set the rule where a word containing N phonemes in the reference transcription ($N > 3$) has to reach at least $N - 2$ hits to be accepted for training, a heuristic that we found to be appropriate from our previous work [31]. Once the data has been selected, they are added to the training set and the whole procedure is repeated in a second iteration.

4 A case study for Spanish disabled users

In previous work [14], we have successfully ported different ASR models from different Slavic languages, such as Polish, Croatian, Slovak or Russian, to Czech. In this paper, we propose to use a similar approach to build a Spanish model with our ASR system using the resources available for Czech (source language) and Spanish (target language) and adapt it for the recognition of impaired speech (adaptation of the target language).

Although Czech and Spanish are not as similar as the Slavic languages considered in our previous works, we can exploit the Czech resources to speed-up the preparation of the Spanish training data (which is a very time-consuming task). Once we have sufficient amount of the Spanish data, we can leave out the original Czech data and use the Spanish resources only.

The ASR system we have used was originally proposed for processing Czech, a highly inflective Slavic language [12,32], so it supports large vocabularies (it can operate with 500,000 vocabulary items in the on-line mode). The inputs are converted via the FFmpeg codec to the standard 16 kHz pulse-code modulation (PCM) wave format (16 bits per sample), this way the system supports most audio (and video) input formats. The parametrization uses standard 39-MFCC vectors computed on 20 ms frames with 10 ms overlap, and the feature vector is processed with the cepstral mean subtraction (CMS) normalization.

We cover 42 phonemes and 8 non-speech events (e.g. click, breathe, silence, hesitation). The output of the recognizer comprises the written form of the detected word, the detected phonetic form (words usually have several phonetic alternatives), and the time stamps (beginning and end of each word and noise). Additionally, it can be used on-line, when it can also run a post-processing that formats the output as it was pronounced.

Continuous speech recognition is very powerful, but it can also be very demanding for disabled users who might not be able to pronounce a whole sentence correctly but still be able to say it word by word. That is why we employ a keyword spotter (KWS), which is a speech recognition technology used for the detection of isolated words of interest (keywords) from an audio stream. Typical applications include smart homes, industrial enhancements, making audio-archives accessible, or security purposes. The specific implementations and their accuracies may differ between on-line and off-line applications as shown in [31].

As the algorithms mostly rely on the acoustic similarity between keywords and features of the audio stream, we must pay attention when choosing the keywords. If we were interested in detecting acoustically similar words or words that are substrings (one word is included in the other) the system will raise many false alarms or confuse the output. This problem is particularly significant in Slavic languages where the words often differ in the ending only [33].

As we demand the ability of on-line response, we use our KWS system derived from the continuous Czech speech recognition system described previously. It can be used in both on-line and off-line modes. Another advantage is that this KWS uses the same parametrization and acoustic model as the continuous speech recognizer. The difference is in the language model and the vocabulary. It employs phoneme-based *n*-grams to model speech and build the filler model. Such defined fillers compete with keywords from the vocabulary and with the non-speech events and noises. The performance is controlled by language model parameters and penalties so we can tune the ratio between false alarms and missing detections.

As indicated in 'Adaptation to atypical voices' section, we propose to constrain the minimal length of a keyword to three phonemes and the minimum difference between keywords in at least two phonemes. This ensures there will be no substrings in the vocabulary but still there can be false alarms caused by a combination of two (or more) words in the audio stream that form together one of the keywords. This problem does not appear in the case of isolated word utterances.

4.1 Cross-lingual adaptation from Czech to Spanish

To build the Spanish acoustic model we used the Albayzin [34] corpus. The corpus comprises two sub-corpora with 6,800 utterances each: one based on texts extracted from novels and the other based on queries to a geography database. The utterances were recorded under good acoustic conditions (quiet offices, with the same set of professional microphones) and were pronounced by 304 speakers (152 female, 152 male), whose age varied from 18 to 55 years. The Albayzin corpus represents 12 h 52 min of annotated speech in 13,600 gender and phonetically balanced sentences.

To obtain the training data, we carried out the following preparations. The first step was the conversion of the original audio data from original .ses format to the .wav format - we used (16 kHz, mono 16 bit per sample PCM). From the transcriptions, we removed the punctuation marks, replaced the numbers with their word forms (e.g. 512 = 'quinientos doce'), and processed some special symbols (mainly units of areas or distances). To annotate noises, we employed our forced aligner module (see 'Acoustic and language model' section), which was able to detect and annotate the noises and select the best alternative phonetic representation for each word.

For the cross-lingual adaptation, we employed the proposed G2P conversion in two stages. The first stage was the conversion from the Spanish orthographic form to the Spanish phonetic form. In order to carry out the conversion, we used several rules that varied from unigrams to trigrams. The converter sequentially parsed the orthographic form finding the longest fitting rule. Then, the output was slightly modified to reflect the voicing assimilation and some more phenomena by a set of regular expressions.

The second stage was a substitution of the phonetic inventories: we substituted the phonemes of the target language (Spanish) with the phonemes contained in the phonetic set of the source language (Czech). As discussed before, we had already done this between Slavic languages [14] where the phonetic inventories were quite similar. As shown in Table 1, for the adaptation between Spanish and Czech, we substituted 30 Spanish phonemes with 28 Czech phonemes. So, not all the Czech phonemes are used, and a few Spanish phonemes were substituted by the same Czech phoneme. The phonetic inventory can be optimized after the prototyping phase is finished.

Table 1 The proposed mapping between Spanish and Czech phonemes

Spanish phoneme [adapted X-SAMPA] → Czech phoneme [PAC]				
z→s	d→d	ñ→ň	J→j	á→á
k→k	f→f	p→p	é→é	u→u
C→č	j→X	r→r	í→í	o→o
g→g	l→l	s→s	ó→ó	i→i
y→Č	m→m	t→t	ú→ú	e→e
b→b	n→n	x→ks	0→0	a→a

For example, the conversion of the Spanish sentence 'Guillermo y Yolanda practicaban ciclismo con Jaime' to Czech phonetic form is as follows:

- ESP_{text}: guillermo y yolanda practicaban ciclismo con Jaime
- ESP_{phon}: giyermo i yolanda praktikaban ziklismo kon jaJme
- ESP_{PAC}: giČermo i Čolanda praktikaban siklismo kon Xajme

Where the rules used for the first three items in the sentence are as follows: 'gui' → 'gi' ; 'll' → 'y' ; 'e' → 'e' ; 'r' → 'r' ; 'm' → 'm' ; 'o' → 'o' ; ' y' → 'i' ; 'y' → 'y' ; 'o' → 'o' ; 'l' → 'l' ; 'a' → 'a' ; 'n' → 'n' ; 'd' → 'd' ; 'a' → 'a' ; ' ' → ' '. As can be observed, rules may differ if the beginning or end of a word is encountered.

To build a language model and vocabulary for continuous Spanish recognition, it was necessary to retrieve and process a large amount of Spanish texts. As we were interested in rapid development, we used daily Spanish and international news from different web pages. We downloaded 11.7 GB of texts from http://elpais.com/, http://www.20minutos.es/, and http://spanish.news.cn/.

As the text corpus was gathered from downloaded articles, we carried out a careful post-processing to prepare the corpus for training the target language model (statistical bigram model of Spanish). This way, we used different scripts to remove all HTML tags, foreign (non-Spanish) characters, English words, and other parts of the text that were not suitable for our purpose (e.g. currency rates, information from stock exchange and sports results). We also replaced all numbers with their orthographic transcription. For example, instead of 'in 1926' we had 'in nineteen hundred and twenty six' (in Spanish: 'en 1926' - 'en mil novecientos veintiséis').

Once the data was processed, we computed the bigram language model. For preparing the vocabulary, we employed all words that occurred more than 10 times in the corpus. We decided to use collocations (several words that usually go together - for example 'Los Ángeles') for definite and indefinite articles (e.g. 'el pan', 'un profesor') as short items are disadvantageous for speech recognition. Using this approach, we generated a vocabulary with 54,217 words and word collocations.

4.2 Acoustic models generated

As we trained and compared several acoustic models, we decided to list them here together with their features. We have quantified the amount and sources of training data, so they can be easily compared. In all the models, the Spanish data are used twice - first with floating CMS, then with CMS computed over all the recordings.

AM_CZ is the model made from 200 hours of Czech recordings already available from our previous work [14] (as previously discussed, we consider Czech our source language).

AM_cz&ES_cross denotes five different models. All consist of 1.45 h of Czech recordings (chosen to cover the Czech phonemes not used in Spanish phonetic inventory) and approximately 10 h of Spanish data from the Albayzin corpus (for details, see 'Cross-lingual adaptation from Czech to Spanish' section).

AM_CZ&es denotes the acoustic model consisting of 133 h 24 min of Czech recordings, 12 h 52 min of continuous Spanish speech (whole Albayzin corpus), and 69 min of isolated words uttered by disabled people.

AM_cz&ES consists of 1 h 27 min Czech data and all the Spanish data mentioned in *AM_CZ&es*. The Czech training data mostly covers phonemes unused in the Spanish vocabulary (as we experimented with the best pairing of Czech and Spanish phonemes). This can be considered equivalent to a 'Spanish only' model.

AM_cz&ES_2 is similar to *AM_cz&ES* but including more isolated words uttered by the disabled people (140 min).

AM_cz&ES_sent consists of the *AM_cz&ES_2* and 28 min of continuous speech uttered by disabled people.

4.3 Adaptation to Spanish atypical voices

To gather a corpus of impaired speech, we have worked together with two associations of people with learning disabilities in Southeastern Spain: JABALCÓN[a] and APAFA[b]. Both associations are based in mainly rural areas and have around 100 affiliated persons. They work for the social integration of the learning disabled through different programs of professional development such as wood workshops. People in these associations are mainly adults from the towns and villages nearby who visit the centres during the day.

Due to the characteristics of the users, the recordings had to be carried out in special conditions. Firstly, we had several meetings with the professionals who work with the users on a daily basis, in order to select the individuals that would participate in the recordings. The selection was carried out according to the following criteria:

- To select subjects with a wide range of phoniatric problems.
- To select only subjects for which the participation in the recordings would not impact their wellness, as certain disabilities imply that a change in the person's agenda can be very disruptive.
- To select only subjects who were willing to participate voluntarily with the consent of their families and/or caregivers.

Following this approach, 42 subjects were selected. As the subjects could not participate in long recording sessions, we had to make several visits to the associations to make the recordings. During these visits, we carried out three types of recordings: single words, sentences, and conversations. The first group of recordings were frequent words from their daily activities from a vocabulary that was agreed with their caregivers and categorized in to the following six scenarios: 'street' (street, coin, house...), home (bed, sofa, table...), food (apple, meat, fork...), 'family' (father, mother, sister...), 'dressing' (trousers, jersey, coat...), and 'me' (cold, happy, hungry...). The second group were basic sentences containing words in this vocabulary (e.g. 'The fork is on the table', 'I am cold', 'Open the door'...). Finally, the third group was comprised of open conversations about their daily activities: activities inside the association, visits around the village, sports (especially football), summer activities and family.

During the recording sessions, there was an assistant from our team, a caregiver from the association, and the subject being recorded. Although the sessions were planned to be adapted to the subjects, they could be interrupted if the subject was tired or was not willing to continue with it, or if the caregiver decided for some reason to stop the session. In the first two groups of recordings, the assistant or the caregiver (depending on the subject being recorded) would read the word or sentence aloud and the subject would repeat it. Our aim was to record only the subject, but in some cases they would start speaking while the other person was reading aloud, which forced us to process the corpus. In the last group, the assistant asked them questions that they could respond to without restrictions. In this case, the questions were not always the same, as the idea was to facilitate conversation-asking questions about the topics for which each subject seemed to be more talkative. The recordings were not cut and the whole conversation was recorded, from which we discarded the bits not corresponding to the disabled speakers. Table 2 shows a summary with the number of recordings of each type.

The recordings took place in the associations, so that the participants did not have to travel and the recordings contained the environmental noise that would also surround the users when employing the generated recognizers. Due to the high number of activities carried out by the associations, the same rooms were not always available during the recording sessions, also the recordings were done at different times of the day, so the levels of acoustic noise vary and in some cases, there appear some events that produced louder noises such as opening/closing doors. We believe that these situations are desirable to build acoustic models that consider the noise that will be present during the usage of the speech recognizer. State-of-the-art databases of impaired speech (some of them are described in [35]), are usually recorded in laboratory conditions, which makes it more difficult to employ the recognizers trained with them in real settings.

With respect to the single words, the accepted data (chosen by the automatic classification) were used for training a new acoustic model, and we repeated this procedure with the adapted model in several iterations. In the first round, we used 6,156 recordings to obtain 2,268 words suitable for training. The second round (with the improved acoustic model) chose 921 more words for training. Then, from the other 6,900 recordings, we chose other additional 3,300 samples.

With respect to the sentences, they consist of 45 min of annotated speech in 940 utterances. We automatically prepared the phonetic annotations using the speech, its orthographic transcription and the G2P technique described in 'Cross-lingual adaptation' section, and corrected them as well as the orthographic transcriptions in the cases in which they were inaccurate (i.e. they did not correspond to what was pronounced). For adapting the acoustic model, we employed 618 utterances with 28 min of speech.

Finally, the conversations were split by speakers. We used only the parts with disabled speakers. From 47 min of conversation between disabled speakers and moderators, we separated 16 annotated min for test purposes.

Table 2 Number of elements and impaired subjects recorded

Group of recordings	Number of recordings	Number of users
Single words	13,056 words	42 users (30 male, 12 female)
Sentences	940 utterances (45 min)	18 users (13 male, 5 female)
Conversations	47 min	18 users (13 male, 5 female)

5 Experimental evaluation

To evaluate our proposal, we have carried out different experiments corresponding to the recognition of Spanish disabled users with Czech as the source language (the case study described in the 'A case study for Spanish disabled users' section). Concretely, we have studied two speech recognition technologies (keyword spotter and continuous speech recognizer) using the different acoustic models described in the 'Acoustic models generated' section and also varying the vocabulary and language models.

5.1 Evaluation metrics

We have used four evaluation metrics: ACC, DR, FA, and CORR. ACC stands for accuracy (Eq. 1), DR stands for detection rate (Eq. 2), FA stands for false alarm rate as the number of false positives per keyword per hour (Eq. 3), and CORR stands for correctness (Eq. 4). In the following equations, TP means true positives (correctly detected items), FP means false positives (false alarms), FN means false negatives (items that were not detected), N_{kw} denotes the number of keywords in the vocabulary, Dur stands for the total duration of recordings, and N_{rec} is the number of words in the reference transcription that really appear in each audio recording.

$$\mathrm{ACC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \times 100\ [\%] \tag{1}$$

$$\mathrm{DR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\ [\%] \tag{2}$$

$$\mathrm{FA} = \frac{\mathrm{FP}}{N_{\mathrm{kw}} \times \mathrm{Dur}}\ [1/\mathrm{kw/h}] \tag{3}$$

$$\mathrm{CORR} = \frac{\mathrm{TP}}{N_{\mathrm{rec}}} \times 100\ [\%] \tag{4}$$

5.2 Keyword spotter for common speakers

This experiment was the first to be done. As we needed to decide which ratio between Czech and Spanish training data was the best for training the acoustic models, we made a comparative experiment between the *AM_CZ*, *AM_CZ&es* and *AM_cz&ES_cross* acoustic models (Table 3). The latter represents a group of five models that were cross-verified: we split the Albayzin corpus into five equivalent subparts and used one for testing and the remaining four for training. From the text transcription, we chose all the words without substring occurrence (every vocabulary item had to have at least three phonemes and differ from each other in two phonemes at least). The main idea behind this rule is that we need to distinguish completely mispronounced words and somehow distorted words. Given a word that should

Table 3 Comparison of acoustic models with different balance of Czech and Spanish training data evaluated over typical Spanish voices

Acoustic model	ACC (%)	DR (%)	FA [1/kw/h]
AM_cz&ES_cross	54.10	85.88	1.94
AM_CZ&es	52.30	81.42	2.03
AM_CZ	45.75	57.00	0.46

be pronounced as x and was pronounced as x', the N-2 threshold means that there are only N-2 out of N correctly pronounced phonemes in x'. This usually implies that x' (e.g. 'boleto'/ticket) is a completely different word from x (e.g. 'coleta'/ponytail) and thus, it is not as restrictive as it may appear.

In average, there were 350 items chosen for the spotting. In this case, we spot the words from a continuous speech so we also checked if the combination of words could substitute a vocabulary item. The phonetic forms were generated automatically by the G2P module (only one alternative for each item). As our work is focused on disabled speakers who have troubles with spoken communication, we set the system for high DR (even if it implies possibly higher FA). Although false alarms may raise errors, it is possible to recover from them in the application that employs the ASR by establishing interaction contexts or scenarios, and also providing N-best lists from which the users may select the most appropriate response. This may be more suited for disabled users than having a recognizer with a lower DR that gives the impression to be not responding to the user's inputs.

From all points of view, the *AM_cz&ES_cross* achieved the best results. This was a surprise for us because when we were porting Slavic languages, it was advantageous to use more data (even from relative languages and not the source or target one) as it guaranteed the robustness of the model. In this case, when porting a Roman language, the mapped phonemes were so different that it was better to use the Czech data only for uncovered phonemes (phonemes not used in Spanish). The *AM_CZ* model had low FA, but the DR was insufficient. The results of this experiment (with almost no Czech data) are very promising in comparison with our former work with Czech KWS [31] so we decided to minimize the usage of Czech acoustic data in the acoustic model training.

5.3 Keyword spotter for disabled speakers

We made two groups for these experiments: pack1, with 131 items in the lexicon, and pack2 with 177 items. We launched the KWS using the *AM_cz&ES* acoustic model and created two lexicons for each group: the initial lexicon (*pack1_base* and *pack2_base*) and a lexicon with alternative phonetics automatically obtained with a set of rules based on expert observation of the disabled users (*pack1_alter* and *pack2_alter*). Phonetic conflicts and similarities were revealed by the MED algorithm [30] and corrected. As there was a large portion of data that was not marked as suitable for training (approximately one half), we decided to test if there is a difference between the recognition of suitable (*suit*) and unsuitable (*unsuit*) data. The results are shown in Table 4.

As can be observed, the results are not very encouraging but we must realize there are only 69 min (approximately 35 min of pure speech) to adapt the acoustic model to disabled speakers. However, we can state that the algorithm that chooses the training data was correct, as it was able to correctly discriminate the unsuitable data from the suitable samples. We can also clearly see the impact of the alternative phonetics in the vocabulary (labelled *alter*). The only drawback is the increase of the false alarm rate. But as stated before, in this domain, it is preferable to have a higher detection rate than to lower the false alarm rate.

Even though the ratio between suitable and unsuitable data is practically the same in *pack1* and *pack2*, the results of *pack2* are better. The reason lies in the vocabulary (although there are 35% more items in *pack2*, there are less phonetic conflicts). That is why we decided to split the whole vocabulary into the six scenarios described in the 'Adaptation to Spanish atypical voices' section so only a subset of the vocabulary is used each time (street, home, food, dressing, me, or family). The results are shown in Table 5. As can be observed, narrowing the vocabulary helped to obtain much better recognition results.

We also wanted to verify that increasing the amount of the training data pronounced by the disabled speakers improves the KWS performance. We ran the KWS with *AM_cz&ES_2* over the *pack2* recordings and the accuracy was 25.30% when using one phonetic only and 29.60% when using the alternative phonetics. This shows the negative impact that the reduced amount of utterances by disabled speakers had in the previously discussed experimental results.

Table 4 Impact of alternative phonetics and recordings' pronunciation quality of the atypical voices

Test data set	ACC (%)	DR (%)	FA [1/kw/h]
*pack*1_*base*$_{\text{all}}$	07.60	08.10	2.75
*pack*1_*base*$_{\text{suit}}$	13.60	14.90	3.80
*pack*1_*base*$_{\text{unsuit}}$	00.80	00.80	1.70
*pack*1_*alter*$_{\text{all}}$	08.80	10.10	5.98
*pack*1_*alter*$_{\text{suit}}$	15.60	18.30	7.25
*pack*1_*alter*$_{\text{unsuit}}$	01.20	01.40	4.71
pack2_base$_{\text{all}}$	11.80	12.40	1.56
pack2_alter$_{\text{all}}$	14.10	15.80	4.15

Table 5 Impact of dividing vocabulary into scenarios - atypical voices

Scenario	ACC (%)	DR (%)	N_{kw}
Street	23.20	25.27	47
Home	25.00	28.54	78
Food	24.40	27.54	77
Family	30.70	30.77	6
Dressing	23.10	26.07	65
Me	21.70	22.96	22

Our last experiment with the KWS was the evaluation of the possible improvements gained via speaker (and simultaneously channel) adaptation, as the parameters of the recording differed between the Albayzin corpus and the devices used to record the disabled speakers. We used the *pack2* data which were found suitable for training to make the adaptation and compared the results when using *AM_cz&ES* with and without the adaptation.

Maximum likelihood linear regression (MLLR) [36] was employed for adaptation. We adapted the constrained version of this method, known as constrained MLLR (CMLLR) [37], where the transformation matrix applied for adaptation of means has to be the same as the one used for adaptation of variances. Therefore, the adaptation can also be performed in the feature space and the adapted feature vector $\hat{\mathbf{o}}$ can be expressed as

$$\hat{\mathbf{o}} = \mathbf{W}\xi \tag{5}$$

where $\mathbf{W}$ is the extended transformation matrix, $\xi = \left[\, \omega \;\; o_1 \;\; o_2 \; \ldots \; o_n \,\right]^{\top}$ is the extended vector of features, n is the dimension of data and ω represents a bias offset.

The matrix $\mathbf{W}$ has to be calculated within an iterative process, where the likelihood of adaptation data with known transcription is maximized [37]. Note that in our case, only one global transform was estimated for all Gaussian components of the system. Hence, it was not necessary to include the Jacobian of the transformation in the likelihood calculation.

There exist two basic approaches on how to perform adaptation for a target speaker, supervised and unsupervised [38]. The former utilizes adaptation data that is annotated manually by a human expert. The latter employs a speech recognizer, which creates these transcripts automatically.

We have employed a supervised adaptation approach. At first, the manual orthographic transcripts of adaptation data were available. Then, we performed forced alignment to adapt the utterances using a speech recognizer operating with the baseline speaker independent model and the lexicon containing all pronunciation variants of all words occurring in the orthographic transcripts. As a result of this process, we obtained accurate phonetic transcripts with labelled noises produced by speakers (breathing, various hesitation sounds, cough, lip-smack, etc.). Finally, general speaker-specific transformations were estimated using this annotated adaptation data.

As the impact of adaptation differs between speakers, it is not reasonable to measure an average improvement. For speakers who were badly processed by the baseline, the gain varied. For some of them ACC decreased from 11.37% to 8.75%; for others, increased from 2.92% to 6.07%. For better recognized speakers, the behaviour was much more predictable: ACC increased by approximately 20% (e.g. 40.19% increased to 58.52%, and 59.85% increased to 73.14%). Thus, the speaker-channel adaptation can be very useful but, on the other hand, it is necessary to identify the user.

5.4 Continuous speech recognition

For testing the language model and vocabulary, we randomly chose 1,500 testing utterances from the Albayzin corpus. These utterances contained 1 h and 24 min of continuous speech. We did not use texts from these utterances for preparation of the language model or for the vocabulary. Thus, there could appear out of vocabulary (OOV) words. Accounting for the impact of the number OOV words is very important, as if it is high, it indicates that the vocabulary should be bigger and there is space for improving the recognition results. Table 6 shows the results from experiments in which we used the language model and vocabulary with several acoustic models. The worst results were obtained with the *AM_CZ* model that consisted only on Czech training data. The inclusion of Spanish data increased the results to a great extent, obtaining a maximum accuracy of 63.29% with *AM_cz&ES_2*. The number of OOV words was very high, so there is possibly a good opportunity to improve our vocabulary and reach even better accuracy. In the case of the non-disabled speakers, a part of Albayzin was obtained using a geography database, the OOV were mainly substantives describing geographical accidents and proper names of Spanish areas, rivers and mountains. In the case of the learning disabled, the OOV were mainly related to the way they talk, which is very informal and also affected by the way of speaking in the

Table 6 Results for continuous speech recognition with different acoustic models - typical voices

Acoustic model	CORR (%)	ACC (%)	OOV (%)
AM_CZ	31.59	29.39	23.18
AM_cz&ES_sent	65.08	63.03	23.18
AM_CZ&es	64.55	62.45	23.18
AM_cz&ES	65.44	63.24	23.18
AM_cz&ES_2	65.44	63.29	23.18

location of the associations, where many diminutives are employed.

The results of continuous speech recognition with the acoustic model for disabled speakers is shown in Table 7. We used 16 min of impaired speech recorded as a conversation (see more in the 'Adaptation to Spanish atypical voices' section). We performed several experiments: normal language model and vocabulary (*LM_ES*), vocabulary with multiple pronunciations (*LM_ES_multiVOC*), and adding test sentences to the language model once (*LM_ES_DS_multiVOC*) or several times (*LM_ES_nDS_multiVOC*). As can be observed, the results are much worse for disabled speakers than for average speakers. This is not only because of their pronunciation but also because the recordings did not take place under ideal conditions as in the Albayzin corpus, which was gathered with noise isolation and professional equipment.

However, as the continuous speech recognizer accuracy is higher than 60%, it is possible to use this recognizer to semi-automatically obtain new data for training. In order to do this, it is possible to use the approach that we described in [14], specially if we have at least partially transcribed speech. For example, it is possible to gather this type of data from radio broadcasts, as on their web pages there are usually some abstracts about the recordings and in a lot of cases there are some rewritten parts also. With our recognizer, we can automatically transcript the recordings and compare them with the text from the web page. If we find matchings that are long enough (usually a few words), we can extract these parts and use them to build more accurate models in a short time.

6 Conclusions

Speech technology can be a very valuable help for learning disabled users, whose varied conditions result in different dysfunctions of their speech system. Different efforts have been made by the scientific community in order to propose different approaches to the implementation of automatic speech recognizers for these users. However, many of the state-of-the-art approaches focus on particular pathologies or imply an in-depth study of specific dimensions such as certain articulatory features. This makes the development of automatic speech recognizers costly, as there is a need for a large amount of specific data samples to train the recognizer.

Table 7 Results for continuous speech recognition with different language models - atypical voices

Language model	CORR (%)	ACC (%)	OOV (%)
LM_ES	2.75	2.30	11.25
LM_ES_multiVOC	3.32	2.70	11.25
LM_ES_DS_multiVOC	3.72	2.97	11.25
LM_ES_nDS_multiVOC	5.32	4.52	11.25

In this paper, we have presented a cross-lingual approach for the development of speech recognizers for disabled people. Our main aim was to study mechanisms to take the most of the resources already available for average users in different languages for the recognition of atypical voices. Although the generality of the method reduces its performance compared to recognizers trained on large databases of impaired speech, it allows rapid and cost-efficient development, which makes it suitable for fast development of assistive speech applications.

We have evaluated the proposed model preparing a KWS and CSR for Spanish using Czech resources, and adapting the Spanish resources obtained with the cross-lingual approach to fit learning disabled speakers.

The experimental results show that the KWS for common Spanish speakers achieves results comparable to those reached by the source KWS when used for Czech. The CSR obtains accuracies over 60%. Although it does not seem very satisfactory, we would like to emphasize that there were more than 20% OOV words. The high OOV rate demonstrates we have to increase the vocabulary for CSR (and concurrently retrain the language model). Taking all these aspects into account, the results for the recognition of common Spanish speakers are very promising, and when the new vocabulary is ready, we should be able to launch the almost automatic improvement procedure proposed in [14]. This procedure will give us new data to improve the robustness of the acoustic model for Spanish.

Both technologies (KWS and CSR) obtained lower accuracy rates with the disabled speakers. In the case of CSR, we believe that limiting the vocabulary and changing the system behaviour from continuous speech recognition to recognition of isolated words would provide several benefits. On the one hand, the speakers (who are not used to speak for long periods of time) will have the opportunity to relax their vocal tract and pronounce the words better and on the other hand isolated word dictation is more reliable than CSR. In the case of KWS, although the results are not sufficient for real operation, we have shown the positive impacts of the tested enhancement techniques. We show the importance of a careful choice of keywords as well as context-dependent limitation of the vocabulary, together with the use of proper alternative phonetics. We have also demonstrated the improvement achieved through the supervised speaker (and channel) adaptation.

We have proved that for common speakers a cost-efficient cross-lingual adaptation can be done even with a training dataset smaller than the usual databases for training speech recognizers. Especially, the forced alignment tool has proven to be very useful. In the case of disabled speakers, the task itself is challenging.

However, we succeeded in the automatic elimination of the data which was not suitable for training, and generally, we can say that the proposed method leads to significant reduction of time-consuming expert work.

For future work, we plan to improve the quality of phonetic alternatives using a data-driven weighted finite state transducer (WFST)-based approach. The WFST can be trained directly from a vocabulary where the input data consists of (*Spanish word - Czech phonetic*) pairs [39]. Along with the general-purpose approach, it is also possible to adapt the vocabularies to fit the speakers individually, this WFST-based system should be able to 'learn' the rules created by an expert and to propose alternatives replacing the actual G2P module.

Another promising guideline comes from our ongoing work. We are currently making experiments replacing the physical state decoder of the speech recognizer by a neural network. This network has the advantage that it uses seven subsequent parametrized frames to classify the middle one, so it uses more information. The experiments show promising improvement especially for the data with low initial recognition score. Another advantage of this approach is that it uses the same training data as the current HMM-based decoder. We want to apply this change for the existing speech recognizer and also to prepare another KWS based on these neural networks which output would be processed directly by the weighted finite state transducers.

Although the number of recordings in our database of atypical voices is in the same order as other state-of-the-art corpora (see a review of existing corpora in [35] and [40]), it would be desirable to record new data and/or manually annotate the data marked as 'not suitable for training'. The latter solution has the drawback of requiring a large amount of expert work, but we can presume it would help to recognize the more severely impaired speakers.

As the emphasis in the paper has been to take advantage of the existing resources to help to develop new assistive technologies for disabled people, we offer our collaboration and resources to interested researchers. In the near future, we plan to test the adapted recognizer under real conditions integrating it in an application for tablets, the results of this new stage of our research will also be at the disposal of the scientific community.

Endnotes

[a] Asociación pro discapacitados psíquicos Jabalcón, www.jabalcon.org.

[b] Asociación de padres, familiares y amigos de personas con discapacidad intelectual del norte de Almería, www.apafa.es.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors have read and approved the manuscript.

Acknowledgements

This research was supported by the project 'Favoreciendo la vida autónoma de discapacitados intelectuales con problemas de comunicación oral mediante interfaces personalizados de reconocimiento automático del habla', financed by the Centre of Initiatives for Development Cooperation (Centro de Iniciativas de Cooperación al Desarrollo, CICODE), University of Granada, Spain. This research was supported by the Student Grant Scheme 2014 (SGS) at the Technical University of Liberec. The authors want to thank the staff from the associations JABALCON and APAFA and the participants in the recordings for their contribution to make this work possible.

Author details

[1] Institute of Information Technology and Electronics, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic. [2] Department of Languages and Computer Systems, University of Granada, CITIC-UGR, Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain.

Received: 1 February 2014 Accepted: 1 October 2014
[redacted]

References

1. J Sigafoos, RW Schlosser, GE Lancioni, MF O'Reilly, VA Green, NN Singh, GE Lancioni, NN Singh, in *Assistive Technology for People with Communication Disorders*. Autism and Child Psychopathology Series (Springer, New York, 2014), pp. 77–112
2. O Saz, J Simón, W-R Rodríguez, E Lleida, C Vaquero, Analysis of acoustic features in speakers with cognitive disorders and speech impairments. EURASIP J. Adv. Signal Process. **2009**, 159–234 (2009)
3. TH Falk, W-Y Chan, F Shein, Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. **54**, 622–631 (2012)
4. MA Neerincx, AHM Cremers, JM Kessens, DA van Leeuwen, KP Truong, Attuning speech-enabled interfaces to user and context for inclusive design: technology, methodology and practice. Univers. Access. Inform. Soc. **8**, 109–122 (2009)
5. WR Rodríguez, O Saz, E Lleida, A prelingual tool for the education of altered voices. Speech Commun. **54**, 583–600 (2012)
6. MS Hawley, SP Cunningham, PD Green, P Enderby, R Palmer, S Sehgal, P O'Neill, A voice-input voice-output communication aid for people with severe speech impairment. IEEE Trans. Neural Syst. Rehabil. Eng. **21**, 23–31 (2013)
7. Ed Joode, Cv Heugten, F Verhey, Mv Boxtel, Efficacy and usability of assistive technology for patients with cognitive deficits: a systematic review. Clin. Rehabil. **24**, 701–714 (2010)
8. GE Lancioni, NN Singh, MF O'Reilly, J Sigafoos, D Oliva, in *Assistive Technology for People with Severe/Profound Intellectual and Multiple Disabilities*. Autism and Child Psychopathology Series (Springer, New York, 2014), pp. 277–313
9. KF McCoy, JL Arnott, L Ferres, M Fried-Oken, B Roark, Speech and language processing as assistive technologies. Comput. Speech Lang. **27**, 1143–1146. (2013-09)
10. O Chia Ai, M Hariharan, S Yaacob, L Sin Chee, Classification of speech dysfluencies with MFCC and LPCC features. Expert Syst. Appl. **39**, 2157–2165 (2012)
11. J Borg, S Larsson, P Östergren, The right to assistive technology: for whom, for what, and by whom? Disabil. Soc. **26**, 151–167 (2011)
12. J Nouza, K Blavka, P Červa, J Zdansky, J Silovsky, M Bohac, J Prazak, Making czech historical radio archive accessible and searchable for wide public. J. Multimed. **7**, 159–169 (2012)
13. P Červa, J Nouza, in *Proceedings of the Conference of the International Speech Communication Association Interspeech: 27-31 August 2007; Antwerp, Belgium, (ISCA, France)*. Design and development of voice controlled aids for motor-handicapped persons, (2007), pp. 2521–2524

14. J Nouza, P Červa, M Kucharová, Cost-efficient development of acoustic models for speech recognition of related languages. Radioengineering. **22**, 866–873 (2013)
15. P Lal, S King, Cross-lingual automatic speech recognition using tandem features. **21**, 2506–2515 (2013)
16. L Besacier, E Barnard, A Karpov, T Schultz, Automatic speech recognition for under-resourced languages: a survey. Speech Commun. **56**, 85–100 (2014)
17. F Rudzicz, Production knowledge in the recognition of dysarthric speech. PhD thesis, University of Toronto (2011)
18. SA Borrie, MJ McAuliffe, JM Liss, GA O'Beirne, TJ Anderson, A follow-up investigation into the mechanisms that underlie improved recognition of dysarthric speech. J. Acoust. Soc. Am. **132**, 102–108 (2012)
19. J-P Hosom, T Jakobs, A Baker, S Fager, in *Proceedings of the 11th Conference of the International Speech Communication Association (Interspeech): 26-30 September 2010; Makuhari, Japan (International, Speech Communication Association, France)*, ed. by T Kobayashi, K Hirose, and S Nakamura. in *Automatic speech recognition for assistive writing in speech supplemented word prediction*, (2010), pp. 2674–2677
20. WK Seong, JH Park, HK Kim, in *Dysarthric Speech Recognition Error Correction Using Weighted Finite State Transducers Based on Context-Dependent Pronunciation Variation*. LNCS, ed. by Miesenberger K, A Karshmer, P Penaz, and W Zagler (Springer, Heidelberg, 2012), pp. 475–482
21. I Kraljevski, G Strecha, M Wolff, O Jokisch, S Chungurski, R Hoffmann, in *Cross-Language Acoustic Modeling for Macedonian Speech Technology Applications*, ed. by S Markovski, M Gusev. Advances in Intelligent Systems and Computing (Springer, Berlin, 2013), pp. 35–45
22. D Imseng, H Bourlard, J Dines, PN Garner, M Magimai-Doss, Applying multi- and cross-lingual stochastic phone space transformations to non-native speech recognition. IEEE Trans. Audio Speech Lang. Process. **21**, 1713–1726 (2013)
23. T Schultz, K Kirchhoff, *Multilingual Speech Processing*. (Academic Press, USA, 2006)
24. D Imseng, P Motlicek, PN Garner, H Bourlard, in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU): 8-12 December 2013; Olomouc, Czech Republic (IEEE, USA)*. Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition, (2013), pp. 332–337
25. P Xu, P Fung, Cross-lingual language modeling for low-resource speech recognition. IEEE Trans. Audio Speech Lang. Process. **21**, 1134–1144 (2013)
26. M Bohac, K Blavka, Text-to-speech alignment for imperfect transcriptions. LNCS: Text, Speech Dialogue. **8082**, 536–543 (2013)
27. J Zhang, F Pan, Y Yan, in *Proceedings of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC): 26-27 August 2012; Nanchang, China (IEEE, U.S.A.)* An LVCSR based automatic scoring method in English reading tests, (2012), pp. 34–37
28. DP Córdova Lucero, DT Toledano, in *Proceedings of the Joint 7th Spanish Speech Technology Workshop and the Iberian SLTech Workshop: 21-23 November 2012; Madrid, Spain (Springer, Germany)*. Preliminary results of alignment of text and audio in news and songs, (2012), pp. 59–68
29. J Nouza, P Červa, J Chaloupka, in *Proceedings of the International Conference on Health Informatics (HEALTHINF - BIODEVICES): 26-29 January 2011; Rome, Italy (SciTePress, U.K.)* Rainbow bridge: Training center based on voice technology for people with physical disabilities, (2011), pp. 529–533
30. RA Wagner, MJ Fischer, String-to-string correction problem. J. ACM. **21**, 168–173 (1974)
31. M Bohac, in *Proceedings of the 54th International Symposium ELMAR: 12-14 September 2012; Zadar, Croatia (IEEE, USA)*. Performance comparison of several techniques to detect keywords in audio streams and audio scene, (2012), pp. 215–218
32. J Nouza, J Zdansky, P Červa, J Silovsky, Challenges in speech processing of Slavic languages (case studies in speech recognition of Czech and Slovak). LNCS. **5967**, 225–241 (2010)
33. M Bohac, J Nouza, K Blavka, Investigation on most frequent errors in large-scale speech recognition applications. LNCS: Text, Speech Dialogue. **7499**, 520–527 (2012)
34. Albayzin corpus in the European Language Resources Association. http://catalog.elra.info/product_info.php?products_id=746. Accessed 10 October 2014
35. D-L Choi, B-W Kim, Y-W Kim, Y-J Lee, Y Um, M Chung, in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC): 23-25 May 2012; Istanbul, Turkey (IEEE, USA)*, ed. by N Calzolari (Conference Chair), K Choukri, T Declerck, MU Dogan, B Maegaard, J Mariani, Odijk J, and S Piperidis. Dysarthric speech database for development of QoLT software technology, (2012), pp. 47-50
36. MJF Gales, PC Woodland, Mean and variance adaptation within the MLLR framework. Comput. Speech Lang. **10**, 249–264 (1996)
37. MJF Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**, 75–98 (1998)
38. P Červa, J Nouza, Supervised and unsupervised speaker adaptation in large vocabulary continuous speech recognition of Czech. LNCS (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **3658**, 203–210 (2005)
39. M Bohac, J Malek, K Blavka, in *Proceedings of the 36th International Conference on Telecommunications and Signal Processing (TSP): 2–4 July 2013; Brno, Czech Republic (IEEE, U.S.A.)* Iterative grapheme-to-phoneme alignment for the training of WFST-based phonetic conversion, (2013), pp. 474–478
40. F Rudzicz, Using articulatory likelihoods in the recognition of dysarthric speech. Speech Commun. **54**, 430–444 (2012)

doi:10.1186/s13636-014-0039-0
Cite this article as: Bohac *et al.*: **A cross-lingual adaptation approach for rapid development of speech recognizers for learning disabled users.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:39.