Cost-Efficient Development of Acoustic Models for Speech Recognition of Related Languages

Jan NOUZA, Petr ČERVA, Michaela KUCHAŘOVÁ

SpeechLab, Faculty of Mechatronics, Technical University of Liberec, Studentská 2, 46117 Liberec, Czech Republic

jan.nouza@tul.cz, petr.cerva@tul.cz, michaela.kucharova1@tul.cz

Abstract. *When adapting an existing speech recognition system to a new language, major development costs are associated with the creation of an appropriate acoustic model (AM). For its training, a certain amount of recorded and annotated speech is required. In this paper, we show that not only the annotation process, but also the process of speech acquisition can be automated to minimize the need of human and expert work. We demonstrate the proposed methodology on Croatian language, for which the target AM has been built via cross-lingual adaptation of a Czech AM in 2 ways: a) using the commercially available GlobalPhone database, and b) by automatic speech data mining from HRT radio archive. The latter approach is cost-free, yet it yields comparable or better results in experiments conducted on 3 Croatian test sets.*

Keywords

Speech recognition, acoustic model, cross-lingual adaptation, Slavic languages.

1. Introduction

Modern systems for large-vocabulary continuous speech recognition (LVCSR) are designed in the way that allows for easy separation of language dependent and language independent components. The former include an acoustic model (AM), a lexicon with pronunciations, a language model (LM) and an optional text pre-processing and post-processing module. The latter part consists namely of a signal processing front-end and a decoder. When an existing system is to be ported to another language, only the former have to be developed. Thus, there is a natural demand to make this porting in a fast and cost-efficient manner.

As an example, we can mention the efforts of the LIMSI team to adapt their LVCSR system (developed originally for French and English [1]) to other major languages, like e.g. Arabian [2], as well as to those spoken by much smaller population, like Finish [3] or even Luxembourgish [4]. Another research team with a strong focus on multi-lingual speech processing works at the Karlsruhe Institute of Technology. It has collected a large database of spoken data in 20 languages known as GlobalPhone [11] and used it for the development of multi-lingual LVCSR systems. While the creation of the lexicon and the corresponding LM for the target (European) language is the easier task - thanks to digital text resources available via Internet ([5], [6]) - the development of the proper AM requires a large amount of speech records and their phonetic transcriptions. The latter can be provided either manually by an expert (a phonetician) or by automated procedures combined with a varying degree of human supervision [7].

Our research in this field has been motivated by the fact that during the last decade we developed an LVCSR system for Czech that proved to be practically usable in off-line as well as on-line applications, such as broadcast news (BN) transcription [8], spoken archive processing [9] or voice dictation. Later, the system was adapted also to Slovak [10], and recently, we are working on other related languages, like Polish, Russian and Croatian. Our focus on Slavic languages has several rational reasons: a) we can utilize the existing LVCSR system tailored specifically for inflected languages with very large vocabularies, b) we can benefit from the fact that these related languages share some similar and specific patterns in phonetics, lexical inventories, morphology and grammar, c) these languages have attracted less interest from the world-wide research community, so far.

In this paper, we describe the methodology that helped us in a rapid and cost-efficient development of AMs for these languages. In the next section, we provide a brief review of main approaches used for cross-lingual AM adaptation. After that, we propose two methods that reduce the amount of human work in the process of acquisition of phonetically annotated speech data and that can be applied without an expert familiar with the target language. The methods are evaluated experimentally on Croatian, which has been the most challenging language from the above mentioned ones, namely due to limited text and speech resources (as Croatian is spoken by some 5 million people). Yet, the results obtained on 3 different test sets show that the AM created during a several-week period of mostly automated work is applicable for demonstrating a potential of the Croatian LVCSR system.

2. Related Work

Before starting the training of an AM for an LVCSR system, a certain amount of speech from various speakers must be collected and annotated on the phonetic level. For some languages, annotated speech databases suitable for AM training are available, usually on a commercial base. If this is not the case, the phonetic transcriptions must be created, either manually by skilled annotators or by some automated procedures. The well-known Forced Alignment algorithm is the best option when precise orthographic transcriptions are available. In case of large multi-lingual databases, like e.g. GlobalPhone [11], phonetic annotations are missing and the orthographic ones may contain various errors or inconsistencies: from completely or partially wrong texts or corrupted audio, to minor mistakes, like omitted, added or switched words. In such a situation, the transcription process must include a procedure that is capable of discovering and handling these errors. It is usually done by incorporating the iteratively evolving LVCSR system as a checking tool [12].

When an AM for a new language is developed within a multi-lingual environment, the process generally starts by a bootstrapping phase where either one or more existing AMs serve for initializing HMMs of phonemes and noises. The initial model is used to transcribe the data in the target language, which is followed by a series of iterative re-training steps with gradually increasing amount of data. The maximum-likelihood training approach is usually combined with model and feature adaptation techniques as shown, e.g. in [13]. One of the most recent methods, which seems promising particularly for low-resource languages, is based on sharing acoustic data from multiple resources and representing the target AM by subspace Gaussian mixture models [14]. Last but not least, it should be mentioned that the phonetic transcriptions can be omitted, if the AM is built on graphemes rather than on phonemes. The results published for Russian [15] or Slovak [16] show, however, that the classic phoneme-based HMMs always outperform the grapheme-based ones.

3. Developing AM in Efficient Way

The goal of our work is to develop AMs for various Slavic languages with minimum costs. Yet, we want the performance of these AMs to be as high as possible, as it will allow us to use unsupervised training and adaptation techniques in later stages when more data is available.

We start the AM process building with bootstrapping from a Czech phoneme-based AM and then we utilize two schemes. One is applicable to speech data with orthographic (but not necessarily error-free) annotations, and its main goal is to use the existing LVCSR system to generate phonetic transcriptions, to check these annotations and identifying possible inconsistencies in them. The other approach is based on searching for publically available audio data that contain speech and for which some additional text information (e.g. in form of summaries, captions or quotations) can be found. By matching the text resources with the output of the LVCSR we identify the portions with a high level of agreement and utilize them for iterative retraining of the target AM.

3.1 Generating Phonetic Transcriptions for Imperfect Speech and Text Data

When using a speech database provided by a third party, we should be prepared for the situation that not all audio and text data are perfect. Some major errors, like missing files, missing parts of utterances or their transcriptions, can be discovered early, but smaller errors caused either by speakers or annotators are hard to be detected without an expert in the target languages. If the degree of inconsistence is high (which may be true even for some established databases as shown in Section 4), a straightforward application of the forced alignment technique would not be the best option for generating phonetic transcriptions. In this case, it is important to apply a procedure that is capable of checking the audio and reference text content and identifying potential problems. In our scheme, it is the developed target LVCSR system itself that plays the role of the expert who automatically checks the files, generates the transcriptions, and provides hints to a human supervisor where he or she should intervene.

3.1.1 Basic Scheme

The scheme runs iteratively, with the following steps:

1. Preparation. It consists in preparing the existing LVCSR for running with the lexicon and LM of the target language. Pronunciations in the lexicon are temporarily mapped to the phonetic inventory of the source language.

2. Initialization. The existing AM from the source languages is used. All files in the database are labeled as *NotChecked*.

3. Transcription. All *NotChecked* files are transcribed using the LVCSR system and the current AM.

4. Matching. For each audio file, the recognizer's output is matched to the reference text. To quantify the agreement on the word level, we use the standard Word Error Rate (WER) measure:

$$WER = (N_S + N_D + N_I)/N.100\% \quad (1)$$

where N_S, N_D, N_I are the numbers of substitutions, deletions, insertions and N is the total number of words, respectively. As the recognizer produces also a phonetic transcription, we can match it to the phonetic transcription generated from the reference text using the lexicon or a grapheme-to-phoneme transducer (G2P). By applying the same matching procedure as for the words, we get a similar measure denoted as Phoneme Error Rate (*PER*).

5. Classification. If an utterance yields $WER = 0$, it is (almost) sure that the reference text is correct, the audio

file is uncorrupted and the automatically generated phonetic transcription is appropriate. If $WER > 0$ but $PER = 0$, it means that the text disagreement is caused either by homophones or spelling variants. Each utterance is classified into one of the three classes: *Accepted* (if $WER = 0$ or $PER = 0$), *ToBeChecked* (if $WER < T_W$) and *NotChecked* (otherwise).

6. Manual Check. The *ToBeChecked* utterances are those with little disagreement. It can be 1 to 3 words (depending in the utterance length) when we set threshold value $T_W = 10$ %. Using a simple check program, we can visualize the differences, listen to them and correct either the reference text or the LVCSR output (both the orthographic and the phonetic one). After that, the checked utterance get label *Accepted*.

7. Checkpoint. If there are no new *Accepted* utterances in the current iteration, the procedure is stopped here.

8. Retraining. A new AM is trained using phonetic transcriptions of all *Accepted* utterances.

9. Repeat. The procedure goes back to step 3.

3.1.2 Enhanced Scheme

The above described scheme can be further improved to get faster progress with less human work.

a) The WER values will be reduced if the lexicon and LM used in step 1 are better fitted to the given speech database. It can be done by adding (temporarily) the database specific words to the lexicon and similarly by adding the reference texts to the LM training corpus.

b) If we are not sure about the pronunciation of some words in the target language, or about the proper mapping of target language phoneme set, we can use multiple variant pronunciations and let the recognizer decide which one is more appropriate or statistically more frequent.

c) Although the amount of human work needed in step 6 is significantly smaller compared to full manual check, it still can be reduced if more utterances get the *Accepted* label. This can be achieved by utilizing several different AMs in step 3 of each iteration. These AMs can differ only slightly, e.g. by the number of HMM mixtures or by using global or sliding-window Cepstral Mean Subtraction (CMS) parameterization. It is possible to use also an AM which is trained on the mix of the already *Accepted* utterances (from the target language) with some amount of speech from the source language. It is very likely that each of these different AMs will produce a slightly different set of *Accepted* utterances, and hence their total number in each iteration will be increased. We demonstrate the positive effect of this idea in Section 4.

d) If the size of the training data in the target language is large enough, the transcriptions and the lexicon are remapped back to the original phoneme set. (After this step, however, only the target language AM can be employed.)

3.2 Automatic Speech and Text Data Mining from Web

The main problem of training speech databases is that their size is limited and they are available only for some languages. Though, one can find a lot of audio files containing speech on Internet, e.g. in publicly accessible archives of radio and TV broadcasters, on web pages of some institutions, like parliaments, senates, courts, etc. In some cases, these audio files are accompanied also with texts. The ideal situation occurs when these texts are verbatim transcriptions of speech files. In this trivial case, no special procedure is needed to align them and to make them a part of the training data. In most cases, however, the text differs to some extent from the speech. A high degree of correspondence occurs, e.g. between broadcast speech and attached close captions, which has been often utilized for so called lightly supervised AM training (e.g. in [17]). In other cases, the accompanying texts may be just summaries of what was spoken, news articles containing quotations, or documents that were discussed e.g. during a parliament session, etc. However, even these loosely related texts can serve for collecting data suitable for AM training.

If a source (usually a web page) containing both text and speech is found, one of the four situations illustrated in Fig. 1 can occur: a) the text has nothing in common with the audio file, b) the text and the speech share some common words (usually prepositions, conjunctions, pronouns) that are randomly scattered, c) the text and speech contain coincident phrases (strings of few words), and d) the two sources are related in the way, that some spoken utterances occur as written (not necessarily verbatim) sentences in the text. The last case is a good opportunity for automatic acquisition of new training data. Yet, most of the potential Internet sources are mixes of the four cases, with case d) often being the least frequent one. Anyway, if the source is large and the used data mining method is robust, we can collect a considerable amount of new training material.

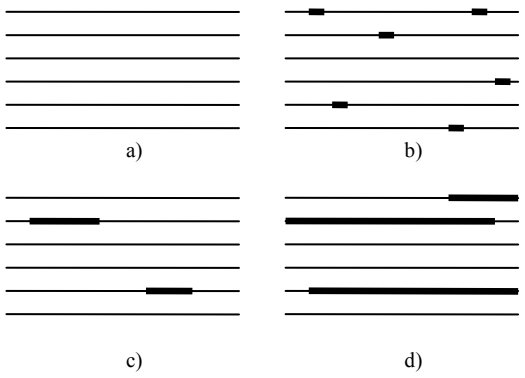

Fig. 1. Web page text and its parts found in audio (bold) a) no correspondence, b) randomly scattered words, c) shorter phrases, d) longer utterances.

LVCSR *raw*	...	skijašice	[noise]	startaju	sutra	prva	vožnja	počinje	u	15			druga	u	18	[noise]	muški	…
LVCSR w_j	…	skijašice		startaju	sutra	prva	vožnja	počinje	u	15			druga	u	18		muški	…
TEXT r_i	…	skijašice		startaju	sutra	prva	vožnja	počinje	u	15	sati	a	druga	u	18			…
LABEL		H	-	H	H	H	H	H	H	H	D	D	H	H	H	-	I	

START STOP

Fig. 2. Example of alignment of a part of LVCSR output (raw and text-only) with a part of reference text (in Croatian). Here, 2 words occurring in text were not found by LVCSR (words "sati" and "a" were not spoken) and recognized word "muški" did not appear in text. Symbols START and STOP denote endpoints of an eligible word sequence. The actual cut points are moved to the nearest occurrence of silence.

3.2.1 Searching for Speech Related to Text

The first step consists in finding a large publicly accessible Internet source that is structured into smaller units, usually web pages, containing audio files and some text. Then, for each page, we search if there are speech and text segments that correspond to each other. The search is based again on matching the text (it will be further referred to as a *reference*) to the output of the currently available LVCSR system. For the alignment of the two strings we have proposed a variant of the Minimum Edit Distance (MED) algorithm, inspired by [18], that prefers local rather than global alignment of word sequences.

The algorithm searches for the optimal alignment of two sequences, the reference comprised of J words r_j and the recognizer output comprised of I words w_i (numbers I and J can differ significantly). This type of tasks is usually solved by the dynamic programming approach, using distance matrix **A** as a space where the solution is searched. The procedure starts with initialization:

$$\begin{aligned} A(i,0) &= P_D \cdot (i-1), 1 \le i \le I \\ A(0,j) &= P_I \cdot (j-1), 1 \le j \le J \end{aligned} \tag{2}$$

and continues with the recursive computation of all other $A(i,j)$ values:

$$\begin{aligned} A(i,j) = \min[&A(i-1,j-1) + d(r_i, w_j) - b_{i-1\,j-1}; \\ &A(i,j-1) + P_I; A(i-1,j) + P_D] \end{aligned} \tag{3}$$

where

$$d(r_i, w_j) = \begin{cases} 0 & \text{if } r_i = w_j \\ P_S & \text{if } r_i \ne w_j \end{cases} \tag{4}$$

and

$$b_{ij} = \begin{cases} 0 & \text{if } r_i \ne w_j \\ b_{i-1\,j-1} + 1 & \text{if } r_i = w_j \end{cases} \tag{5}$$

Constants P_D, P_I and P_S are penalties associated with word deletion, insertion and substitution, respectively. Their optimal values depend on which of the cases shown in Fig. 1 are typical for the source data. We use a small subset of this data to determine them experimentally. Auxiliary values b_{ij} help us to keep a track of non-interrupted hit sequences. In (3), these obtain a small bonus.

When all cells in matrix **A** are computed, the best alignment path is revealed by a standard backtracking procedure from final point (*I,J*) to starting point (1,1). Each word in reference is assigned one of the labels: Hit (H), Substitution (S), Deletion (D) or Insertion (I).

The next step consists in identifying word sequences where the reference text and the LVCSR output are either same or differing only slightly. The algorithm goes through the word labels and searches for sequences with dominating hits. Formally expressed, we search for a string of words W_1, W_2,... W_N that meets the following constraints:

$$\begin{aligned} &N_{\min} < N < N_{\max} \\ &Label(W_1) = Label(W_2) = Label(W_N) = Hit \\ &N_H > (N_S + N_I + N_D) \end{aligned} \tag{6}$$

The sequence should have minimum length N_{min} and for practical reasons it should not be longer than N_{max} (otherwise it is split). The first, second and last words must be labeled as hits, and the total number of hits N_H in the sequence should be higher than the rest. The last constraint may seem weak but let us note that at this level we search for data that will be later processed with an LVCSR system whose performance will improve in time and some non-hit terms get a chance to be classified correctly.

In the last step, the utterances belonging to the eligible sequences are cut out from the original (often very long) audio files and stored with the corresponding text. The cut points are derived from the time stamps associated to each word (and non-speech event) during the LVCSR procedure. To minimize problems with inaccurate cuts at the beginning and end of the utterance, the actual cut points are moved to the center of the nearest noise event (usually silence or breath). The whole process is illustrated in Fig. 2.

3.2.2 Making Training Database from the Mined Data

After completing the process described above, we get a collection of audio files with reference texts, i.e. data

similar to many standard speech databases. Obviously, we must be aware of the fact that the text transcriptions can contain errors but the same happens also to official databases. What is missing is information about speakers. Yet, there are ways to cope with this problem. In case of broadcast archives, many web pages mention the name of the editor, who often is the main speaker in the audio file. On parliament or senate pages, the speaker name is often explicitly stated. Another alternative consists in utilizing speaker recognition methods to identify different speakers. This speaker clustering is necessary if we want to apply a limit for the amount of the data provided by a single speaker. After this step, the speech database is ready for the process described in Section 3.1.

4. Evaluation on Croatian LVCSR

The methods proposed in the previous section have been successfully applied for the development of LVCSR systems in four languages (Slovak, Russian, Croatian and Polish). In the following text we will focus only on their evaluation on Croatian, as it has been the most challenging language so far, mainly because of very limited resources.

4.1 LVCSR System Applied to Croatian

The evaluation experiments were conducted on the standard LVCSR system originally developed for Czech and recently described e.g. in [19]. Its front-end processes 16 kHz audio data, converts them into 39 MFCC features, applies global or floating CMS, and HLDA. The Czech AM uses triphone HMMs to represents 41 Czech phonemes and 7 types of noise. Its recent version has been trained on 320 hours of speech (of various types). The decoder runs in real-time with vocabularies up to 500K words and a bigram LM smoothed by Kneser-Ney method is used in standard one-pass mode.

When preparing the system for Croatian, we collected from Internet a large corpus (940 MB) of newspaper text. We used it to compile a 255K lexicon and a bigram LM based on 28M different word-pairs. Three Croatian specific phonemes (represented by graphemes 'ć', 'đ' and 'lj') were mapped to the closest Czech counterparts. More details on these basic preparation steps can be found in [20].

4.2 Speech Data for Training and Testing

In this study we used 3 sources of Croatian data, the GlobalPhone set, the COST set and the HRT web resource.

4.2.1 GlobalPhone - HR

This data is part of a large multi-lingual speech corpus collected by the team at the University of Karlsruhe [11]. Recently it includes 20 languages from various parts of the world and its subsets are distributed on commercial base via ELRA [21]. Unlike the other language sets in the GlobalPhone (GP) collection, the Croatian one has some specific features. First, its size is smaller compared to the other sets. It contains 4499 utterances that were recorded by 92 speakers. (Most other language sets contain about 10,000 recordings from 100 speakers). Second, the distribution of the recordings among the speakers is not balanced, as some speakers recorded less than 30 sentences while some others contributed more than 100 ones. Third, the speech is supposed to be read but in many cases the speakers did not read given sentences fluently, they mispronounced words, repeated them, made false starts, or they uttered words different from those in the text form. These mistakes and the fact that most speakers were actually speaking Bosnian (using different words, e.g. 'hiliada' instead of 'tisuća', and slightly different pronunciation) complicates automatic processing of the recordings. Obviously, the database as it is can be used for training the AM applicable for Croatian LVCSR experiments as it was shown in [5]. However, in this case, native human annotators (who are able to discover and fix the errors) are necessary.

4.2.2 COST278 - HR

This is another multi-lingual speech database. It was created within European COST278 project to support international collaboration on broadcast news processing, namely in speaker segmentation and clustering tasks [22]. It includes 5 to 10 complete TV shows in 9 languages (about 3 hours per each), including Croatian. Each show is manually segmented and orthographically transcribed.

4.2.3 HRT Radio Speech Data

When searching for additional speech resources we discovered the web archive of the major public broadcaster in Croatia, HRT. Its regional stations have their own web sites, with pages devoted to short local news. The news is described by text and occasionally also by audio. In most cases, the correspondence between the text and speech resembles situations a), b) or c) illustrated in Fig. 1. However, the amount of available audio (several hundred hours) and text (about 10K files) allows for experimenting with the method proposed in Section 3.2. For this purpose we have chosen data covering the 1/2010 to 7/2012 period.

4.2.4 Test Sets

Set	Speech style and recording year	Size in minutes	#words	OOV [%]
GP	speech produced by amateurs (1998)	59	7386	1.96
COST	read/planned speech by professionals (2003)	35	5052	1.18
HRT	read/planned speech by professionals (2013)	27	4088	1.13

Tab. 1. Description of three Croatian test sets.

For evaluation, we used the following test data: a) utterances of speakers 02, 03, 04, 06, and 07 from the GlobalPhone set, b) 307 speech segments from 2 COST278

TV shows and c) 104 utterances mined from HRT radio station Pula (news broadcasted in January 2013).

4.3 Bootstrapping with Czech AM

In the first series of experiments, we tried to measure, what performance can be achieved with a purely Czech AM. The second question was which type of Czech AM is optimal for bootstrapping a Croatian system. We compared an AM tuned for the best performance in Czech LVCSR with two AMs represented by a lower number of parameters (physical states). The results are in Tab. 2.

	WER [%] for 3 test sets		
AM parameters	**GP**	**COST**	**HRT**
5575 states, discrim. training	32.36	25.12	28.40
4044 states, EM training	32.07	24.78	26.54
2041 states, EM training	28.47	23.65	26.39

Tab. 2. Performance achieved with 3 Czech acoustic models.

The figures show that all AMs, and especially those less fitted to Czech, have an acceptable performance in the initial tests. We concluded that for bootstrapping, the 2041-state model was the optimal choice.

4.4 AM Trained on Croatian GlobalPhone Set

As explained earlier, the Croatian GlobalPhone set is a really challenging speech resource. Its precise phonetic annotation would be a difficult task even for a native speaker or a skilled phonetician. The main problems stem from low acoustic quality of some recording sessions, influent speech interrupted by many restarts, incorrect orthographic transcriptions, inconsistent pronunciation and the use of Bosnian language by more than half speakers. There are also occasional background voices or audible prompts from the recording supervisors. Hence, the application of the transcription method described in Section 3.1 promises to save a lot of tedious manual work.

Before launching the proposed iterative procedure we slightly adapted the general-purpose lexicon and the LM to better fit the GlobalPhone utterances. This step was necessary, as the database was recorded in 1998 and most utterances deal with war and post-war events of that period. About 200 most frequent OOV words were added to the vocabulary and all sentences (except those used for testing) were included in the LM training corpus.

The transcription process run according to the enhanced scheme described in Section 3.1. In each iteration loop, we used several available AMs. To illustrate their effect, let us compare performance of two of them: one based entirely on the already transcribed Croatian data, the other trained on the mix of the same data and 10 hours of randomly chosen Czech training sentences. In Tab. 3, we can observe that the additional Czech data helped to improve the AM and to reduce WER values, especially in initial stages. Another advantage of the mixed model is that its Czech part supplies training data for (so far) rarely seen phoneme context and, in particular, for noise models. Not only these two AMs but also their variants differing in numbers of mixtures (32 or 16) or in the application of the CMS normalization (global or floating) were utilized in each iteration.

	WER [%]									
Hours	1	2	3	4	5	6	7	8	9	10
HR	29.2	24.9	23.7	24.4	22.1	21.8	21.4	20.7	20.3	20.0
HR+CZ	25.3	22.8	21.5	21.7	21.0	20.5	20.7	20.5	19.8	19.7

Tab. 3. WER obtained on GP test set for AMs trained on increasing amounts of Croatian speech (HR) and with 10 hours of Czech (HR+CZ).

The amount of transcribed data after each iteration is shown in Fig. 3. We can see that, e.g. after the first iteration (in which bootstrapped Czech AMs were used), we got 1.7 hours of phonetically annotated data. From this amount, 1.3 hours were obtained automatically, 0.4 hours required small manual corrections related to 1 or 2 words (either in the reference text or in the LVCSR output). We can also notice that the largest gain occurred during the first 4 iterations. The process was stopped after the 12th iteration, when 11.5 hours were transcribed. The remaining amount (1.7 hours) from all the 13.2 hours allocated for training was not used, as these data had either bad acoustic quality or they were hard to be corrected by a non-expert. The whole procedure consumed mainly computer time, while the total amount of required manual work took just a small portion of it.

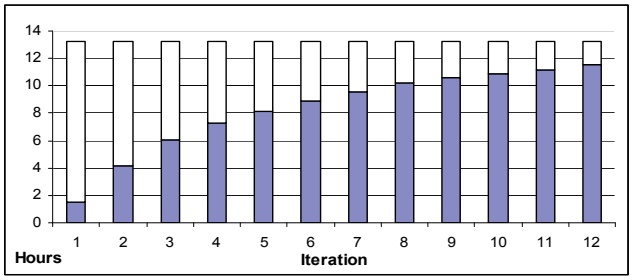

Fig. 3. Amount (in hours) of transcribed GlobalPhone data during the iterative procedure described in Section 3.1.

In each iteration step there were approx. 100-200 utterances that passed the threshold $T_W = 10\,\%$ and that were eligible for human check. We designed a simple tool that highlights the difference between the reference text and the LVCSR output. The tool can play either the whole utterance or the selected part. In most cases, the human supervisor just needs to decide which word is correct (either the reference or the recognized one) and make the adjustment by one click. No prior knowledge of the target language is needed for this type of action. If the annotator is not sure about the correct word (because it is unknown to him/her, or its pronunciation is unclear or incomplete, or it is masked by noise, etc) he/she can skip the utterance and remove it from further processing. One of the biggest benefits of this scheme is that the human work is focused only on those utterances that require minimum effort. The time spent by correcting the Croatian recordings labeled as *ToBeChecked* was about 2-3 hours in each iteration.

The effect of the AM trained on these 11.5 hours was evaluated on the 3 test sets. The results presented in Tab. 4 can be directly compared with those in Tab. 2. We can see that the improvement in performance is significant.

	WER [%] for 3 test sets		
AM	**GP**	**COST**	**HRT**
Trained on 11.5 hours of GP data	19.93	16.29	18.13

Tab. 4. WER values obtained with AM trained on GlobalPhone data.

4.5 AM Trained on HRT Radio Archive Data

The goal of this experiment was to verify how good can be an AM that is trained entirely on data automatically collected from web. As explained in Section 4.2.3, our source was HRT radio archive. We found 11,851 web pages that contained both audio files and text. On a small subset of these data (200 pages randomly chosen) we analyzed the correspondence between the audio and text content. Unfortunately, the most favorable case (that depicted in Fig. 1d) was very rare. In most cases, the alignment between the text and speech revealed no common sequences. During a parameter optimization process, we set the following constants: $P_S = 15$, $P_D = 10$, $P_I = 3$, $N_{min} = 10$ and $N_{max} = 25$. Using the bootstrapped Czech AM we run the method described in Section 3.2. It found 3,694 segments that met the constraints defined by (6). Their total duration was 8.8 hours. After that, this data passed through the same iterative procedure as applied to the GlobalPhone data. In this case, only 5 iterations were necessary to transcribe (mostly automatically) the complete set and to train the final AM. The smaller number of iterations can be explained by the fact that the selected segments contained mainly clean speech produced by professionals in studio. Moreover, many manual interventions dealt with the reference text rather than the recognized one.

The results achieved with this AM are listed in Tab. 5. When comparing them to those in Tab. 4, we can see that with one exception (the GP test set), the performance is better, in spite of a smaller amount of the training data. Let us also remind that this data are cost-free. Obviously, the process of the audio data mining could be repeated with the new Croatian AM and it is expected that more data would be acquired.

	WER [%] for 3 test sets		
AM	**GP**	**COST**	**HRT**
Trained on 8.8 hours of HRT data	19.98	14.89	15.14

Tab. 5. WER values obtained with AM trained on HRT data.

4.6 AM Trained on All Available Data

In the last experiment, we made a natural step and put all the available training data together: GlobalPhone (11.5 hours), HRT (8.8 hours) and 1.3 hours acquired through the same transcription scheme from the remaining part (3 TV shows) of the COST database. We trained the final AM on these 21.6 hours of Croatian data coming from three different sources and three time periods (1998, 2003, 2010-2012). The results are summarized in Tab. 6. We can notice a consistent improvement for all the three test sets.

	WER [%] for 3 test sets		
AM (# physical HMM states)	**GP**	**COST**	**HRT**
Trained on 21.6 hours of HRT, GP and COST data (1541states)	17.55	14.12	14.28

Tab. 6. WER values obtained with AM trained on all available data.

5. Discussion and Conclusions

We have proposed and evaluated two schemes that can save a significant portion of human work in developing acoustic models for languages that are related to one with an existing AM. Both schemes utilize a LVCSR system as a tool that performs the two functions: The first is to check the validity of orthographic transcriptions that are provided either explicitly, e.g. as a part of a speech database, or that can be acquired from public sources like Internet. The second function is to generate phonetic transcriptions by using a lexicon (or a G2P transducer), choosing between alternative pronunciations, and identifying and labeling non-speech sounds.

We have also shown that an acoustic model for a new language can be trained without a dedicated, commercially distributed speech database. The data we acquired automatically from publicly available Internet sources enabled us to train an AM whose performance is better than that made of the Croatian part of the GlobalPhone database.

Both schemes have been already used in practice: for Croatian - as documented in this paper - and also for Slovak, Russian and Polish. (Let us note that the quality of the Russian and Polish GlobalPhone subsets was significantly better compared to the Croatian one.) The availability of the AMs for the other Slavic languages allows us to further enhance the proposed methods, for example by utilizing multiple and multi-lingual acoustic models within the bootstrapping phase. To examine the idea, we have run a simple experiment, in which five AMs, each developed for one language, were tested on the three Croatian sets. From the results presented in Tab. 7, we can observe that the Slovak AM would be even better in the bootstrapping phase than the Czech one was.

	WER [%] for 3 test sets		
AM (# physical HMM states)	**GP**	**COST**	**HRT**
Czech (2041 states)	28.47	23.65	26.39
Slovak (3764 states)	26.09	19.58	22.36
Polish (2035 states)	32.37	27.04	26.22
Russian (3382 states)	33.54	27.85	30.35
Croatian (1541 states)	17.55	14.12	14.28

Tab. 7. WER values obtained with AMs representing 5 languages.

The AMs developed for the four Slavic languages represent a good starting point for demonstrating the potential of an LVCSR in tasks like broadcast news tran-

scription. In each of the four languages, we got close to the 15-percent-WER level, at least for read speech. This level allows for running a system that can monitor broadcast news programs and save the data for further AM improvements, via lightly supervised or even unsupervised techniques, which is our main research direction in this field, recently.

Acknowledgements

This work was supported by the Czech Science Foundation (project no. P103/11/P499) and by the Technology Agency of the Czech Republic (project no. TA01011204).

References

[1] GAUVAIN, J-L., LAMEL, L., ADDA G. The LIMSI broadcast news transcription system. *Speech Communication*, 2002, vol. 37, no. 1-2, p.89-108.

[2] LAMEL, L., MESSAOUDI, A., GAUVAIN, J-L. Automatic speech-to-text transcription in Arabic. *ACM Transactions on Asian Language Information Processing*, 2009, vol. 8, no. 4, p. 1-17.

[3] LAMEL, L., VIERU, B. Development of a speech-to-text transcription system for Finnish. In *Proc. of SLTU 2010*. Penang (Malaysia), 2010, p. 62–67.

[4] ADDA-DECKER, M., LAMEL, L., ADDA, G. A first LVCSR system for Luxembourgish, an under-resourced European language. In *Proc. of LTC'11 Workshop*. Poznan (Poland), 2011, p. 47-50.

[5] VU, N. T., SCHLIPPE, T., KRAUS, F., SCHULTZ, T. Rapid bootstrapping of five eastern European languages using the rapid language adaptation toolkit. In *Proc. of Interspeech 2010*. Makuhari (Japan), p. 865–868.

[6] PROCHAZKA, V., POLLAK, P., ZDANSKY, J., NOUZA, J. Performance of Czech speech recognition with language models created from public resources. *Radioengineering*, 2011, vol. 20, no. 4, p. 1002-1008.

[7] SCHULTZ, T., BLACK, A. Rapid language adaptation tools and technologies for multilingual speech processing. In *Proc. of ICASSP 2008*. Las Vegas (USA).

[8] NOUZA, J., ZDANSKY, J., CERVA, P., KOLORENC, J. Continual on-line monitoring of Czech spoken broadcast programs. In *Proc. of Interspeech 2006*. Pittsburgh (USA), p. 1650-1653.

[9] NOUZA, J., et al. Voice technology to enable sophisticated access to historical audio archive of the Czech radio. *Multimedia for Cultural Heritage*. Springer Berlin Heidelberg, 2012, CCIS vol. 247, p.27-38.

[10] 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*. Brisbane (Australia), p. 2683-2686.

[11] SCHULTZ, T. GlobalPhone: A multilingual speech and text database developed at Karlsruhe University. In *Proc. of ICSLP 2002*. Denver (USA), 2002, p. 345–348.

[12] SCHULTZ, T., WAIBEL, A. Language-independent and language-adaptive acoustic modelling for speech recognition. *Speech Communication*, 2001, vol. 35, no 1–2, p. 31–51.

[13] LÖÖF, J., GOLLAN, C., NEY, H. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system. In *Proc. of Interspeech 2009*, Brighton (UK), p. 88-91.

[14] BURGET, L., et al. Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In *Proc. of ICASSP'10.* Dallas (USA).

[15] STÜKER, S., SCHULTZ, T. A Grapheme based speech recognition system for Russian. In *Proc of Specom 2004*. St. Petersburg (Russia), September 2004.

[16] MIRILOVIC, M., JUHAR, J., CIZMAR, A. Comparison of grapheme and phoneme based acoustic modeling in LVCSR task in Slovak. *Multimodal Signal: Cognitive and Algorithmic Issues*. Springer, LNAI, vol. 5398, 2009, p. 242–247.

[17] LAMEL, L., GAUVAIN, J-L., ADDA, G. Lightly supervised and unsupervised acoustic model training. *Computer, Speech & Language*. 2002, vol. 16, no. 1, p. 115–129.

[18] HIRSCHBERG, D. S. Algorithms for the longest common subsequence problem. *J. of the ACM*, 1977, vol. 24, no.4, p. 664–675.

[19] NOUZA J., et al. Making Czech historical radio archive accessible and searchable for wide public. *Journal of Multimedia,* 2012*, v*ol. 7, no. 2, p. 159-16.

[20] NOUZA, J., CERVA, P., ZDANSKY, J., KUCHAROVA, M. A study on adapting Czech automatic speech recognition system to Croatian language. In *Proc. of Elmar 2012*. Zadar (Croatia), 2012, p. 227-230.

[21] ELRA catalogue, GlobalPhone - HR, ref. number ELRA-S0195.

[22] ZIBERT, J., et al. The COST278 broadcast news segmentation and speaker clustering dvaluation. Overview, methodology, systems, results. In *Proc. of Interspeech 2005*. Lisbon (Portugal), 2005, p. 628-631.

[23] HRT archive available at http://www.hrt.hr/

About Authors ...

Jan NOUZA (*1957) received his M.Sc. and Ph.D. degrees at the Czech Technical University (Faculty of Electrical Engineering) in Prague in 1981 and 1986, respectively. Since 1987 he has been teaching and doing research at the Technical University in Liberec. In 1999 he became a full professor. His research focuses mainly on speech recognition and voice technology applications, such as voice-to-text conversion, dictation, broadcast speech processing and design of voice-controlled tools for handicapped persons. He is the head of SpeechLab group at the Institute of Information Technology and Electronics.

Petr ČERVA (*1980) received the Master degree and the Ph.D. degree from the Technical University of Liberec (TUL), in 2004 and 2007, respectively. He is currently an assistant professor at the Inst. of Information Technology and Electronics at TUL. His research interests are speaker adaptation and speech recognition.

Michaela KUCHAŘOVÁ (*1987) received her MSc degree in Information Technology from TUL in 2011. She joined SpeechLab as a PhD student and recently she works mainly on linguistic topics, with a special interest in multilingual issues.